On Guidelines for Comparative Effectiveness Research Using Nonrandomized Studies in Secondary Data Sources

Sebastian Schneeweiss, MD, ScD
Division of Pharmacoepidemiology, Department of Medicine, Brigham and Women’s Hospital and Harvard Medical School, Boston, MA, USA

Keywords: comparative study, guidance, health-care decision-makers, methodology.

Generating knowledge on the comparative effectiveness of medical products with nonrandomized studies is challenging and will not infrequently result in biased effect estimates if epidemiologic principles are not followed [1]. Nevertheless, a recent study by Academy Health suggested that currently, about 85% of studies producing comparative effectiveness research (CER) are nonrandomized and I would not expect too much change in the near future as much as we all love randomized effectiveness trials [2]. The sizable and sudden influx of research funds to perform CER via the American Recovery and Reinvestment Act in the United States is a unique chance to help jump-start and establish the field. This challenge comes with a nontrivial risk. What if after all the money is spent, it turns out that a good proportion of the studies produced invalid results or results that will not answer the questions that decision-makers struggle with?

The nonrandomized world of research on the effectiveness of therapeutics is complex, certainly more complex than the world of randomized trials. The now published guidelines on “Good Research Practices for Comparative Effectiveness Research with Nonrandomized Studies of Treatment Effects using Secondary Data Sources” [3–5] by the corresponding ISPOR Task Force will serve as a welcome reminder that we should keep the basics in mind when embarking in CER: be clear on what questions we want to answer and make sure we answer it in a way that policymakers and physicians can meaningfully act upon the new evidence. Start out with an incidence user design, define an exposure risk window that is appropriate given a product’s biological action, thoughtfully choose a comparison group with identical or similar indication and of practical relevance, and pay utmost attention to confounding by indication caused by the channeling of patients into treatment groups based on their progression related to the study outcome [6]. These considerations paired with solid clinical knowledge and understanding of real-world utilization will avoid many grave mistakes. Most of the advanced analytic methods discussed in the guidelines are important additions to the epidemiologist’s toolbox but will not overcome errors in fundamental design consideration.

How are guidelines different from pharmacoepidemiology textbooks? They might be more up-to-date but less detailed. A major similarity, however, is that one will have to read guidelines, learn more about specific issues depending on one’s background, and study new methods. If researchers are not willing to go back to school with these guidelines, their mission is largely in vain. This raises a fundamental issue that is critical for the establishment of a functioning and productive CER community: the capacity to produce researchers trained in the advanced methods of pharmacoepidemiology and technology assessment using nonrandomized studies. Such training is available in some but not many academic institutions in North America and Europe, but not enough graduates are produced to satisfy the growing demand. The National Institutes of Health does not have a T32 training grant mechanism for our discipline, which seems to fall between the cracks of an organization largely organized by disease areas. Rigorous training is the most effective way to ensure a high quality of nonrandomized research on the effectiveness of medical products.

The authors of these guidelines are to be applauded for their work, and now, it is up to the practitioners to study them and brush up on some newer methodologies, while the guidelines can go into round 2 of an ongoing revision cycle.

Disclosure: Dr. Schneeweiss is the president of the International Society for Pharmacoepidemiology and is the principal investigator of the Brigham and Women’s Hospital DEcIDE Center for Comparative Effectiveness Research funded by the Agency for HealthCare Quality and Research. The opinions stated in the article are his own and not representing any of those organizations.

References
ISPOR Health Policy Council Proposed Good Research Practices for Comparative Effectiveness Research: Benefit or Harm?

Til Stürmer, MD, MPH, Tim Carey, MD, MPH, Charles Poole, MD, ScD
Department of Epidemiology, UNC Gillings School of Global Public Health, Chapel Hill, NC, USA

There are increasing calls for better understanding of “what works” in health care [1]. One of the means for assessing what works is through “comparative effectiveness research” (CER) [2]. Ideally, the needed data would come from randomized controlled trials (RCTs) or from natural experiments. RCTs would need to be large, practical clinical trials that compare interventions head-to-head in real clinical settings [3,4], using novel approaches to assess clinically relevant outcomes.

Nonexperimental studies of intended drug effects have been criticized because confounding by indication (selective channeling of patients to treatment modalities based on outcome predictors such as severity of disease) can almost never be ruled out [5,6]. Recent developments in pharmacoepidemiologic methods limit the potential for bias, and thus increase the value of non-experimental comparisons of intended drug effects. These methods include instrumental variable methods [7,8], the new user design [9], the use of a comparator drug with a similar indication to that of the index drug [10], propensity scores [11], and simple improvements such as eliminating immortal person-time [12] and reducing selection bias by not censoring follow-up when a person stops taking a drug [13,14]. Much remains to be done, however, including the study of heterogeneity of treatment effects at the intersection between personalized medicine and pharmacoepidemiology. In addition, there remains an unresolved tension between emulating RCTs (increasing internal validity based on increased restrictions [15]) and enhancing generalizability (external validity).

CER is an interdisciplinary endeavor in which the disciplines are linked by the need for information and the development of methods. The involvement of ISPOR in this enterprise is welcome. Like drugs, the Good Research Practices proposed by the working group of the ISPOR Health Policy Council and published in this issue of the journal [16–18] need to be evaluated by their potential benefits and harms. The potential benefits are obvious. Someone unfamiliar with performing non-experimental comparisons of drugs and their outcomes will find valuable discussion in these documents of issues to be considered. Common to all such documents, however, there is the potential for harm when the recommendations are used as a cookbook without understanding their interplay. References to standard textbooks of pharmacoepidemiology [e.g., 19] could help alleviate this problem.

It is inevitable for any detailed overview to contain questionable- or outmoded recommendations. Recommendations of the proposed documents that some experienced pharmacoepidemiologists might find argueable include:

1. The requirement to report the results from all ex ante analyses. Some such analyses will have been abandoned because the researchers discovered that they are biased.
2. The assessment of the importance of biases based on how they affect the acceptance or rejection of the null hypothesis. Biases are best measured by their effects on the magnitude, direction and precision of effect-measure estimates.
3. The recommendation for propensity score models to include variables that are only weakly related to treatment selection (but unrelated to the outcome per the following recommendation). It is unclear why any variable that is unrelated to the outcome should be included in a propensity score [20].

Some core issues in the design and analysis of non-experimental comparisons are not addressed in enough detail. One is the importance of the role of various “stakeholders” in CER. Another is the distinction between confounding (e.g., by indication) and selection bias (due, for instance, to non-adherence, drop-out of “sick stoppers,” etc. [13]). The potential to separate these forms of bias is one of the main advantages of the new user design [9].

Given the expected continuation of the rapid development of pharmacoepidemiologic methods over the past 5 years, the proposed Good Research Practices may become outdated very rapidly [21]. We found no indication of how ISPOR intends to keep these guidelines up to date. In an era of guideline proliferation, one might ask what the proposed ISPOR document will add to the existing ones in this field, especially the Good Pharmacoepidemiologic Practice document published and continuously updated by the International Society for Pharmacoepidemiology (ISPE) [22]. Finally, harmonization of the ISPOR documents with others, including those proposed by ISPE, the US Institute of Medicine, the US Agency for Healthcare Research and Quality, the UK National Institute for Health and Clinical Excellence should be considered. Such harmonization would prevent confusion and nit picking by groups opposed to CER.

In our view, the benefit-to-harm balance of ISPOR’s proposed documents on Good Research Practices favors the benefit side. It will help to spread the news that non-experimental treatment comparisons are possible given careful design, analysis, and interpretation. We congratulate ISPOR for providing guidelines for CER that emphasize the potential benefits without giving CER a black box warning for its potential harms.

Address correspondence to: Til Stürmer, Department of Epidemiology UNC Gillings School of Global Public Health Epidemiology, McGavran Greenberg Hall, CB # 7435, Chapel Hill, NC 27599-7435, USA. E-mail: til.sturmer@post.harvard.edu 10.1111/j.1524-4733.2009.00653.x

Dr. Tim Carey has no conflict to declare; he receives salary support from the Agency for Healthcare Research and Quality.

Dr. Charles Poole receives salary support from the UNC-GSK Center of Excellence in Pharmacoepidemiology and Public Health and has received speaking fees from GSK.

Source of financial support: Dr. Til Stürmer receives investigator-initiated research funding and support as principal investigator and coinvestigator from the National Institute on Aging at the National Institutes of Health. He also receives research funding as coprincipal investigator of the...
UNC-DeCIDE center from the Agency for Healthcare Research and Quality. Dr. Stürmer does not accept personal compensation of any kind from any pharmaceutical company, though he receives salary support from the UNC-GSK Center of Excellence in Pharmacoepidemiology and Public Health and from unrestricted research grants from pharmaceutical companies to UNC.

References

Objectives: Health insurers, physicians, and patients worldwide need information on the comparative effectiveness and safety of prescription drugs in routine care. Nonrandomized studies of treatment effects using secondary databases may supplement the evidence based from randomized clinical trials and prospective observational studies. Recognizing the challenges to conducting valid retrospective epidemiologic and health services research studies, a Task Force was formed to develop a guidance document on state of the art approaches to frame research questions and report findings for these studies.

Methods: The Task Force was commissioned and a Chair was selected by the International Society for Pharmacoeconomics and Outcomes Research Board of Directors in October 2007. This Report, the first of three reported in this issue of the journal, addressed issues of framing the research question and reporting and interpreting findings.

Results: The Task Force Report proposes four primary characteristics—relevance, specificity, novelty, and feasibility while defining the research question. Recommendations included: the practice of a priori specification of the research question; transparency of prespecified analytical plans, provision of justifications for any subsequent changes in analytical plan, and reporting the results of prespecified plans as well as results from significant modifications, structured abstracts to report findings with scientific neutrality; and reasoned interpretations of findings to help inform policy decisions.

Conclusions: Comparative effectiveness research in the form of nonrandomized studies using secondary databases can be designed with rigorous elements and conducted with sophisticated statistical methods to improve causal inference of treatment effects. Standardized reporting and careful interpretation of results can aid policy and decision-making.

Keywords: comparative effectiveness, health policy, nonrandomized studies, secondary databases.

Background to the Task Force

In September 2007, the International Society for Pharmacoeconomics and Outcomes Research (ISPOR) Health Science Policy Council recommended that the issue of establishing a Task Force to recommend Good Research Practices for Designing and Analyzing Retrospective Databases be considered by the ISPOR Board of Directors. The Council’s recommendations concerning this new Task Force were to keep an overarching view toward the need to ensure internal validity and improve causal inference from observational studies, review prior work from past and ongoing ISPOR task forces and other initiatives to establish baseline standards from which to set an agenda for work. The ISPOR Board of Directors approved the creation of the Task Force in October 2007. Task Force leadership and reviewer groups were finalized by December 2007 and the first teleconference took place in January 2008.

Task Force members were experienced in medicine, epidemiology, biostatistics, public health, health economics, and pharmacy sciences, and were drawn from industry, academia, and as advisors to governments. The members came from the UK, Germany, Austria, Canada, and the United States. Beginning in January 2008, the Task Force conducted monthly teleconferences to develop core assumptions and an outline before preparing a draft report. A face-to-face meeting took place in October 2008 to develop the draft, and three forums took place at the ISPOR meetings to develop consensus for the final draft reports. The draft reports were posted on the ISPOR website in May 2009 and the task forces’ reviewer group and ISPOR general membership were invited to submit their comments for a 2-week reviewer period. In total, 38 responses were received. All comments received were posted to the ISPOR website and presented for discussion at the Task Force forum during the ISPOR 12th Annual International Meeting in May 2009. Comments and feedback from the forum and reviewer and membership responses were considered and acknowledged in the final reports. Once consensus was reached, the manuscript was submitted to *Value in Health*.

Introduction

Health insurers, physicians, and patients worldwide need information on the comparative effectiveness and safety of prescription drugs in routine care. Nonrandomized studies of treatment effects using secondary databases may supplement the evidence based from randomized clinical trials and prospective observational studies. Recognizing the challenges to conducting valid retrospective epidemiologic and health services research studies, a Task Force was formed to develop a guidance document on state of the art approaches to frame research questions and report findings for these studies.
Drugs in routine care. Although randomized clinical trials (RCTs) are the gold standard to determine a drug's efficacy against placebo, it is well recognized that results of such studies may not accurately reflect effectiveness of therapies delivered in typical practice [1–3]. In addition, clinical decisions usually involve choices among therapies yet sponsors of drug trials have limited motivation to test new drugs against existing therapies [4]. Routinely collected and electronically stored information on healthcare utilization in everyday clinical practice has proliferated over the past several decades. Large computerized databases with millions of observations of the use of drugs, biologics, devices, and procedures along with health outcomes may be useful in assessing which treatments are most effective and safe in routine care without long delays and the prohibitive costs of most RCTs.

There is controversy, however, on how to best design and analyze nonrandomized studies on comparative treatment effects using secondary databases, including claims databases, patient registries, electronic medical record databases, and other routinely collected health-care data. Challenges of conducting epidemiologic and health services research studies from secondary data sources include concerns about the adequacy of study design, the relevance of the population and timeframe available for study, approaches to minimize confounding in the absence of randomization, and the specificity of clinical outcome assessment. Such threats to validity limit the usefulness of these studies and adoption of findings into policy and practice. With proper research design and application of an array of traditional and newer analytic approaches, such concerns can be addressed to improve our understanding of treatment effects. (See Parts II and III of this Task Force Report, also in this issue [5,6].) This report will suggest that to optimize the validity of findings from observational studies designed to inform health-care policy decisions, researchers employ a priori hypotheses in written protocol and data analysis plans before study implementation, that they follow reporting standards that make transparent to readers if, why, and how their analytic plans evolved, as well as provide a justification of the suitability of the database to test their hypotheses.

Although we recognize that exploratory analyses and data mining of large datasets are often used to generate hypotheses regarding the effectiveness and comparative effectiveness of treatments, stricter criteria for the design and execution of studies as well as transparency in their reporting are required to justify the conclusion that such findings are robust enough to warrant changes in clinical practice or to influence policy decisions.

Thus, the objective of this report is to lay out good research practices for comparative therapeutic effectiveness studies using secondary databases. We present the report in three sections: Defining, Reporting and Interpreting Nonrandomized Studies; Design Issues; and Analytical Issues. By describing best practice, this report will serve to improve future research, assist in evaluating the validity of existing studies, and suggest how these studies should be interpreted for decision-making; it may also be of use to journal editors who are responsible for the peer-review process for publication. We do not seek to be complete in our discussion of analytic options, nor will we fully explain all methods, but rather focus on the issues surrounding the most relevant designs and analytic techniques for secondary databases.

It is important to be explicit about the definition of comparative effectiveness as it is applied by the authors of this report. With respect to the term comparative, this report will focus on the majority of circumstances when comparison can be made between two or more active treatments rather than comparisons made between an active treatment and “no treatment.” With respect to the term effectiveness, this report will focus on the benefits of therapies evaluated rather than harms (as extensively examined in the field of pharmacoepidemiology) or costs (as extensively examined in pharmacoconomics and health services research).

The assessment of the comparative benefits of various treatment options through the analysis of secondary databases is controversial. Nevertheless, policymakers, payers, and providers are increasingly turning to the analysis of large secondary databases to answer a variety of questions. For example, when there are no head-to-head RCT or prospective observational study data available, these data sets are used to examine whether and to what magnitude there are differences in benefit associated with various treatments including drugs in the same or different drug classes. Even if there are published head-to-head clinical trials, there may be a reason to suspect that there may be clinically or policy important differences in treatment effectiveness in real-world usage in comparison with the outcomes observed in RCTs—perhaps driven by differences in target population, adherence/compliance, or other important factors. Additionally, RCTs are frequently designed to examine intermediate or surrogate measures as outcomes; thus, there may be the desire to examine the magnitude of benefit when assessed on true outcome measures over longer observation durations (e.g., mortality, disability). Finally, when decision-makers want comparative effectiveness data to inform trade-offs driven by different profiles of benefits, harms and costs among treatment options, secondary databases provide a highly valuable source of information.

Defining the Question

As we have moved from the early 20th century practice of medicine—a cottage industry based more on anecdotal experience and characterized by enormous practice variation—to a 21st century practice that is based upon treatment guidelines and accountability for quality—a more systematic approach in which inappropriate practice variation will be discouraged—the inter-relationship between health-care policy decision-makers and those generating the evidence to support their decisions has become more complicated.

Fueling this evolving relationship is the continued rise in health-care costs at a rate faster than the overall economy. Thus, governments and payers are making health policy decisions regarding access and reimbursement of new health-care technologies that have not always delineated separate answers to the three cardinal questions of evidence-based technology evaluation (as set forth by Archie Cochrane): “Can it work? Will it work? Is it worth it?” [7]. Answering the latter question comprises judgments about both comparative effectiveness (What are the advantages—from the perspective of the patient, the provider, and the payer—that a new technology provides over the available standard of care?) and contextual value considerations (Can we afford it? Do we get a good “bang for the buck” relative to alternatives? How does this square with precedent and our collective preferences in allotting scarce resources to health care?).

To adequately answer these questions, the development of evidence is ideally an iterative process between decision-makers, those who generate evidence and those who evaluate and summarize the body of evidence relevant to particular health policy questions [8]. Considering these questions individually—especially the value question—is critical to a transparent and fair appraisal process and can be enhanced by appropriate structuring of the process [9].

Questions regarding comparative effectiveness are often restructured into operational terms such as “How does this drug
compare to other similar drugs on the formulary in terms of clinical outcomes?" "What role should a new drug play in the treatment of a particular condition and how will this be described in a treatment guideline?" and questions of value are often worded in terms such as "should we grant reimbursement for a new drug and at what contract terms or price?"

Because the majority (if not all) of health policy decisions are made with imperfect information to inform these questions—due to either imperfect evidence or the absence of the desired evidence—the decision-making process is confounded further. When is the evidence “good enough” to make a recommendation? The alternative—waiting for perfect evidence—is usually not acceptable, because we never have perfect evidence and we are incapable (either due to cost or feasibility) to perform “gold standard” RCTs to answer the myriad questions posed for a forever-growing armamentarium of health-care technologies. Hence, we must update the traditional evidence hierarchy [10] and look to ways to optimize the use of observational data that will increasingly be automated into health-care delivery.

Currently, there is reluctance by many health policy decision-makers to use observational data—especially data from retrospective analysis of large data sets—to inform their deliberations. Many decision-makers are uncertain about the reliability and robustness of results derived from observational studies, primarily because of concerns about confounding and selection bias. This distrust is derived, at least in part, from the lack of generally accepted good research practices and lack of standardized reporting; it is also due to discordance of the results in examination of clinical effectiveness between some observational studies and randomized controlled clinical trials. Understanding the source of these discordances, in turn, also relies upon a rigorous approach to the design and analysis of observational studies. Thus, we believe, that creation and adoption of good observational research practices will augment their use and credibility.

Prospective Specification of Research and Protocol Development

Arguably, the most important and challenging part of research is to clearly and precisely articulate the objective of a study in the form of a focused research question [11,12] before the design and execution of a study (i.e., a priori). One strength of clinical trials is the requirement for a study protocol which specifies inclusion criteria for subjects, primary and secondary outcomes, and analytic approach. Although there are differing views in medical science regarding a priori specification of a research hypothesis when conducting observational research [13], prior specification minimizes the risk of "cherry-picking" interesting findings and a related issue of observing spurious findings because of multiple hypothesis testing [14]. For these reasons, we recommend the practice of a priori specification of the research question and study design in a formal study protocol and data-analysis plan is strongly advised to assure end-users that the results were not the product of data-mining. (Note: this is not an indictment of data-mining per se; rather that data-mining is more appropriate for hypothesis generation, rather than hypothesis testing).

As part of the protocol, the rationale for the observational study should be explicitly stated. For example, there are no direct comparative data on the effectiveness of various treatment options or that available data have only examined the short-term consequences of treatment and decision-makers were seeking information on long-term outcomes. When defining the research question, four primary characteristics are proposed:

Relevance and Rationale

The research questions and hypotheses should be highly topical and meaningful from a clinical, policy, or research methodology perspective not only at time of study conception but, perhaps more importantly, at the anticipated time of submission for publication or presentation to the relevant audience.

Specificity

The research question should be concise yet unambiguous, should relate to the stated research objectives where relevant, should state the intervention and outcome of interest where relevant, should identify the patient population, and should focus on one primary end point. Existing data sources must be adequate to provide valid identification of the appropriate patients, interventions, and outcomes. The protocol methods section should discuss the strengths and weaknesses of a secondary database with respect to its suitability in answering the primary research questions.

Novelty

Proposals should clearly identify what a new study can add to existing knowledge. At one extreme, there may be an absence of literature that directly relates to the proposed study question thereby making the proposed research question novel. Alternatively, the proposed study design for the given research question may improve on previous studies. Previous findings may have been inconclusive, conflicting or questioned because of study limitations. Finally, even when some research exists (including clinical trials), there may be a need to validate findings. As the number of well-designed studies addressing a specific question whose findings are consistent with each other increases, the value of an additional study addressing this question diminishes.

Feasibility

Investigators should recognize that conducting a rigorous observational study can be as challenging as conducting trials and should ensure that studies are feasible with respect to power of the study to answer a question, time and resources required, and ability to link necessary data sources. There should also be adequate numbers of patients and events to yield sufficient power for the primary analysis. Timing can be important because some areas change so rapidly that the answers may no longer be relevant if it takes several years to collect and analyze data. Finally, even where data already exist, there can be substantial hurdles to linking data from different systems to conduct the intended analysis.

In formulating the research question with the above-mentioned characteristics, two suggestions may be helpful: 1) “begin at the end” [11]; and 2) know the limitations of the available data. Envisioning the one key table or figure required to answer the research question is extremely helpful in focusing the research question and understanding what can feasibly be extracted from the available data. Also, a sound understanding of data limitations will also help to understand which research questions should or should not be studied with the available data sources.

Once the research question has been defined, a sound study protocol should be developed with this study question in mind [12]. Key components of a study protocol include study background and rationale, research question/objective, study design, study population, data sources and storage where relevant, study timeframe, specific study definitions, one prespecified primary end point, secondary end points, statistical analysis (including
sample size and power where relevant), informed consent process where relevant, and mock output tables and/or figures [12]. A written detailed data analysis plan (DAP) should also accompany the protocol; a good DAP will include definitions of outcomes, measures of treatments, and identify all covariates. The DAP should provide general specification of any modeling that is contemplated. We recognize that analytic plans often require adjustment once researchers begin to analyze secondary datasets. We recommend that researchers be transparent about their ex ante analytic plans, provide justification for subsequent changes in analytic models, and report out the results of their ex ante analytic plan as well as the results from its modifications. In addition, researchers may wish to establish explicit limits on the evolution of the analytic plan—beyond which any results should be considered hypothesis-generating—and not appropriate for making clinical practice or policy recommendations. For example, one might consider establishing the boundary when a hypothesis-testing study changes into a hypothesis-generating study. Following extraction of the analytic dataset and completion of prespecified primary analyses, researchers frequently discover “bugs” in their analyses—perhaps because of coding problems in the data or because of the algorithms applied to define exposure; appropriate correction of these “bugs” is well accepted. Nevertheless, if important flaws become evident in the analytic approach such that different analytic approaches must be applied, this should signal that the study should be considered hypothesis-generating and not hypothesis-testing.

Recommendations

- A priori specification of the research question and study design in a formal study protocol and data-analysis plan is strongly advised.
- Be transparent about ex ante analytic plans, provide justification for subsequent changes in analytic models, and report out the results of their ex ante analytic plan as well as the results from its modifications.

Selection of Study Design Appropriate to the Study Question

Although numerous epidemiologic and econometric study designs exist [15–21], the choice of study design is almost always determined by both the research question and feasibility constraints [22,23]. It is crucial to be absolutely uncompro-mising about design aspects of a study that might hamper its validity [24]. A detailed review of typical study designs used in clinical research is provided elsewhere [15–21]. Several key study designs used in observational research are outlined in the subsequent sections. Guidelines have recently been proposed on the reporting of observational studies, specifically as it relates to cross-sectional, cohort, and case-control studies [22,23].

Cross-Sectional Designs

The cross-sectional study examines a “snapshot” of data and typically either describes the data available in that snapshot or attempts to make correlations between variables available in the dataset. Although this study design can provide some valuable information, it is typically limited by its inability to characterize temporality—it is often uncertain whether the exposure preceded the outcome of interest or vice versa. In research questions where temporality of exposure and outcome are important, alternative designs should be selected.

Cohort Designs

In a cohort study, groups of patients (i.e., cohorts) exposed to drug therapies are followed over time to compare rates of the outcomes of interest between the study cohorts. Temporal relationships between exposure and outcome can be well characterized in a cohort study and both relative and absolute risks can be reported directly with the use of this design. Consequently, this design may be of particular interest for research questions requiring absolute risk estimates and where the temporal nature of associations is important to characterize.

Case-Control Designs

Case-control designs involve the identification of individuals who experience an outcome of interest (i.e., the cases) and those who do not (i.e., controls). Exposure to an intervention of interest in a period before the designation of case or control status is then compared between cases and controls. This design has historically been used when the outcome of interest is rare, maximizing the capture of such precious outcomes. Analysis of case-control designs typically provide estimates of relative risk but do not directly provide absolute risk estimates.

Case-Time-Control Designs

A primary challenge of cohort and case-control studies is the selection of comparable comparison groups. In case-crossover studies, individuals serve as their own controls. Only those individuals who experience the outcome of interest (i.e., cases) and were exposed to a treatment of interest within a certain time before the outcome date are included. Exposure to the treatment of interest in the period immediately before the outcome is compared with the exposure prevalence in a period more distant to the date of the event of interest in the same individual. Exposure prevalences are then compared between more recent and distant exposure windows to arrive at a risk ratio. Case-crossover designs are ideally suited for transient exposures that result in acute events but require sufficient numbers of patients who have both an event and are exposed to the drug of interest in either the nearby or more distant exposure windows. This design may be particularly attractive for research questions involving the comparison of groups that are extremely different in their clinical profiles (i.e., where major selection bias may exist) and involve transient exposures and immediate outcomes.

Case-Time-Crossover Designs

A limitation of case-crossover designs is temporal confounding where the prevalence of treatment exposure is higher in the exposure window closer to the event date relative to the exposure window more distant to the event date simply because of naturally increasing treatment uptake over time rather than a truly casual relationship. To circumvent this issue, a control group of individuals who do not experience the event of interest is created and analyzed in a manner similar to the cases to estimate the “natural” increase in treatment exposure prevalence over time—the exposure prevalence in the exposure window closer to the event date is compared with the exposure prevalence in the exposure window in a more distant period to arrive at a risk ratio amongst controls. The “case” risk ratio is then divided by the “control” risk ratio to arrive at an overall risk ratio. This design also requires sufficient numbers of patients who have both an
event and exposure to the treatment of interest in either of the predefined exposure windows and issues of selection bias in comparing case to controls may still be problematic [24].

**Interrupted Time Series Designs**

Interrupted time series analysis typically involve cross-sections of data over time both before and following an event of interest. Actual trends in exposures or outcomes following an event of interest are then compared with expected trends based on patterns of historical data before the event of interest. For example, in assessing the impact of a drug policy on drug utilization, historical trends would be used to establish an expected drug utilization rate in the absence of the policy change [25]. This expected drug utilization rate would then be compared with observed rates occurring following the implementation of the drug policy using advanced statistical approaches. The benefit of conducting a time series analysis is the minimization of problematic selection bias. Challenges, however, include issues related to temporal confounding (i.e., other events that may have occurred simultaneously at the time of intervention) and the typical need for relatively large effect sizes. This design may be particularly relevant for research questions aimed at assessing the impact of events on drug utilization and immediate outcomes.

Although the above-mentioned descriptions serve a basic overview of selected study designs in pharmacoepidemiology, health services and outcomes research, study designs are not necessarily mutually exclusive. For example, the case-crossover design is inherent in the case-time-control design, a nested case-control study may involve a formal cohort study as part of its case ascertainment [26], and previous research has embedded cohorts of patients in time series analysis [27].

Explicit in the research question are the exposure and/or outcome of interest. The nature of the association between the exposure and outcome is often implied in the research question. For example, if the research question suggests the measurement of incidence of an event, a cohort study design may be preferred over a case-control study design.

Although the research question establishes the key parameters of the association being assessed, feasibility constraints such as small numbers of available patients and outcomes in the dataset, data quality, level of funding, and skill level of the researcher team may significantly influence the study design to be used in the analysis. For example, in measuring the association between an exposure and outcome, if the outcome stated in the research question is extremely rare and a modest budget is available for prospective data collection, a case-control study may be preferred over a cohort study.

**The Study Question Dictates the Choice of Data Source**

The data source must be able to adequately answer the study question using the selected research design. Several characteristics of the data source must be taken into consideration including the breadth and depth of the data in the database, the quality of the database, the patient population that contributes data to the database, and duration of information contained in the databases. For example, if the study question includes a highly specific, well-defined outcome, this outcome must be captured and well coded in the database being used in the research.

Two primary types of databases for observational research exist: medical records databases and administrative databases [28]. Data in the former are recorded as part of the process of clinical outpatient care while data in the latter are recorded as a by-product of financial transactions. Consequently, although administrative databases typically contain more general information on very large numbers of patients, medical records databases typically contain much more detailed clinical information on its patients (e.g., smoking status, lab results, and body mass index) that are often lacking in administrative databases. Medical records data may provide more extensive data for comorbidity adjustment for research studies that may be particularly susceptible to selection bias whereas administrative claims data, if considerably larger in numbers of patients captured, may be better suited for research questions that involve rare outcomes.

Electronic medical records (EMR) are emerging as a promising source of data for clinical research but come with their own sets of challenges [29], including the development and harmonization of data standards. Although these EMR-based datasets can provide a rich base of clinical information that is often not afforded by administrative databases, challenges typical of observational research such as selection bias will still persist.

Merging clinical and administrative datasets also provides the opportunity to leverage the strengths of each type of data. For example, rich clinical information for defined sets of patients can be merged into administrative data to limit the need for prospective follow-up of outcomes that are routinely collected in administrative databases [30]. Although the practice of merging such datasets has been increasing, the process of merging, privacy issues, and data quality and transferability must all be considered as part of the process.

Ultimately, the selected data source will need to have the required breadth and/or depth, duration, and quality of information dictated by the research question to provide findings meaningful to society.

**Reporting**

**Structured Abstract**

Reporting of results is a critical step in the conduct of scientific studies. It permits end users to make independent assessment of the strength and limitations of a study as well as to judge the robustness of the findings. In turn, this informs their assessment about the relevance and weight that study findings should be given in subsequent decision-making. This is particularly true for observational studies. Reporting of observational studies should allow users to understand clearly the primary question, reasons for choosing the particular data, the quality of the data sources, the processes to reduce bias, and the potential for results to be explained by factors other than a causal effect. Interpretation of the results should be placed in the context of other studies, especially randomized studies, and differences explained.

To this end, a standardized approach to reporting of observational studies should be adopted, similar to the CONSORT (Consolidated Standards of Reporting Trials) recommendations [31] which has been modified by STROBE (Strengthening the reporting of observational studies in epidemiology) statement [23]. The CONSORT and STROBE recommendations could be adapted as follows:
null hypothesis. We acknowledge that journal editors may not confounding that would have driven your result back toward the intentional confounding before study execution, or estimate the level of possible, provide an estimate of the expected magnitude of potential confounders.

To the extent well as the ultimate study results. If you had to make compromises to the modification of analyses or models. Report—at least in analytic plan evolved over time—explain what you found that led to the modification of analyses or models. Report—at least in summary terms—the results of the prespecified analytic plan as well as the ultimate study results. If you had to make compromises in the goals of the study, what did you do and why? To the extent possible, provide an estimate of the expected magnitude of potential confounding before study execution, or estimate the level of confounding that would have driven your result back toward the null hypothesis. We acknowledge that journal editors may not allow space for this level detailed reporting; nonetheless, we believe that this will enhance transparency of the research process and could be included in an appendix.

### Interpretation of Results

To interpret the results of observational studies, they must be put into the larger evidentiary context. When results of an observational study conflict with a well-conducted RCT, possible reasons for the discrepant findings should be systematically examined. These may include:

- Significant confounding present, whether or not it can be identified. These may include differences in adherence, confounding by indication, the impact of out-of-pocket costs, etc.
- Data quality is poor—biased to null.
- Different question: A different population was studied that exhibited a different response to therapy.
- Different question: Differential effects on outcomes used in observational studies and those used in the RCT (effectiveness vs. efficacy).

If there is no RCT to compare to, then should the results run counter to current understanding of biology and disease processes, they should not be considered as definitive but warranting further investigation. Indeed, in general, observational studies can be used to generate hypotheses worthy of additional study. As discussed elsewhere in this report, different observational study designs are better suited to hypothesis generation, such as cross-sectional studies. In contrast, a good case-control study may be ideal and informative of causal inferences under some circumstances (e.g., food borne outbreaks of disease with large relative risks). Well-conducted time-series studies can provide quite compelling data to support relative effectiveness in some circumstances (e.g., cervical cancer screening) and cohort studies are often good for assessing risks—but are frequently poor at assessing the effectiveness of interventions.

Second, reproducibility is a hallmark of robust evidence. Ideally, observational studies using the same analytic approach should result in substantially similar findings when applied to similar populations or different databases. Of course defining similar can be difficult. Simply demonstrating equivalence in the distribution of age, gender, ethnicity, or disease prevalence may not be enough. Differences in outcomes may be explained by variation in social and economic factors. Thus, the finding of reproducibility should enhance confidence in using the outcomes to inform decision-making; nevertheless, the absence should not rule out their use in assessing the body of relevant evidence.

Third, one can have greater confidence in the findings of an observational study, if an analysis of its cohort that is restricted to a subpopulation that is comparable with that of a published RCT provides similar results. Studies that examine an inception cohort of new users reduce the biases introduced by focusing on existing users, because nonresponders and those suffering adverse effects of therapy won’t be represented. This really is quite helpful because RCTs have higher internal validity and observational studies have higher external validity. Thus, they complement each other and enhance confidence in using the outcomes to inform decision-making.

Fourth, if an observational study is examining the relative effectiveness and safety of two different interventions, be suspicious of the robustness of small differences. If the point estimates of effect by the two treatments do not seem clinically compelling, then the evidence should be interpreted with caution. Differences

<table>
<thead>
<tr>
<th>SECTION AND TOPIC</th>
<th>ITEM #</th>
<th>DESCRIPTOR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Study Design</td>
<td>1</td>
<td>Description of Study Design</td>
</tr>
<tr>
<td>• Background</td>
<td>2</td>
<td>Scientific background and explanation of rationale including discussion of the suitability of the database employed</td>
</tr>
<tr>
<td>Methods</td>
<td>3</td>
<td>Clearly defined goals of the study with description of specific sub-questions. Description of study design and why it was chosen; Description of strengths and weaknesses of data source and how study groups were identified, including description of critical variables: diagnostic criteria, exposures, and potential confounders</td>
</tr>
<tr>
<td>• Measurement of treatment effects</td>
<td>4</td>
<td>Discuss how treatment effects were measured and how classification bias was addressed.</td>
</tr>
<tr>
<td>• Classification bias</td>
<td>5</td>
<td>Discuss how outcomes were measured and how classification bias was addressed.</td>
</tr>
<tr>
<td>• Confounding</td>
<td>6</td>
<td>Discuss the potential for confounding, both measured and unmeasured, and how this was assessed and addressed</td>
</tr>
<tr>
<td>Discussion</td>
<td>7</td>
<td>Interpretation of results, taking into account confounding and imprecision of results.</td>
</tr>
<tr>
<td>• Internal validity</td>
<td>8</td>
<td>Generalizability (external validity) of the study findings.</td>
</tr>
<tr>
<td>• Overall evidence</td>
<td>9</td>
<td>General interpretation of the results in the context of current evidence.</td>
</tr>
</tbody>
</table>
in the system of health-care delivery can have important effects on patient outcomes. Ideally, the investigators should define a priori what would be considered a clinically meaningful difference and interpret their findings in the light of this definition.

**Recommendations**

- A standardized approach to reporting of observational studies should be adopted.
- If there is no RCT to compare to, then should the results run counter to current understanding of biology and disease processes, they should not be considered as definitive but warranting further investigation.
- The finding of reproducibility should enhance confidence in using the outcomes to inform decision-making; nevertheless, the absence should not rule out their use in assessing the body of relevant evidence.
- One can have greater confidence in the findings of an observational study, if an analysis of its cohort that is restricted to a subpopulation that is comparable with that of a published RCT provides similar results.
- Be suspicious of the robustness of small differences. If the point estimates of effect by the two treatments do not seem clinically compelling, then the evidence should be interpreted with caution.

**How Findings Should Be Interpreted in Light of Policy Questions**

Health policy decisions vary widely in their scope. Increasingly, governments, payers, and providers are trying to base their decisions on the best available evidence. Nevertheless, numerous factors affect how findings are interpreted and whether they are incorporated into clinical or policy decisions. These include the direct relationship between the available evidence and the research question being asked, the magnitude of the observed effect, the generalizability of the research findings to broader populations, the limitations of the study, and the consistency of the findings with other available information. There are additional factors (for example, political and economic factors) that may alter uptake of research evidence. (We note that health policy decision-makers must also be informed of potential conflicts-of-interest involved in the generation of information—a subject beyond the scope of this report.)

Research that appears directly relevant to the policy question at hand is more likely to be used in decision-making. Although research is often performed by independent researchers, it is incumbent upon decision-makers to play a critical role in both defining the key questions and information characteristics that will be employed in making policy decisions. This will separately guide the evidence synthesis and decision-making framework [32,33]. Once a research question is agreed upon by both policymakers (the ones who need the information) and researchers (the ones who design and conduct the retrieval of information), the design must be rigorous in scientific principles and feasible in its implementation. Mock output in the form of figures and tables should also be agreed upon by decision-makers and researchers before research initiation to set expectations on both sides. Thus, in an ideal world, the process of understanding and answering the research question in a manner that will be useful for the decision-maker involves considerable investment of time by both the researchers and decision-makers at the very early stages of the research process. This is rarely the case today. Nevertheless, utilizing this approach would greatly enhance the impact of comparative effectiveness research on policy.

Researchers should put their findings into an appropriate context for policymakers. First and foremost, the findings of the study should have a logical relationship to the available relevant scientific literature. Findings that contradict the preponderance of evidence should be viewed with caution. Second, policymakers should be wary of small differences in effectiveness that may be statistically significant when found through analysis of very large datasets. Establishing beforehand what degree of clinical difference would be important from a policy perspective (for example, a 0.5% absolute reduction in HbA1c level) can help prevent over interpreting small differences. Third, policymakers should have a clear understanding of the strengths and limitations of a particular piece of research as it was conceived, designed, and executed; following our reporting recommendations will provide this level of transparency for decision-makers. Special attention should be paid to the generalizability of the results, the magnitude of confounding factors in the analysis, and the extent and degree to which the analytic plan required adaptation during study execution. One useful test is to see if one can reproduce findings of clinical trials when the population is restricted to subjects who would have been eligible for the clinical trial. If so, policymakers can have greater confidence in the direction and magnitude of differences observed between effectiveness and efficacy.

Our assessment of the strengths and weaknesses of the available evidence should be based on scientific principles of research design—for example, the appropriateness of the study designs, how well it was executed, the potential sources of bias—and the magnitude of potential biases, and consistencies of findings across multiple studies. At the same time, our willingness to make decisions based on evidence from nonrandomized studies will depend on the specific decisions we are making. In any decision-making, the risks of acting “too soon” (e.g., acting on findings subject to Type 1 error—mistaking a chance effect for a real one) are always weighed against those of acting “too late” (e.g., not acting based on findings subject to Type 2 error—missing a real effect because of studies that are underpowered). Different types of policy decisions may present different tradeoffs in the tension between the quality of available evidence and the need to make a decision.

At one end of the spectrum are regulatory decisions such as drug approval. The criteria for Food and Drug Administration (FDA) approval—at least two independent randomized trials with significant effects at the 0.05 level and with independent review of study protocols and data—seek to minimize the chances of allowing ineffective drugs on the market. These considerations make it unlikely that observational studies would play a large role in the initial approval process for pharmaceuticals. At the same time, recent studies of diabetes and lipid lowering drugs have raised concerns about relying on clinical trials that employ intermediate physiologic end points (e.g., glucose or lipid levels) rather than clinical outcomes (e.g., cardiovascular events) [34,35]. Careful postapproval observational studies may provide a more practical and politically acceptable alternative for validating effects on clinical outcomes than the alternative of requiring hard clinical end points in pivotal trials for new drugs, which would greatly increase the time and expense of the approval process.

At the other end of the spectrum, there are situations where the risks of acting “too late” may look greater than the risk of acting based on imperfect evidence. Many public health interventions are based primarily on observational evidence, in part because of the impracticality of requiring evidence based on randomized studies for interventions such as tobacco restrictions or seat belt laws. Similarly, because the imitations of RCTs for assessing drug safety are well known, most signals about safety
risks are derived from large epidemiologic studies. Responses to safety concerns can fall on a continuum that includes drug or device withdrawal (e.g., IUDs), restrictions on access (e.g., Accutane), prominent “black box” warnings (Celebrex) to clinical advisories (e.g., Champix/Chantix). These reflect both how strong and consistent the safety signal is, the potential for bias, the risks posed to the public and the consequences of limiting access to a potentially beneficial intervention.

Somewhere in the middle lie a range of decisions including developing recommendations in clinical practice guidelines and making coverage decisions for public and private insurers. Professional societies or other organizations producing evidence-based guidelines explicitly characterize the strength of individual recommendations and supporting evidence, and the evidence-hierarchies traditionally consider evidence from nonrandomized studies to be weaker than that from randomized trials. The Grading of Recommendations Assessment, Development and Evaluation (GRADE) process [36], which has been adopted by a growing number of international organizations, enumerates a number of factors that allows one to “upgrade” the quality of nonrandomized evidence. Equally important, GRADE (as well as the US Preventive Services Taskforce) make a distinction between quality of evidence and strength of recommendation, noting that one can make strong recommendations even when evidence is not high quality, for example when potential benefits far outweigh any potential harms or costs [37]. Of note, many guideline processes limit search strategies to clinical trials, especially in areas where trials are more numerous. Although more efficient, this process runs the risk of missing the potential value of large databases to answer these additional questions of generalizability and balance of harms and benefits observed in typical practice. Clinical recommendations in guidelines have not yet advanced very far in incorporating understanding of patient preferences and issues such as adherence and persistence to treatments. As none of these are well represented in efficacy trials enrolling highly selected volunteers, data from large observational databases could be useful for this process.

To have a consistent approach to the use of observational studies, decision-makers must understand in advance their tolerance for error in decision-making. When making comparisons, it is critical that the comparator chosen be reasonable and relevant to the decision at hand (What is the best available treatment alternative? What is the standard of care?) Using certainty and magnitude of benefit as key dimensions, a multistakeholder EBM Workgroup has developed a framework for describing judgments of comparative clinical effectiveness evidence into a matrix as shown in Figure 1 (with increasing certainty on the vertical axis and increasing comparative net health benefit on the horizontal axis [38].

Using such an approach, decision-makers may explicitly take into account limitations of evidence, including that from observational studies, with respect to the magnitude of perceived benefit and the robustness of the findings. This model is currently being used in comparative effectiveness reviews by the Institute for Clinical and Economic Review (ICER, http://www.icer-review.org).

**Figure 1** Framework for describing judgments of comparative clinical effectiveness evidence.

A = “Superior” [High certainty of a substantial comparative net health benefit]

B = “Incremental” [High certainty of a small comparative net health benefit]

C = “Comparable” [High certainty of a comparable comparative net health benefit]

D = “Inferior” [High certainty of a negative comparative net health benefit]

U/P = “Unproven with Potential” [Moderate certainty of a small or substantial comparative net health benefit]

This category is intended to represent bodies of evidence that provide a best estimate in comparative net health benefit as small or substantial but without enough precision to judge which is more likely. The U/P category also implies that there is a relatively small possibility that future evidence would demonstrate that the true net comparative benefit is inferior to other alternatives for many or all patients.

I = “Insufficient” The evidence does not provide high certainty that the net health benefit of the technology is at least comparable with that provided by the comparator(s).
Conclusion

Information regarding comparative effectiveness of therapies is increasing in importance. Nonrandomized studies using secondary databases can be designed with rigorous elements and conducted with sophisticated statistical methods to improve causal inference of treatment effects. The next two sections of our report will address design and analysis issues directly. When results from these studies are obtained, we suggest standard methods to report them, and reasonable caution in interpreting them.

Source of financial support: This work was supported in part by the Department of Veteran Affairs Health Services Research and Development grant HFP020-90. The views expressed are those of the authors and do not necessarily reflect the views of the Department of Veteran Affairs.

References


Emily Cox, PhD,1 Bradley C. Martin, PharmD, PhD,2 Tjeerd Van Staa, PhD, MD, MSc, MA,3 Edeltraut Garbe, MD, PhD,4 Uwe Siebert, MD, MPH, MSc, ScD,5 Michael L. Johnson, PhD6

1Express Scripts, St. Louis, MO, USA; 2Division of Pharmaceutical Evaluation and Policy, College of Pharmacy, University of Arkansas for Medical Sciences, Little Rock, AR, USA; 3General Practice Research Database, London, UK; 4Department of Clinical Epidemiology, Bremen Institute for Prevention Research and Social Medicine, Bremen, Germany; 5Department of Public Health, Medical Decision Making and Health Technology Assessment, University of Health Sciences, Medical Informatics and Technology, Hall, Austria; Adjunct Professor of Public Health Policy and Management, Harvard University, Cambridge, MA, USA; 6University of Houston, College of Pharmacy, Department of Clinical Sciences and Administration, Houston, TX, USA. E-mail: mikejohnson@uh.edu

Objectives: The goal of comparative effectiveness analysis is to examine the relationship between two variables, treatment, or exposure and effectiveness or outcome. Unlike data obtained through randomized controlled trials, researchers face greater challenges with causal inference from observational studies. Recognizing these challenges, a task force was formed to develop a guidance document on methodological approaches to address these biases.

Methods: The task force was commissioned and a Chair was selected by the International Society for Pharmacoeconomics and Outcomes Research Board of Directors in October 2007. This report, the second of three, reported in this issue of the Journal, discusses the inherent biases when using secondary data sources for comparative effectiveness analysis and provides methodological recommendations to help mitigate these biases.

Results: The task force report provides recommendations and tools for researchers to mitigate threats to validity from bias and confounding in measurement of exposure and outcome. Recommendations on design of study included: the need for data analysis plan with causal diagrams; detailed attention to classification bias in definition of exposure and clinical outcome; careful and appropriate use of restriction; extreme care to identify and control for confounding factors, including time-dependent confounding.

Conclusions: Design of nonrandomized studies of comparative effectiveness face several daunting issues, including measurement of exposure and outcome challenged by misclassification and confounding. Use of causal diagrams and restriction are two techniques that can improve the theoretical basis for analyzing treatment effects in study populations of more homogeneity, with reduced loss of generalizability.

Keywords: comparative effectiveness, epidemiology, nonrandomized studies, research design, secondary databases.

Background to the Task Force

In September 2007, the International Society for Pharmacoeconomics and Outcomes Research (ISPOR) Health Science Policy Council recommended that the issue of establishing a task force to recommend Good Research Practices for Designing and Analyzing Retrospective Databases be considered by the ISPOR Board of Directors. The Council’s recommendations concerning this new task force were to keep an overarching view toward the need to ensure internal validity and improve causal inference from observational studies, review prior work from past and ongoing ISPOR task forces and other initiatives to establish baseline standards from which to set an agenda for work. The ISPOR Board of Directors approved the creation of the task force in October 2007. Task force leadership and reviewer groups were finalized by December 2007 and the first teleconference took place in January 2008.

The task force members were experienced in medicine, epidemiology, biostatistics, public health, health economics, and pharmacy sciences, and were drawn from industry, academia and as advisors to governments. The members came from the UK, Germany, Austria, Canada, and the United States.

Beginning in January 2008, the task force conducted monthly teleconferences to develop core assumptions and an outline before preparing a draft report. A face-to-face meeting took place in October 2008, to develop the draft and three forums took place at the ISPOR meetings to develop consensus for the final draft reports. The draft reports were posted on the ISPOR website in May 2009 and the task forces’ reviewer group and ISPOR general membership were invited to submit their comments for a 2-week reviewer period. In total, 38 responses were
Introduction

The goal of comparative effectiveness analysis is to examine the relationship between two variables, treatment or exposure, and effectiveness or outcome. The advantages of using secondary databases to examine this relationship are easily recognized by researchers in the field. Compared with data obtained through randomized controlled trials (RCTs), secondary data sources provide a low-cost means of answering the research question, answers can be obtained in a relatively short time frame, the data are more representative of routine clinical care and large cohorts of patients can be followed over long time periods [1]. However, researchers should be mindful of data limitations that, in some instances, preclude their use. In this section, we will address issues of validity with respect to secondary data sources and, where appropriate, provide researchers with tools to help mitigate threats to validity.

Researchers have been writing about the challenges that secondary data sources pose for more than two decades now [2–4], and although challenges still exist, the methodological approaches to address these challenges have greatly improved [5,6]. Key in contributing toward inaccuracies in administrative data is the fact that they were built for billing and record keeping purposes, not for research. Therefore, the potential for error occurs at many points along the record keeping process [7]. The implication for researchers is that both systematic and random error can occur in the identification of treatment exposure and outcome.

In RCTs, identifying and measuring exposure is done with a great deal of accuracy and precision. For example, in a clinical trial evaluation of drug treatment, not only is it known who has received the active drug, but also the degree of exposure—dose, duration and compliance with therapy. Similarly, outcomes—or measures of effectiveness—are measured with a great deal of accuracy and precision. Various devices and laboratory tests are used to measure and record both surrogate (blood pressure, cholesterol levels, tumor staging) and final end points (e.g., myocardial infarction [MI], stroke, and even death). This same level of measurement—exposure

| Table 1 Secondary data sources and measurement of drug exposure |
|----------------------------------|-----------------|--------------------|-----------------|
| Data source                      | Measurement     | Type of exposure measured | Limitations                               |
| Outpatient prescription claims   | NDC or therapeutic classification system (i.e., GCN, ATC, AHFS, etc.) | Incidence and prevalence use and intensity of exposure | See expanded discussion on misclassification |
| Medical records/ charts          | Drug name, dosage and regimen for prescribed and OTC agents | Binary drug exposure (incidence/prevalence) | Incomplete capture of patients’ medication history; does not capture degree of exposure (i.e., duration) |
| Outpatient medical claims        | Health Care Procedure Codes only for select medications | Binary drug exposure (incidence/prevalence) and persistency | Limited to only those medications administered in the physician’s office |

measurement—Outcome

This paper first addresses how exposure and outcomes are measured using secondary data sources, discuss instances of misclassification and ways to mitigate these biases. This is followed by a discussion on confounding in epidemiological research and methodological approaches researchers should consider to control for confounding.

Measurement—Exposure

Secondary data sources measure drug exposure with varying degrees of accuracy. Table 1 highlights these data sources, the level of measurement, and inherent limitations in using these data sources for drug exposure.

Considered the most accurate and most commonly used measure of drug exposure is outpatient prescription claims. Prescription claims data provide a wealth of information on drug exposure including date of service, dispensing pharmacy, drug name, quantity, dose, and duration (days supply), and are considered by many to be the gold standard for measuring drug exposure [16]. It should be noted that days’ supply can be unreliable for some drug classes (i.e., injectables or medications dosed on an as needed basis) and outside the United States, measures of duration may not be available.

Several options are available to identify drugs from outpatient prescription claims files. First is the National Drug Code (NDC), a 10-digit coding system established by the Food and Drug Administration (FDA) to uniquely identify drug, dosage, and package size. The FDA provides a complete listing of NDCs on its website however the drug lists can become cumbersome to manage, are time sensitive—changing with new drug entries or exits from the market. They can also be quite cumbersome to code particularly when a large number of NDCs codes are used. For example, using only the first nine digits of the NDC, which ignores package size, there are over 280 NDCs for the beta-

|AHFS, American Hospital Formulary Service; ATC, Anatomical Therapeutic Chemical; GCN, generic code number. |
Table 2  Secondary data sources for measuring outcomes

<table>
<thead>
<tr>
<th>Data source</th>
<th>Measurement</th>
<th>Outcome measured</th>
</tr>
</thead>
<tbody>
<tr>
<td>Medical records</td>
<td>Manual or automated (electronic medical records) extrapolation of diagnoses, procedures and treatments, biomarkers and other laboratory data</td>
<td>Used alone or with other data sources to identify disease progression, surrogate, or final end points</td>
</tr>
<tr>
<td>Outpatient medical claims</td>
<td>ICD-9 or ICD-10-CM, OXMIS. This paper first addresses how exposure and outcomes are measured using secondary data sources, discuss instances of misclassification and ways to mitigate these biases. This is followed by a discussion on confounding in epidemiological research and methodological approaches researchers should consider to control for confounding, CPT-4, OPCS-4, laboratory testing, diagnostic tests</td>
<td>Used alone or with other data sources to identify disease progression, surrogate or final end points</td>
</tr>
<tr>
<td>Eligibility files</td>
<td>Death&lt;sup&gt;a&lt;/sup&gt;</td>
<td></td>
</tr>
<tr>
<td>Inpatient medical claims</td>
<td>ICD-9 or ICD-10-CM, OXMIS, The Read Codes, CPT-4, OPCS-4, laboratory testing, diagnostic tests, discharge status</td>
<td>Used alone or with other data sources to identify disease progression, events or final end points</td>
</tr>
</tbody>
</table>

<sup>a</sup>May not be documented as such in all cases.


blocker Atenolol. To simplify drug identification, researchers can purchase a therapeutic classification system such as the American Hospital Formulary Service Pharmacologic-Therapeutic Classification, Red Book, Anatomical Therapeutic Chemical classification system or Medi-Span’s Generic Product Identifier. These systems link NDC to drug classes, which allow for more manageable coding of unique drugs or therapy class.

Medical records can be another data source to identify drug exposure recording whether the physician prescribed medication for the patient, the dose, and intended regimen. However, medical records do not record whether the patient obtained the medication from the pharmacy, or typically the degree of exposure (i.e., compliance). Additionally, the medical record (either inpatient or outpatient) does not record all prescribed medications taken by patients and is generally not considered a valid source for identifying drug exposure. However, medical records may be considered as a source for capturing over-the-counter (OTC) agents, typically not covered or captured in the prescription claim record. In the United States, it should be noted that the FDA will not accept e-medical records as a source for measuring drug exposure.

Drug exposure can also be measured using outpatient medical claims for a limited number of medications dispensed and administered in the physicians’ office. In the United States, these are captured using Health Care Procedure Codes (HCPCS). However, drug use identified from HCPCS codes do not indicate dosage and are not immediately assigned to newer agents. Additionally, medical billers often use miscellaneous J-codes when billing for medications administered in physician offices, which does not allow for accurate identification of the drug administered.

Another challenge faced by researchers in measuring exposure is accounting for switching in the assignment to exposure groups. Switching from one drug therapy to another often occurs naturally as a result of treatment failure or systematically from changes in benefit design [17] or programmatic features such as formulary status changes. Researchers should establish criteria a priori for treatment group assignment, be transparent in methods, and conduct sensitivity analysis to determine the impact of treatment identification on study results.

**Measurement—Outcomes**

For a given disease or condition, various measures of clinical effectiveness exist. For example, in the treatment of high cholesterol, measures of clinical effectiveness include both intermediate measures, such as the biomarker low-density lipoprotein cholesterol and cardiovascular end points including stroke or MI. Outside of the RCT environment, researchers face limitations in measuring effectiveness, particularly those that involve intermediate biomarkers or self-reported symptom scales and measures of patient functioning. Among secondary data sources, medical records are typically considered the gold standard for capturing intermediate and final outcomes (Table 2). Other secondary data sources, although providing a wealth of information on treatment patterns and medical events are more limited in measuring effectiveness. Administrative claims data can identify final end points such as fractures, stroke, or MI but are limited to proxy measures at best in the measurement of intermediary outcomes. Using a combination of diagnostic, procedure, or facility codes, researchers are beginning to develop proxy measures of intermediary outcomes with some success. For example, a study examining disease severity for chronic obstructive pulmonary disease used diagnostic and inpatient hospital stays to classify severe or moderate COPD and found moderate accuracy to medical charts [18].

There is growing use of laboratory results data linked to administrative claims data to measure intermediate outcomes. However, these data are as yet to be made available on a large scale in the United States.

**Classification Bias**

Systematic and random errors can occur in measuring both exposure and outcome resulting in the violation of internal validity. This error is termed classification bias—identifying subjects as being exposed to drug when they are not or not exposed when they are. Classification bias is further categorized as differential or nondifferential and unidirectional or bidirectional. Nondifferential misclassification occurs when the likelihood of misclassification is the same across the exposed or outcome groups. For example, exposure misclassification for a low-cost medication using prescription claims data would be equally likely regardless of outcome. However, differential misclassification is present when the likelihood of misclassification is different between exposed or outcome groups. An example of differential misclassification for drug exposure is when those who are exposed have a lower likelihood of outcome misclassification because to receive medication they have to enter the health-care system, which increases their likelihood of recording a diagnosis. Those not exposed are much more likely to be misclassified as not having the disease, which is an artifact of not entering the health care system. Unidirectional misclassification occurs when the
direction of the misclassification is in the same direction. Bidirectional misclassification occurs when the likelihood of misclassification is in both directions—there is a probability that cases appear as controls and controls appear as cases. For a more complete discussion see Hartzema and Perfetto [19]. As a researcher, one should consider and state the direction of potential sources of misclassification and how that could influence the rejection of the null hypothesis [7].

An important data element influencing classification bias of both drug exposure and outcomes when using secondary data sources is member eligibility. In the United States, many administrative datasets are linked to employment and natural transitions in the labor market can influence classification bias. If eligibility is not accounted for in the measure of medication compliance, for example, those not continuously eligible may be incorrectly classified as noncompliant when in fact the lack of drug exposure was caused by the loss of eligibility. Statistically controlling for length of eligibility or limiting to continuously eligible in these instances may be most appropriate. Lack of appropriate time for follow-up because of drops in eligibility is also a concern for outcomes misclassification if member follow-up does not allow for capture of the clinical event. Eligibility must be controlled for and lack of this information precludes comparative effectiveness research.

**Drug Exposure Misclassification**

Many factors can lead to misclassification with respect to drug exposure. With outpatient prescription claims, a greater number of opportunities for misclassification in the direction of not exposed exist given the multiple channels by which members can receive their medications outside of the reimbursement arrangements of third-party payers. Other means for obtaining prescription drugs that would preclude claims capture include physician samples, patient assistance programs (PAP), paying out of pocket, inpatient hospital stays, taking a medication belonging to someone else, secondary insurance coverage, or fraudulent behavior. The likelihood of this misclassification can be influenced by patient demographics and plan design. For example, the elderly and lower-income patients or those facing higher out-of-pocket payments may be more likely to participate in PAP programs or obtain samples from their physician, leading to systematic misclassification.

Various trend and utilization management programs can also lead to misclassification. Programmatic features including prior authorization policies, caps, or maximum limits on coverage, and pharmaceutical step therapy programs can influence not only the measure of exposure but assignment to exposure category. This information, although not always readily available to researchers, would represent a major limitation if not documented. More recently, the proliferation in the United States of no-cost or low-cost generic programs offered by retail chain pharmacies is increasing the likelihood of misclassification since these claims are not captured by the health plan. This could lead to bias depending upon the drug comparators, study sample, or geographic region given that market penetration of these programs differ by region.

Differences in formulations, or the list of covered drugs, can lead to misclassification. Systematic errors in exposure classification can occur when the treatments being compared have different formulary status or are on different tiers. If drug A is a second tier product being compared with drug B, which is a third tier product where members pay a higher copayment, differential classification bias could result, assuming higher copayments lead to lower compliance, which could impact outcomes. Methods to address these issues are covered in later sections.

Additionally, for administrative claims data, classification bias is present when measuring exposure for OTC medications or medications with limits or coverage exclusions (medications used to treat cosmetic indications). For ambulatory comparative effectiveness analysis, hospital stays (or other inpatient stays) must be accounted for in the statistical analysis [20]. However, the random bias that occurs when patients use other channels to receive medication can only be addressed as a potential study limitation.

The level of exposure misclassification can also be influenced by the study design. One important choice in the design of database studies is the time-window during which patients are considered “exposed.” This will impact misclassification of not only exposure but also outcome measurement. A study that is based on prescription information can use, for example, a 3-month time-period following each prescription in order to assess the outcome and estimate the risk of the outcome during this time-period. Although this is not always recognized, the choice of this exposure time-window is of major importance. Since misclassification of the relevant exposure time will lead to a nondifferential bias toward the null, the choice of the exposure time-window should not be based on the actual drug intake, but rather on the time-period during which the medication may cause the outcome and the duration of the pathogenic process [21,22].

As an example, a study of the effects of a medication on the risk of malignancies may suffer from a major exposure misclassification if the exposure time window would be based on the time period of drug intake and the study would include many short-term users. On the other hand, a study of allergic reactions would also suffer from exposure misclassification if the exposure time window goes beyond what is considered clinically relevant. Different approaches to improve the characterization of the exposure time window include efforts to validate the relationship or sensitivity analysis, repeating the analysis with different exposure time windows.

The focus of exposure misclassification has been in the direction of not exposed. However, it should be noted that the direction of misclassification can also be toward exposure. It cannot be assumed that presence of a claim indicates that the patient actually took the medication. For example, patients may obtain a medication for antibiotic or pain therapy, and take only if symptoms appear increasing the likelihood for misclassification toward exposed.

**Outcome Misclassification**

Several factors can lead to misclassification of diagnostic or procedure codes including plan payment systems, diagnoses, and the specificity of coding in the database [1,7]. Reimbursement systems based upon capitated payment arrangements where providers are less incentivized to submit claims documenting care compared with fee for service payment arrangements are more prone to classification bias. Under capitated payment systems, researchers should proceed with caution and attempt to validate claims data with external data sources (i.e., medical chart review).

Misclassification has been shown to vary by disease state with hypertension and diabetes having the highest rates of sensitivity (60.6 and 62.6, respectively) and chronic liver disease, peptic ulcer disease, or acute myocardial infarction with some of the lowest levels of sensitivity (27.6, 27.6, and 25.4, respectively) [8]. This variability can be caused by multiple factors including clinical ambiguity in diagnoses, stigma associated with the diagnoses or coding used for rule out diagnostic procedures. Using a longer look back period and requirements of at least two diagnoses or...
inclusion of medical treatment can increase specificity [23]. Also being explored is the use of algorithms using drug, medical and patient demographic information to increase the accuracy of diagnostic information [24]. Systematic error in classification of outcomes can occur if the researcher fails to take into account changes in codes resulting from updates or brought about by the transition from the International Classification of Diseases (ICD)-9 to ICD-10 coding systems.

When considering various approaches, researchers should seek out definitions that have been validated with external sources, such as chart review. When there are several approaches without a clear empirical direction, sensitivity analyses should be explored to understand the implications of the various definitions on the results. For example, MI may be defined using two diagnoses or one diagnosis and a hospital stay, which will alter the incidence of MI detected in the study. When measuring comorbidity ideally one should select a measure that has been validated in a population most similar to the study and for the outcome under investigation.

**Recommendations**

1. State the direction of potential sources of misclassification and how that could influence the acceptance or rejection of the null hypothesis.
2. Eligibility must be controlled for and lack of this information precludes comparative effectiveness research.
3. For ambulatory comparative effectiveness analysis, hospital stays (or other inpatient stays) must be accounted for in the statistical analysis.
4. The choice of the exposure time-window should not be based on the actual drug intake, but rather on the time-period during which the medication may cause the outcome and the duration of the pathogenic process.
5. Definitions that have been validated with external sources, such as chart review, should be used as the primary method in defining the measure. When there are several approaches without a clear empirical direction, sensitivity analyses should be explored to understand the implications of the various definitions on the results.
6. When measuring comorbidity, select a measure that has been validated in a population most similar to the study and for the outcome under investigation.

**Confounding and Causal Graphs**

Issues surrounding misclassification is not the only bias that researchers are faced with when using retrospective secondary data sources. Confounding also comes into play. Confounding is classically defined as a bias that distorts the exposure-disease or exposure-outcome relationship [25]. Frequently used definitions of confounding and standard textbook methods to control for confounding state that a confounder is an independent (causal) risk factor for the outcome of interest that is associated with the exposure of interest in the population, but that is not an intermediate step in the causal pathway between the exposure and the outcome [26,27].

**Confounding by Indication for Treatment**

A common and pernicious problem endemic to pharmacoepidemiologic studies is confounding by indication of treatment. For example, when the choice of therapy is affected by the severity of illness, and physicians prescribe one therapy over another depending on the severity and the perceived effectiveness of one drug compared with another for patients with differing severity levels, then confounding by indication for treatment occurs (assuming that the severity of disease also is a risk factor for the outcome of interest). In this case, apparent (i.e., estimated) treatment effects are confounded, that is, they are not causal but they may actually be caused by the severity of illness that led to patients being prescribed a given treatment.

**Measured versus Unmeasured Confounding**

Confounders may be measured or unmeasured. Secondary databases of a variety of sources may contain a wide and rich variety of information that can be used to measure an array of potentially confounding factors. However, even the most detailed and complete data sources may fail to include information on potential confounding factors, and these remain unmeasured and hence uncontrolled in a given study leading to residual confounding. Methods to address both measured and unmeasured (residual) confounding factors have been developed to address these concerns and will be detailed in the third series of the task force’s report.

**Time-Dependent Confounding**

The more complicated (but probably not less common) case of time-dependent confounding refers to variables that simultaneously act as confounders and intermediate steps, that is, confounders and risk factors of interest mutually affect each other. Confounding by indication, may take the form of time dependent confounding. An example is the effect of aspirin use (treatment) on risk of MI and cardiac death (outcome). Prior MI is a confounder for the effect of aspirin use on risk of cardiac death, because prior MI is a cause of (subsequent) aspirin use, and is also a causal risk factor for (subsequent) cardiac death. However, (prior) aspirin use also causally prevents prior MI. Therefore, prior MI simultaneously acts as confounder (causing aspirin use) and intermediate step (being affected by aspirin use), and hence is a time-dependent confounder affected by previous treatment.

Traditional textbook techniques to control for time-independent confounding include restriction, stratification, matching, or multivariate regression analysis. However, these methods have been criticized for being inadequate to control for time-dependent confounding. Other methods such as g-computation, marginal structural models, or structural nested models have been suggested as approaches to this problem [28,29].

These analytic methods require repeated measurements of the treatment of interest, potential confounders and the outcome. With the proliferation of longitudinal data sources, where patients are followed up over years of exposure to medical therapies, these analytic methods should be applied.

**Causal Graphs**

To address the issue of confounding in retrospective databases and to be able to do a proper causal analysis, we must answer these questions: 1) which a priori assumptions can we make about the causal relationships between the variables of an epidemiological study?; 2) under these assumptions, are the observed data sufficient to control for confounding?; and 3) what methods are appropriate to control for confounding?

Causal graphs can guide us in answering these questions [30]. Directed acyclic graphs (DAGs) are causal graphs that can be used to understand and explicitly state causal a priori assumptions about the underlying biological mechanisms [31,32]. DAGs consist of a set of nodes and directed links (arrows) that connect
certain pairs of nodes (see Fig. 1). For our purposes, nodes represent variables and arrows denote causal relationships. A set of precise graphical rules for DAGs has been developed, which allows us to determine whether an unbiased effect is estimable from the observed data, which variables must be adjusted for in the analysis, and which statistical methods can be used to obtain unbiased causal effects. Part of these rules is a new and graphically oriented definition of confounding (i.e., the “backdoor criterion”).

Furthermore, DAGs offer a readily accessible approach to understanding complex statistical issues including the fallibility of estimating direct effects (i.e., controlling for intermediate steps), the rationale for instrumental variables, and controlling for compliance in randomized clinical trials (when both “intention-to-treat” [ITT] and “per protocol” analyses can fail to yield the true causal intervention effect). In conclusion, DAGs are a valuable and comprehensive tool that offers epidemiologists and outcomes researchers better insight into confounding and the causal interpretation of their model results.

Another example of time-dependent confounding by treatment is antiviral treatment of HIV infection, where treatment or dose may depend on CD4-count and this dependency may continue over the course of the disease [33].

In the conduct of nonrandomized comparative effectiveness studies, it is strongly recommended to: 1) define the DAG for the base-case analysis before actually starting the analysis; 2) report the DAG for the base-case analysis; and 3) if sensitivity analyses are performed for different assumptions regarding the confounding structure, report the additional DAGs representing the assumptions of the respective sensitivity analyses.

One additional approach for assessing the likelihood that residual confounding may be responsible for an observed treatment effect would be to conduct a sensitivity analysis exploring the effect of the treatments on outcomes that should not be influenced by the treatment in addition to the primary end points. For example, in an analysis to compare different prostate cancer treatments, one could explore recurrent cancer related mortality or all cause mortality as primary end points and additionally compare the treatments on outcomes unrelated to the disease or the treatments such as pulmonary or diabetes-related mortality. If a prostate cancer treatment was found to have a beneficial effect on cancer mortality but no impact on diabetes related mortality, there is greater confidence linking the treatment to the primary end point, however, if a beneficial effect is also found for diabetes mortality, it is less clear if the treatments are influencing the primary cancer benefit or if the therapy is more likely to be prescribed for “healthier” patients that could not be controlled in the analysis [34].

**Recommendations:**

1. Define the DAG for the base-case analysis before actually starting the analysis.
2. Report the DAG for the base-case analysis.
3. If sensitivity analyses are performed for different assumptions regarding the confounding structure, report the additional DAGs representing the assumptions of the respective sensitivity analyses.

**Restriction—Inclusion and Exclusion Criteria**

Although a variety of systematic errors may bias nonexperimental research [35] confounding bias is of particular concern in epidemiologic studies of drug effects [36].

Restricting study cohorts to patients who are homogeneous regarding their indication for the study drug will lead to more balance of patient predictors of the study outcome among exposure groups and thus will reduce confounding but not necessarily eliminate confounding, particularly when there are variables that influence prescribing decisions that are not available in the data. Restricting study cohorts can also increase the likelihood that all included subjects will have a similar response to therapy and therefore reduce the likelihood of effect modification. RCTs commonly restrict their study population to patients with a presumed indication for the study drug and then randomly allocate the actual treatment.

There are many different approaches to restriction in specific studies [37] and it is therefore difficult to provide generic advice that fits specific study designs. However, several guiding principles can be identified that should be considered in a nonrandomized database study on effectiveness and safety of medical interventions [38].

**Exclude Patients with a History of the Study Outcome?**

The decision whether to exclude patients with a history of the study outcome is largely based on the study questions and the chronicity of the outcome under study. Some guiding principles may include:

1. Patients with a history of occasionally or frequently occurring events that are restored to a normal health level with or without treatment may not be candidates for exclusion if their health status has reached a normal level before cohort entry. Examples for such conditions are uncomplicated viral or bacterial infections.
2. Patients with a history of conditions that are markers for an underlying chronic condition will have an increased risk for the study outcome and at the same time may be more likely to take a study medication causing confounding. Examples for such conditions include hip fractures in elderly patients, which are markers for frail health and/or osteoporosis, which put the patient at increased risk for a future event. Similarly, a previous MI is a strong risk factor for future cardiac events. If these conditions are strong risk factors for future events and therefore potentially strong confounders it may be better to exclude these patients from the analysis rather than adjusting for them.

**Study Incident Medication Users Only?**

Usually, an epidemiologic database study is implemented by defining a study period for which subjects are considered. Let us
consider a cohort study of statin use and some health outcome. The most basic cohort definition would be to identify subjects who used a statin at any point during the study period, assigning the date of first observed statin use during that period as an index date. On each statin user’s index date, we sampled a subject who had not used a statin as of that date, i.e., a nonuser, and assigned him or her the same index date.

The population of statin users described earlier consists of a mix of incident drug users, i.e., those starting on a statin, and prevalent users, i.e., those taking a statin for some time.

**Mixed Prevalent and Incident User Cohorts**

Studying mixed prevalent and incident user cohorts will lead to under-ascertainment of early events. Depending on the average duration (chronicity) of use, such cohorts may be composed predominantly of prevalent users and few new users (e.g., statins). The estimated average treatment effect will therefore underemphasize effects related to drug initiation and will overemphasize effects of long-term use [39].

Prevalent users of a drug have by definition persisted in their drug use, similar to the concept of survivor cohorts in chronic disease epidemiology [40]. Being persistent or adherent is a characteristic found more frequently in patients who tolerate the drug well and who perceive some therapeutic benefit. Adherence also characterizes patients with higher educational status and health-seeking behavior particularly if the study drug is treating an asymptomatic condition like statins treating hyperlipidemia, characteristics that are difficult to assess in claims data, and may lead to healthier user bias [41–43].

The duration of use among prevalent users can differ by drug exposure; duration thus may cause bias if it remains unadjusted. Such a scenario is likely when newly marketed drugs are compared with competitors that have been available longer. In database studies, duration of prior use can only be assessed by tracing back a continuous string of prescriptions to the initial prescription.

In studying prevalent users, investigators can assess patient characteristics only after the initial exposure; thus the drug under study may affect those characteristics. Adjusting for such factors that are on the causal pathway of the drug’s action will lead to an underestimation of the drug effects.

**“New User Design.”** One begins an incident user design by identifying all patients in a defined population who start a course of treatment with the study medication. Exposed person-time begins at the start of treatment, which is identified as a dispensing of the index drug without a dispensing of the index drug during the prior year or some other fixed time interval comparable with a wash-out period commonly used in RCTs. The advantage of the so-called “New User Design” has recently been summarized [40]. Although limiting the study population to drug initiators resembles one of several key characteristics of clinical trials, the limited number of incident users requires large source populations like health care utilization databases from which new starters can be identified efficiently. For some patients it may not be the first time they take the study drug, i.e., they are not really naive to the drug. Patients who know from earlier treatment courses that they tolerate the drug and that it is effective for them are more likely to use the same drug again. The chance of an initiator to be a true new user can be increased by requiring longer periods without use of the study drug before the index prescription.

**What Is the Most Adequate Comparison Group?**

Choosing a comparison group is a complex and sometimes subjective issue. The ideal comparison should comprise patients with identical distributions of measured and unmeasured risk factors of the study outcome.

**Patients with the same treatment indication: “Alternative Drug Users.”** Selecting comparison drugs that have the same perceived medical indication for head-to-head comparisons of active drugs will reduce confounding by selecting patients with the same indication (e.g., indication for using celecoxib vs. rofecoxib). Although one can rarely measure the indication directly—in the statin example we would need laboratory values of serum lipid levels that are not routinely available in claims data—we infer the indication by the initiation of a treatment specific to the indication. When studying unintended benefits or risks of drugs, such as exploring the potential cancer preventive properties of nonsteroidal anti-inflammatory drugs (NSAIDs), confounding by indication may be less problematic as physicians are unlikely to prescribe therapies based on a patient’s risk of developing the un-intended outcome (e.g., cancer) assuming the disease(s) for which the therapy are indicated are unrelated to the outcome. However, new competitors within a class are often marketed for better efficacy, slightly expanded indications, or better safety (cyclo-oxygenase-2 inhibitors [coxibs] vs. nonselective NSAIDs) influencing physicians’ prescribing decisions [44]. In this way, new opportunities of confounding by indication can arise.

**“Nonusers.”** In some cases there either is no comparator drug with a reasonably close indication to the study drug or a class effect is suspected such that the entire class is to be tested, requiring comparison subjects who did not use any drug of this class. The most obvious choice may be to identify study subjects who do not use the study drug and then to pick a random date as the index date, possibly matched by time to the index date of the first prescription among active drug users.

Obviously, patients on therapy most likely have a medical indication; by contrast a large proportion of nonusers have no medical indication, i.e., patients initiating statin therapy are more likely to have elevated lipid levels and therefore increased cardiac risks. However, nonusers as defined earlier may differ substantially from users of the index drug for both measured and unmeasured characteristics, even beyond the indication for the index drug.

As a case in point: Although initiators of a new drug have (presumably) been evaluated by a physician just before that prescription, nonusers may not have seen a physician for a while and, in fact, may have less contact with the health care system in general. Differential underrecording of health conditions in the nonuser comparison group makes members of the comparison group appear healthier than they really are and may lead to an overestimation of treatment effects.

Groups will be more comparable regarding access to health care, including health-seeking behavior and disease surveillance, when choosing comparison patients who also had contact with the health-care system in the form of a drug dispensing. Like patients starting the study drug, such patients have just been evaluated by a physician before the initial prescription. Adequate comparison groups for new statin initiators could, for example, be initiators of topical glaucoma drugs or thyroid hormone substitution. Both these classes of pharmaceuticals are unrelated to lowering serum lipid levels and are used for preventing the progression of an initially asymptomatic condition.

**Excluding Patients with Contraindications?**

In studies of the effectiveness of drugs it is questionable whether we want to include patients who have a clear contraindication to
the study drug. Such patients will be few and their experience will be unusual. Prudence dictates, therefore, excluding patients with contraindications or absolute indications, resulting in a situation similar to the therapeutic equipoise required for RCTs [45].

Because reliably identifying contraindications in claims data is unlikely, identifying them empirically is more promising. Propensity scores, a common mechanism for doing this, estimate each patient’s probability of treatment given all measured covariates. These propensity scores follow a distribution between 0 and 1 that differ between actual users and nonusers. On the low end of the propensity score distributions indicating a low propensity for receiving treatment, there will be a range that is only populated by actual nonusers because all users have a higher propensity scores. Such nonusers are likely to have a contraindication for the study medication because no subject with such a low propensity score has actually received treatment. These patients should be deleted from the study population. Analogously, such trimming can be considered at the upper end of the propensity score, excluding patients who will always be treated.

Excluding Patients with Very Low Adherence?

Patients dropping out of RCTs for reasons related to the study drug may cause bias. Noninformative dropout causes bias towards the null in ITT analyses. The medical profession and regulatory agencies accept such a bias because its direction is known and trial results are considered conservative regarding the drug’s effectiveness. Discontinuation of treatment may also be associated with study outcomes. Obviously, reasons are lack of perceived treatment effect or intolerance. Both factors may lead to early stopping but can cause discontinuation at any time later during the course of treatment. Another factor that may lead to discontinuation of medications, particularly those used to treat asymptomatic conditions, is overall frail health status that requires multiple medications to treat the more symptomatic conditions. For example, cancer patients may discontinue statins in order to reduce polypharmacy in favor of more urgently needed drugs [42].

RCTs try to minimize bias from nonadherence by frequently reminding patients and by run-in phases before randomization aimed to identify and exclude nonadherent patients. In routine care, adherence to drugs is unfortunately substantially lower than in RCTs. Studies have shown, that for statin medications, only 50% to 60% of elderly patients refill their prescriptions after 6 months [46].

Starting follow-up after the third fill of a chronic medication will exclude patients who are least adherent. Unlike RCTs in which run-in phases are often done with placebo [47] patients in routine care experience their first exposure to a new drug and may discontinue use because of a lack of effectiveness or intolerance during what may be the most vulnerable period for some medication-outcome relations. As long as that proportion is small and most patients discontinue for reasons not directly related to the study drug(s), this issue should be minor.

Generalizability

To guide our thinking about generalizability, it is useful to specify the patient to whom we wish to generalize our results. From a patient and physician perspective, the most relevant and frequently asked question is, “What is the effectiveness and safety of a particular drug that I am about to start and continue to use, compared with not starting therapy, or compared with starting an alternative drug?” From this viewpoint, restricting studies to initiators of drug therapy does not limit generalizability. Instead, it avoids under-representation of treatment effects that occur shortly after initiation. Patients with known contraindications (or their clinicians) would usually not have to confront this hypothetical question because prescribing the drug in the first place would contravene current medical knowledge. Therefore, excluding patients with known contraindications places little limits on generalizability.

In making a prescribing decision, physicians must assume that patients will take a drug as directed. If clinicians knew beforehand that a patient would not take a prescribed medication, they would not ponder the appropriateness of the drug in the first place. Consequently, excluding patients who are non-adherent to their treatment independent of intolerance or treatment failure—will not limit generalizability to the question raised above. However, the situation is quite different if we restrict the study population by disease severity, comorbidities, polypharmacy, and other risk factors for the study outcome. Data based on such restrictions will limit physicians when making prescribing decisions concerning the excluded patient subgroups. The obvious solution to this problem is to stratify analyses according to relevant clinical subgroups, rather than restricting them out of the analysis altogether, and then testing whether treatment effects differ between groups [48]. The large size of health-care utilization databases can allow performing such subgroup analyses with substantial numbers of subjects, and represents an attractive alternative to wholesale restriction.

Conclusion

Design of nonrandomized studies of comparative effectiveness face several daunting issues, including measurement of exposure and outcome challenged by biases in classification and confounding. We identified a set of restrictions that analysts should consider in studies of the effectiveness of therapies when using large observational databases. Such restrictions will place few limits on generalizability of research finding for most clinically relevant treatment choices. Use of causal diagrams and restriction are two techniques that can improve the theoretical basis for analyzing treatment effects in study populations of more homogeneity, with reduced loss of generalizability.

Source of financial support: This work was supported in part by the Department of Veteran Affairs Health Services Research and Development grant HFP020—90. The views expressed are those of the authors and do not necessarily reflect the views of the Department of Veteran Affairs.

References


Michael L. Johnson, PhD,1 William Crown, PhD,2 Bradley C. Martin, PharmD, PhD,2 Colin R. Dormuth, MA, ScD,4 Uwe Siebert, MD, MPH, MSc, ScD5

1University of Houston, College of Pharmacy, Department of Clinical Sciences and Administration, Houston, TX, USA; Houston Center for Quality of Care and Utilization Studies, Department of Veteran Affairs, Michael E. DeBakey VA Medical Center, Houston, TX, USA; 2i3 Innovus, Waltham, MA, USA; 3Division of Pharmaceutical Evaluation and Policy, College of Pharmacy, University of Arkansas for Medical Sciences, Little Rock, AR, USA; 4Department of Anesthesiology, Pharmacology & Therapeutics, University of British Columbia; 5Pharmacoeconomics Group, Therapeutics Initiative, Vancouver, BC, Canada; 4Department of Public Health, Information Systems and Health Technology Assessment, UMIT, University of Health Sciences, Medical Informatics and Technology, Hall i.T., Austria; Institute for Technology Assessment and Department of Radiology, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA, and Department of Health Policy and Management, Harvard School of Public Health, Boston, MA, USA

ABSTRACT

Objectives: Most contemporary epidemiologic studies require complex analytical methods to adjust for bias and confounding. New methods are constantly being developed, and older more established methods are yet appropriate. Careful application of statistical analysis techniques can improve causal inference of comparative treatment effects from nonrandomized studies using secondaries databases. A Task Force was formed to offer a review of the more recent developments in statistical control of confounding.

Methods: The Task Force was commissioned and a chair was selected by the ISPOR Board of Directors in October 2007. This Report, the third in this issue of the journal, addresses methods to improve causal inference of treatment effects for nonrandomized studies.

Results: The Task Force Report recommends general analytic techniques and specific best practices where consensus is reached including: use of stratification analysis before multivariable modeling, multivariable regression including model performance and diagnostic testing, propensity scoring, instrumental variable, and structural modeling techniques including marginal structural models, where appropriate for secondary data. Sensitivity analyses and discussion of extent of residual confounding are discussed.

Conclusions: Valid findings of causal therapeutic benefits can be produced from nonrandomized studies using an array of state-of-the-art analytic techniques. Improving the quality and uniformity of these studies will improve the value to patients, physicians, and policymakers worldwide.

Keywords: causal inference, comparative effectiveness, nonrandomized studies, research methods, secondary databases.

Background to the Task Force

In September 2007, the International Society for Pharmacoeconomics and Outcomes Research (ISPOR) Health Science Policy Council recommended that the issue of establishing a Task Force to recommend Good Research Practices for Designing and Analyzing Retrospective Databases be considered by the ISPOR Board of Directors. The Council’s recommendations concerning this new Task Force were to keep an overarching view toward the need to ensure internal validity and improve causal inference from observational studies, review prior work from past and ongoing ISPOR task forces and other initiatives to establish baseline standards from which to set an agenda for work. The ISPOR Board of Directors approved the creation of the Task Force in October 2007. Task Force leadership and reviewer groups were finalized by December 2007, and the first teleconference took place in January 2008.

Address correspondence to: Michael L. Johnson, University of Houston, College of Pharmacy, Department of Clinical Sciences and Administration, Houston, TX 77030, USA. E-mail: mikejohnson@uh.edu 10.1111/j.1524-4733.2009.00602.x

Task Force members were experienced in medicine, epidemiology, biostatistics, public health, health economics, and pharmacy sciences, and were drawn from industry, academia, and as advisors to governments. The members came from the UK, Germany, Austria, Canada, and the United States.

Beginning in January 2008, the Task Force conducted monthly teleconferences to develop core assumptions and an outline before preparing a draft report. A face-to-face meeting took place in October 2008, to develop the draft, and three forums took place at the ISPOR Meetings to develop consensus for the final draft reports. The draft reports were posted on the ISPOR website in May 2009 and the task forces’ reviewer group and ISPOR general membership were invited to submit their comments for a 2 week reviewer period. In total, 38 responses were received. All comments received were posted to the ISPOR website and presented for discussion at the Task Force forum during the ISPOR 12th Annual International Meeting in May 2009. Comments and feedback from the forum and reviewer and membership responses were considered and acknowledged in the final reports. Once consensus was reached, the manuscript was submitted to Value in Health.
Introduction
We proceed from the assumption that proper statistical analysis of study data is dependent upon the research question to be answered, and the study design that led to the collection of data, in this case secondarily, to be analyzed [1,2]. We also assume that the data to be analyzed has been appropriately measured, validated, defined, and selected. Many authors have described good research practices in these fundamental areas, including other ISPOR task force groups, and we seek to build upon their work, not reproduce it [3,4]. Recognizing that new methods are constantly being developed, and older more established methods are yet appropriate, we intend to offer a review of the more recent developments in statistical control of confounding.

Stratification
Stratified analysis is a fundamental method in observational research. It involves placing data into subcategories, called strata, so that each subcategory can be observed separately. Its many uses in observational studies include standardization, control of confounding, subgroup analysis in the presence of effect-measure modification, and to address selection bias of the type that occurs in matched case control studies. When a cohort is categorized by follow-up time, stratification can also prevent bias from competing risks and losses to follow-up.

Like any analytical method, stratification has its strengths and limitations. Its strengths are that it is an intuitive and hands-on method of analysis, results are readily presented and explained, and it does not require restrictive assumptions. Its disadvantages include a potential for sparsely populated strata, which reduce precision, loss of information when continuous variables are split into arbitrarily chosen categories, and a tendency to become arduous when the number of strata is large.

Most contemporary observational studies use complex analytical methods such as multivariable regression analysis. Given the ubiquity of those methods, it is tempting to undervalue the role of stratified analysis, but to do so is a mistake because it has important uses that are not well served by other methods. Because stratified analysis can be applied “hands-on,” often in a simple spreadsheet program, it allows investigators to get closer to their data than they otherwise could by using more complex methods such as multivariable regression. Furthermore, stratified analysis in a cohort where observations are categorized by levels of the most influential covariates should provide results that are comparable to estimates from a rigorous multivariable model. If results from a stratified analysis are markedly different from estimates obtained from a regression model, then the discrepancy should serve as a warning that the investigator has possibly made a mistake. For this reason, it is important to do a stratified analysis in studies even when more complex analytical methods are finally used. It is advisable to conduct a stratified analysis prior to undertaking a more complex analysis because of its potential to provide important information on relevant covariates and how they could be optimally included in a model.

Stratified analysis can proceed by categorizing data into strata. In a spreadsheet, 10 to 20 strata should be manageable before the analysis becomes cumbersome. If the analysis requires more strata, then it may be helpful to perform the analysis in stages by first examining levels of a few variables, and then in subsets defined according to those first few variables, perform analyses on more variables. A staged approach can be time-consuming and does not lend itself easily to calculating summary or pooled estimates, but it can be useful for studying effect-measure modification.

Regression
Numerous texts are available to teach regression methodology, and we do not attempt to summarize them here, nor cite a complete list [6–10]. Regression is a powerful analytical technique that can accomplish several goals at once. When more than a few strata are formed for stratified analysis, or when more than a few potential confounding factors need to be adjusted, multiple regressions can be used to determine the unique association between the treatment and the outcome, after simultaneously adjusting for the effects of all the other independent factors included in the regression equation. It is very common to see reports of the parameter estimates, rate ratio (RR), or odds ratio (OR), and their 95% confidence limits, for a given variable after adjustment for a long list of covariates. It is becoming less common to actually see the full regression model adjusted for all the covariates, but failure to present the full model may lead to concerns of a “black box” analysis that has other pitfalls [11,12]. Another very important use of regression is to use the regression equation to predict study outcomes in other patients. This is the primary use of multiple logistic regression when used for propensity scoring, which will be discussed later. Present the final regression model, not only the adjusted treatment effects. If journal or other publications limit the amount of information that can be presented, the complete regression should be made available to reviewers and readers in an appendix.

Variable Selection
One of the critical steps in estimating treatment effects in the observational framework is to adequately assess all the potential confounding variables that can influence treatment selection or the outcome. In order to capture all of the potentially confounding variables and any suspected effect modification or interactions, a thorough literature review should be conducted to identify measures that influence treatment selection and outcome measurement, and a table should be created detailing the expected associations. The analyst should identify those measures available in the data, or good proxies for them, and include them in the regression model irrespective of statistical significance at traditional significance levels. When using administrative data sources with limited clinical information, there are often instances when the analyst will not have access to meaningful
Model Selection
The form of the dependent variable generally drives the choice of a regression model. Continuously distributed variables are generally analyzed with ordinary least squares (OLS) regression, while dichotomous or binary outcomes (yes/no; either/or; dead/alive) can be modeled with logistic regression. In practice, common statistical software programs analyze both types of outcomes using maximum likelihood estimation and assuming the appropriate error distribution (normal, logistic, etc.). Logistic regression has become almost ubiquitous in the medical literature in the last 20 years, coinciding with advances in computational capacity and familiarity with the method. Linear and logistic regressions fall under a broader category of models known as generalized linear models (GLM). GLMs also include models with functional forms other than linear or log-linear to describe the relationship between the independent and dependent variables. Some commonly used link functions and error distributions are shown in Table 1.

Another valuable use of GLM models is their ability to incorporate different specifications of covariation structures when the assumption of the independence between observations is violated. Longitudinal analyses using data sets in which multiple measurements are taken on the same subject over time are common in comparative effectiveness studies. In such analyses, the observations are not independent, and any correlation must be accounted for to obtain valid and precise estimates of effects [15].

Testing Model Assumptions
There are many statistical assumptions that underlie these regression techniques. For the linear and logistic models, the assumptions of normality and linear association are possibly the most important. Fortunately, these models are very robust to the assumption of normality, that is, the outcome variable has to be very non-normal (such as costs) to severely threaten parameter estimates. A more common problem is the use of continuous measures modeled as continuous variables without checking the assumption of linear association. For example, age is usually thought of as a continuous variable, ranging from maybe 18 to 90 or so in many studies of adults with medical conditions. Age in years is actually a categorical variable, with 72 different levels, in this case. It is essential to check the assumption of a linear relationship between continuous independent variables and study outcomes; if the independent variable does not have a linear relation, then nonlinear forms, such as categories should be used in modeling. Thus, it is essential to check and see if the association of age with the study outcome increases/decreases in a relatively constant amount as age increases/decreases, or the study results for age, and the control or adjustment for confounding by age, may not be valid. For example, if relatively few subjects are very old and also have a certain unfavorable response to a drug, and most of the other patients are younger and generally do not have unfavorable responses, a regression model may erroneously show that increased age is a risk factor.

Table 1  Commonly used link functions and error distributions for outcomes with different types of data

<table>
<thead>
<tr>
<th>Examples of outcomes</th>
<th>Data type</th>
<th>Link function</th>
<th>Error distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trends in drug utilization or costs</td>
<td>Continuous</td>
<td>Identity</td>
<td>Gamma</td>
</tr>
<tr>
<td>Predictors of treatment choice, death</td>
<td>Binary</td>
<td>Logistic</td>
<td>Binomial</td>
</tr>
<tr>
<td>Myocardial infarction, stroke, hospitalizations</td>
<td>Count</td>
<td>Log</td>
<td>Poisson</td>
</tr>
</tbody>
</table>

*Assuming Gamma-distributed errors does not require log transformation of utilization and thus increases interpretability of results [13, 14].
†Depending on the skewness of data, we may adjust for over dispersion using the scale parameter.
because it is assuming a linear relation with age that is not actually there. In this case, the variable should be modeled as a categorical variable. Another serious assumption that should be tested is the proportional hazards assumption for Cox regression. The Cox model is based on the assumption that the hazard rate for a given treatment group can change over time, but that the ratio of hazard rates for two groups is proportional. In other words, patients in two treatment groups may have different hazards over time, but their relative risks should differ by a more or less constant amount. It is not a difficult assumption to test, but if violated, study results using this technique become questionable. Thus, the proportional hazards assumption for treatment exposure should always be tested before conducting Cox proportional hazards regression, and if this assumption is violated, alternative techniques such as time-varying measures of exposure in extended Cox models should be implemented.

Performance Measurement

Analyses using regression models should report model performance. For example, OLS regression models should report the coefficient of determination ($R^2$). This is so that the reader can determine if this regression equation has made any realistic explanation of the total variance in the data. In a large database study, the $R^2$ could be very small, but parameter estimates may be unbiased. Although valid, it may be questionable as to the value of intervening on such variables. In logistic regression, the c-statistic or area under the receiver operating characteristic curve (ROC) is a standard output in many statistical software packages to assess the ability of model to distinguish subjects who have the event from those who do not. Qualitative assessment of the area under the ROC curve have been given by Hosmer and Lemeshow 1999 [9]. Performance measures ($R^2$, area under ROC curve) should be reported and a qualitative assessment of these measures should be discussed regarding the explanation of variance or discrimination of the model in predicting outcome. If the emphasis is on prediction, one might care more about model performance than if the emphasis is on multivariable control of several confounding factors.

Diagnostics

All statistical software packages also provide an array of measures to determine if the regression model has been properly produced. Plots of the observed values to predicted values, or the difference between observed and predicted (residual values) can be easily produced to examine for outlier observations that may be exerting strong influence on parameter estimates. With powerful computers in today’s world, on the desktop, or in one’s lap at a sidewalk café, it is very easy to rely on the computer to crank out all kinds of fascinating parameter estimates, RRs, ORs, confidence intervals and $P$-values that may be questionable if not actual garbage. It is very important for the analyst to check the results of regression modeling by at least checking plots of residuals for unexpected patterns. Regression diagnostics including goodness of fit should be conducted and reported.

Missing Data

One of the many challenges that face researchers in the analysis of observational data is that of missing data. In its most extreme form, observations may be completely missing for an important analytic variable (e.g., Hamilton depression scores in a medical claims analysis). However, the issue of missing data is more commonly that of missing a value for one or more variables across different observations. In multivariate analyses such as regression models, most software packages simply drop observations if they are missing any values for a variable included in the model. As a result, highly scattered missing observations across a number of variables can lead to a substantial loss in sample size even though the degree of “missingness” might be small for any particular variable.

The appropriate approach for addressing missing data depends upon its form. The simplest approach is to substitute the mean value for each missing observation using the observed values for the variable. A slightly more sophisticated version of this approach is to substitute the predicted value from a regression model. However, in both instances, these approaches substitute the same value for all patients (or all patients with similar characteristics). As a result, these methods reduce the variability in the data. (This may not be a particularly serious problem if the pattern of missingness seems to be random and does not impact large numbers of observations for a given variable.) More sophisticated methods are available that preserve variation in the data. These range from hot deck imputation methods to multiple imputation [16,17]. The extent of missing data and the approach to handle it should always be reported.

Recommendations

- Conduct a stratified analysis prior to undertaking a more complex analysis because of its potential to provide important information on relevant covariates and how they could be optimally included in a model.
- Present the final regression model, not only the adjusted odds ratio but also presenting the complete regression should be made available to reviewers and readers in an appendix.
- Conduct a thorough literature review to identify all potential confounding factors that influence treatment selection and outcome. Create a table detailing the expected associations.
- When known potential confounders cannot be included in the model, the analyst should acknowledge their missingness as a limitation and describe the anticipated directionality of the bias.
- When in doubt as to the true correlation structure, then an exchangeable matrix should be used.
- Check the assumption of a linear relationship between continuous independent variables and study outcomes; if the independent variable does not have a linear relation, then nonlinear forms, such as categories should be used in modeling.
- The proportional hazards assumption for treatment exposure should always be tested before conducting Cox proportional hazards regression; if this assumption is violated, alternative techniques such as time-varying measures of exposure in extended Cox models should be implemented.
- Performance measures ($R^2$, area under ROC curve) should be reported and a qualitative assessment of these measures should be discussed regarding the explanation of variance or discrimination of the model in predicting outcome.
- Regression diagnostics including goodness of fit should be conducted and reported.
- The extent of missing data and the approach to handle it should always be reported.

Propensity Score Analysis

The propensity score is an increasingly popular technique to address issues of selection bias, confounding by indication or
endogeneity commonly encountered in observational studies estimating treatment effects. In 2007, a PubMed search identified 189 human subjects “propensity score” articles compared to just 2 retrieved in 1997. Propensity scoring techniques are now being used in observational studies to address a wide range of economic, clinical, epidemiologic, and health services research topics. The appeal of the propensity scoring techniques lies in an intuitive tractable approach to balance potential confounding variables across treatment and comparison groups. When propensity scores are utilized with a matching technique, the standard “Table 1” that compares baseline characteristics of treated and untreated subjects often resembles those obtained from randomized clinical trials where measured covariates are nearly equally balanced across comparison groups [18,19]. This transparent balancing of confounders facilitates confidence in interpreting the results compared to other statistical modeling approaches; however, unlike randomization, balance between unmeasured or unmeasurable factors cannot be assumed.

The propensity score is defined as the conditional probability of being treated given an individual’s covariates [20,21]. The more formal definition offered by Rosenbaum and Rubin for the propensity score for subject \( i = 1, \ldots, N \) is the conditional probability of assignment to a treatment \( (Z_i = 1) \) versus comparison \( (Z_i = 0) \) given observed covariates, \( x_i \):

\[
E(x_i) = \Pr(Z_i = 1 | X = x_i)
\]

The underlying approach to propensity scoring uses observed covariates \( X \) to derive a “balancing score” \( b(X) \) such that the conditional distribution of \( X \) given \( b(X) \) is the same for treated \( (Z = 1) \) and control \( (Z = 0) \) [20,21]. When propensity scores are used in matching, stratification, or regression, treatment effects are unbiased when treatment assignment is strongly ignorable [21]. Treatment assignment is strongly ignorable if treatment groups, \( Z \), and the outcome (dependent) variable are conditionally independent given the covariates, \( X \). This independence assumption will not hold in situations where there are variables or at least good proxy measures not included as propensity score covariates that are correlated with outcome events and treatment selection. These situations are fundamentally the same issue associated with omitted variable bias encountered in more classical regression-based methods. The most common approach to estimate propensity scores are logistic regression models; however, other approaches such as probit models, discriminant analysis, classification and regression trees, or neural networks are possible [22,23]. A tutorial by D’Agostino provides a good description of how to calculate propensity scores including sample SAS code [21].

Once a propensity score has been developed, there are three main applications of using the propensity score: matching, stratification, and regression. Matching on the propensity score takes several approaches, but all are centered on finding the nearest match of a treated (exposed) individual to a comparison subject(s) based on the scalar propensity score [24]. Onur Basar described and empirically compared seven matching techniques (stratified matching, nearest neighbor, 2 to 1 nearest neighbor, radius matching, kernel matching, Mahalanobis metric matching with and without calipers) and found Mahalanobis metric matching with calipers to produce more balanced groups across covariates. This was the only method to have insignificant differences in the propensity score density estimates, supporting previous work demonstrating the better balance obtained with this matching technique [23,25]. By using calipers in the matching process, only treated control pairs that are comparable are retained. Persons in which treatment is contraindicated or rarely indicated from the control sample (low propensity) or in which treatment is always indicated in the treatment sample (high propensity) are excluded, thus ensuring the desired feature of greater overlap of covariates. This restriction of ensuring overlap on important covariates is a relative strength of propensity score matched analysis; however, if large numbers of unmatched subjects are excluded, one should note the impact on generalizability and in the extreme case if nearly all subjects go unmatched, the comparison should probably not be made in the first place. A lack of overlap may go undetected using traditional regression approaches where the results may be overly influenced by these outliers. Interestingly, one of the criticisms sometimes leveled at propensity score analysis is that it is not always possible to find matches for individuals in the respective treatment groups—this suggests that these individuals should not be compared in the first place!

In addition to matching techniques, propensity scores can be used in stratified analyses and regression techniques [21]. Propensity scores can be used to group treated and untreated subjects into quintiles, deciles, or some other stratification level based on the propensity score, and the effects of treatment can be directly compared within each stratum. Regression approaches commonly include the propensity score as a covariate along with a reduced set of more critical variables in a regression model with an indicator variable to ascertain the impact of treatment. It should be noted that, because the propensity score is a predicted variable, a correction should be made to the standard error of any propensity score variable included in a regression. This is not standard practice and, as a result, the statistical tests of significance for such variables are generally incorrect.

One of the main potential issues of propensity scoring techniques lies in the appropriate specification or selection of covariates that influence the outcome measure or selection of the treatment. The basis of selecting variables should be based on a careful consideration of all factors that are related to treatment selection and or outcome measures [26]. There is some empirical work to help guide the analyst in specifying the propensity models, but additional research in this area is warranted before variable specification recommendations can be made conclusively. One temptation may be to exclude variables that are only related to treatment assignment but have no clear prognostic value for outcome measures. Including variables that are only weakly related to treatment selection should be considered because they may potentially reduce bias more than they increase variance [23,27]. Variables related to outcome should be included in the propensity score despite their strength of association on treatment (exposure) selection. Because the coefficients of the covariates of the propensity score equation are not of direct importance to estimating treatment effects per se, parsimony is less important and all factors that are theoretically related to outcome or treatment selection should be included despite statistical significance at traditional levels of significance. This is why many propensity score algorithms do not use variable reduction techniques, such as stepwise regression, or use very liberal variable inclusion criteria such as \( P < 0.50 \).

One of the clear distinctions between observational data analyses using propensity scoring and large randomized experiments is the inability to balance unmeasured or unmeasurable factors that may violate the treatment independence assumption critical to obtain unbiased treatment estimates. To gain practical insights into the impact of omitting important variables, an empirical exercise compared two propensity models of lipid-lowering treatment and acute myocardial infarction (AMI). One model included 38 variables and 4 quadratic terms, including laboratory results (low density lipoprotein (LDL), high density
lipoprotein (HDL), triglycerides) commonly not available in claims data, and another full model which included 14 additional variables that are not routine measures incorporated in many analyses such as the number of laboratory tests performed [28]. The reduced propensity model had a very high level of discrimination (c-statistic = 0.86) and could have been assumed to be a complete list of factors; however, it failed to show any benefit of statin initiation while the full model showed a lower risk of AMI with statin therapy, a finding comparable to clinical trials. This case study demonstrates the importance of carefully selecting all possible variables that may confound the relationship and highlights the caution one should undertake when using data sets that have limited or missing information on many potentially influential factors, such as commonly encountered with administrative data. Stürmer et al. have proposed a method of using validation data that contains richer data on factors unmeasured in larger data sets to calibrate propensity score estimates [29]. This technique offers promise to address the omitted variable issue with administrative data but would be difficult to implement on a wide scale as more rich validation samples are not routinely available.

Because propensity scoring largely utilizes the same underlying covariate control as standard regression techniques, the benefits of building a propensity scoring scalar instead of directly using the same covariates in a standard regression technique may not be obvious. Empirical comparisons between regression and propensity score estimates have been reviewed and have shown that estimates of treatment effect do not differ greatly between propensity score methods and regression techniques. Propensity scoring tended to yield slightly more conservative estimates of exposure effect than regression [30]. Despite the lack of clear empirical differences between these approaches, there are several theoretical and practical advantages of propensity scoring [22]. Matching and stratification with propensity scores identifies situations in which there exist little overlap on covariates, and these situations are elucidated clearly with propensity scoring; in matched analyses these subjects are excluded from analysis whereas these differences in exposure would be obscured in regression analyses. Stratified analyses can also elucidate propensity score treatment interactions. Because parsimony is not a consideration in the propensity scoring equation, many more covariates, more functional forms of the covariates, and interactions can be included than would be routinely considered in regression analyses. This issue is emphasized when there are relatively few outcome events where there are greater restrictions imposed on the number of covariates in regression techniques when using rules of thumb such as 8–10 events per covariate. One of the drawbacks of stratified or matched propensity scoring approaches relative to regression approaches is that the influence of other covariates (demographics, comorbidities) on the outcome measure is obscured unless additional analyses are undertaken. Overall, there is no clear superiority of regression or propensity score approaches, and ideally, both approaches could be undertaken.

When operating in the observational framework, omitting important variables because they are unavailable or are unmeasurable is often the primary threat to obtaining unbiased estimates of treatment effect. Propensity scoring techniques offer the analyst an alternative to more traditional regression adjustment and when propensity-based matching or stratification techniques are used, the analyst can better assess the “overlap” or comparability of the treated and untreated. However, the propensity score analyses in of themselves cannot address the issues of bias when there are important variables not included in the propensity score estimation. Instrumental variable (IV) techniques have the potential to estimate unbiased estimates, at least local area treatment effects in the presence of omitted variables if one or more instruments can be identified and measured. An empirical comparison between traditional regression adjustment, propensity scoring, and IV analysis in the observational setting was conducted by Stukel et al. that estimated the effects of invasive cardiac management on AMI survival [31]. The study found very minor differences between several propensity score techniques and regression adjustment with rich clinical and administrative covariates. However, there were notable differences in the estimates of treatment effect obtained with IVs (Fig. 1). The IV estimates agreed much more closely with estimates obtained from randomized controlled trials. This empirical example highlights one of the key issues with propensity scoring when there are strong influences directing treatment that are not observed in the data.

Figure 1 Effects of invasive cardiac management on AMI survival [31].
Marginal Structural Models

Standard textbook definitions of confounding and methods to control for confounding refer to independent risk factors for the outcome, that are associated with the risk factor of interest, but that are not an intermediate step in the pathway from the risk factor to disease. The more complicated (but probably not less common) case of time-varying confounding refers to variables that simultaneously act as confounders and intermediate steps, that is, confounders and risk factors of interest mutually affect each other.

Standard methods (stratification, regression modeling) are often adequate to adjust for confounding except for the important situation of time-varying confounding. In particular, confounding by indication is often time varying, and therefore, an additional concern common to pharmacoepidemiologic studies. In the presence of time-varying confounding, standard statistical methods may be biased [32,33], and alternative methods such as marginal structural models or G-estimation should be examined.

Marginal structural models using inverse probability of treatment weighting (IPTW) have been recently developed and shown to consistently estimate causal effects of a time-dependent exposure in the presence of time-dependent confounders that are themselves affected by previous treatment [34-35]. The causal relationship of treatment, outcome, and confounder can be represented by directed acyclic graphs (DAGs) [36-38].

In the Figure 2a above, A represents treatment (or exposure), Y is the outcome, and L is a (vector of) confounding factor(s). Standard statistical approaches apply.

In the case of pharmacoepidemiologic studies, drug treatment effects are often time dependent, and affected by time-dependent confounders that are themselves affected by the treatment. An example is the effect of aspirin use on the risk of myocardial infarction (MI) and cardiac death [39]. Prior MI is a confounder of the effect of aspirin use on risk of cardiac death because prior MI is associated with (subsequent) aspirin use, and is associated with (subsequent) cardiac death. However, (prior) aspirin use is also associated with (protective against) the prior MI. Therefore, prior MI is both a predictor of subsequent aspirin use, and predicted by past aspirin use, and hence is a time-dependent confounder affected by previous treatment. This is depicted in the DAG graph in Figure 2b above. Aspirin use is treatment A, and prior MI is confounder L.

In the presence of time-dependent covariates that are themselves affected by previous treatment, L(t), the estimates of the association of treatment with outcome is unbiased, but it is a biased estimate of the causal effect of a drug of interest on outcome. This bias can be reduced or eliminated by weighting the contribution of each patient i to the risk set at time t by the use of stabilized weights [Hernan et al. 2000 [35]]. The stabilized weights, \( sw_i(t) = \prod_{k=0}^{t-1} \frac{pr(A(k) = a_i(k)|A^*(k-1) = a_i^*(k-1), V = v_i)}{pr(A(k) = a_i^*(k)|A^*(k-1) = a_i^*(k-1), L*(k) = l_i, v_i)} \)

These stabilized weights are used to obtain an IPTW partial likelihood estimate. Here, A* (k – 1) is defined to be 0. The int(t) is the largest integer less than or equal to t, and k is an integer-valued variable denoting days since start of follow-up. Because by definition each patient’s treatment changes at most once from month to month, each factor in the denominator of sw(t) is the probability that the patient received his own observed treatment at time t = k, given past treatment and risk-factor history L*, where the baseline covariates V are now included in L*. The factors in the numerator are interpreted the same, but without adjusting for any past time-dependent risk factors (L*).

Under the assumption that all relevant time-dependent confounders are measured and included in L*, then weighting by \( sw_i(t) \) creates a risk set at time t, where 1) \( L^*(t) \) no longer predicts initiation of the drug treatment at time t, that is, \( L^*(t) \) is not a confounder, and 2) the association between the drug and the event can be appropriately interpreted as a causal effect (association equals causation).

Standard Cox proportional hazards software does not allow subject-specific weights if they are time-dependent weights. The approach to work around this software limitation is to fit a weighted pooled logistic regression, treating each person-month as an observation [35,40]. Using the weights, \( sw_i(t) \), the model is:

\[
\text{logit } pr[D(t) = 1|D(t-1) = 0, A*(t-1), V] = \beta_0(t) + \beta_1 A(t-1) + \beta_2 V.
\]

Here, \( D(t) = 0 \) if a patient was alive in month t and 1 if the patient died in month t. In an unweighted case, this model is
equivalent to fitting an unweighted time-dependent Cox model because the hazard in a given single month is small [40]. The use of weights induces a correlation between subjects, which requires the use of generalized estimating equations [15]. These can be estimated using standard software in SAS by the use of Proc GENMOD, with a “repeated” option to model the correlation between observations. Results are obtained in terms of the usual log-odds of the event. The final practical problem to solve is actual estimation of the weights. This is accomplished by essentially estimating the probability of treatment at time $t$ from the past covariable history, using logistic regression to estimate the treatment probabilities in the numerator (without time-dependent confounders) and in the denominator (with time-dependent confounders) [41]. The method is related to propensity scoring, where the probability of treatment is $p_t$, given covariables $[20,42,43]$. The IPTW-stabilized weight, $sw(t)$, is the inverse of the propensity score for treated subjects, and the inverse of $1 - p_t$, for untreated subjects [39].

**IVs Analysis**

**Sources of Bias in Treatment Effects Estimates**

There are a variety of sources of bias that can arise in any observational study. For example, bias can be generated by omitted variables, measurement error, incorrect functional form, joint causation (e.g., drug use patterns lead to hospitalization risk and vice versa), sample selection bias, and various combinations of these problems. One or more of these problems nearly always exist in any study involving observational data. It is useful to understand that, regardless of the source, bias is always the result of a correlation between a particular variable and the disturbance or error term of the equation. Economists refer to this problem as endogeneity, and it is closely related to the concept of residual confounding.

Unfortunately, the researcher never knows how big the endogeneity problem is in any particular study because the disturbance term is unobserved and, as a consequence, so is the extent of the correlation between the disturbance term and the explanatory variable. Given its importance, it is not surprising that the topic of endogeneity has long been an important topic in the econometrics literature. The method of IV is the primary econometric approach for addressing the problem of endogeneity. The IV approach relies on finding at least one variable that is correlated with the endogenous variable but uncorrelated with the outcome. IV approaches for addressing the problem of endogeneity date to the 1920s—although the identity of the inventor remains in doubt and will probably never be established for certain [44]. With more than nine decades to accumulate, the theoretical and applied literature on IVs estimation is vast. IVs and endogeneity are described in all of the major econometrics texts [45,46].

**The IVs Approach**

In outcomes research applications, endogeneity often raises its head in the form of sample selection bias. This is the case of nonrandom selection into treatment being due to unmeasured variables that are also correlated with the error term of the outcome equation. Sample selection bias methods developed to address this problem [47] are closely related to IVs. For the purposes of simplifying the discussion, we will consider them to be synonymous.

The first step in the estimation of a sample selection model mirrors that of the propensity score approach [48,49]. A model of treatment selection is estimated (generally using a probit model, rather than logit). Once estimated, this model can be used to predict the probability of selecting treatment $A$ as a function of observable variables, and these predicted probabilities can be compared to the patient’s actual status to calculate a set of empirical residuals. In the second step, the empirical residuals (or, more specifically, a function of these residuals known as the inverse mills ratio) are included as an additional variable. If no endogeneity bias is present, the parameter estimate on the inverse mills ratio will be statistically insignificant. However, if, for example, there are important unmeasured variables that are correlated with both treatment selection and outcomes, the included residuals will not be randomly distributed, and the variable will be either positively or negatively correlated with the outcome variable. Thus, sample selection bias models provide a test of the presence of endogeneity due to nonrandom selection into treatment due to unobserved variables that are correlated with the error term of the outcome equation. Even better, if such endogeneity is present, it is now confined to the IV—like magic the problem is solved!

**Sounds Good but Is IV Really the Holy Grail?**

Despite the appeal of sample selection or IV methods for addressing the many variants of endogeneity that commonly arise in the analysis of observational data, researchers have raised concerns over the performance of IV and parametric sample selection bias models—noting, in particular, the practical problems often encountered in identifying good instruments. It is remarkably difficult to come up with strong instruments (i.e., variables that are highly correlated with the endogenous variable) that are uncorrelated with the disturbance term. As a result, instruments tend to be either weakly correlated with the variable for which they are intended to serve as an instrument, correlated with the disturbance term, or both. As a consequence, researchers tend to gravitate toward the use of weak instruments to reduce the chance of using an instrument that is itself endogenous. Unfortunately, several studies have shown that weak instruments may lead not only to larger standard errors in treatment estimates but may, in fact, lead to estimates that have larger bias than OLS [50–53].

Staiger and Stock [51] note that empirical evidence on the strength of instruments is sparse. In their review of 18 articles published in the *American Economic Review* between 1988 and 1992 using two-stage least squares, none reported first stage $F$-statistics or partial $R^2$s measuring the strength of identification of the instruments. In several applications of IV to outcomes research problems, however, researchers have reported on the strength of their instruments [48,49,54–56]. This is good practice and should always be done to allow the reader to assess the potential strengths and weaknesses of the evidence presented.

Most recently, Crown et al. [Crown W, Henk H, Van Ness D, unpubl. ms.] have conducted simulation studies that show that even in the presence of significant endogeneity problems and when the researcher has a strong instrument, OLS analysis often leads to less estimation error than IVs. This is because even low correlations between the instrument and the error term introduce more bias than it takes away. Given the tendency to identify weak instruments in the first place, it seems unlikely that IV will actually outperform OLS in most applied situations.

This suggests that, despite the appeal of IV methods, researchers would be well advised to focus their efforts on reducing the sources of bias (omitted variables, measurement error, etc.), rather than wishing for a “magic bullet” from an IV. Among others, these methods include propensity score matching
methods, structural equation approaches, nonlinear modeling, and many of the other methods described elsewhere in this document. That said, researchers should always test for endogeneity using standard specification tests such as one of the many variants of the Hausman test [45,46]. In instances where it is possible to identify strong, uncontaminated instruments, IV methods will yield treatment estimates that are unbiased even when endogeneity is present. For excellent introductions and summaries of the IV literature, the reader may wish to consult Murray [56], Brookhart et al. [Brookhart MA, Rassen JA, Schneeweiss S, unpubl. ms.], and Basu et al. [57].

Structural Equation Modeling

In all the statistical methods discussed thus far, dummy variables are generally used to evaluate treatment effects. Although multivariate models attempt to control for other observable (and in the case of IVs, unobservable) variables, they ultimately measure an expected mean difference in the dependent variable between treatment groups. Structural models enable much more detail about the treatment effects to be elicited.

To illustrate this, consider that pharmaceutical treatment for an illness may generally be characterized by three behavioral processes and associated outcomes: 1) the choice of the drug; 2) the subsequent realization of the patient’s medication adherence behavior or drug use patterns; and 3) outcomes (e.g., mortality, survival time, relapse, tumor progression). The conceptual framework that links medical outcomes to drug choice can be represented in the form of a path analysis diagram as follows (Fig. 3).

As seen in the path diagram, we envisage choice of pharmaceutical treatment as having an effect or impact on the patient’s compliance behavior or drug use patterns. The arrow that goes from the first to the second box in the path diagram captures this effect. In turn, we expect that patient’s medication adherence will impact outcomes. The arrow between the second and third boxes in the path diagram captures this relationship.

The relationships sketched in Figure 3 may be summarized in a general way as follows:

Drug choice (D) \( D = f_0(X, T, Z, \mathbf{H}) \)
Medication adherence (A) \( A = f_1(D, X, T, Z, \mathbf{H}) \)
Outcomes (O) \( O = f_2(D, A, X, T, Z, \mathbf{H}) \)

Note that the major concepts of interest to us (drug choice, medication adherence, and outcomes) appear on the left, and the relationships among these concepts and their predictors are summarized on the right.

With this notation, \( f_0 \sim f_2 \) refers to the relationships among drug choice, medication adherence, outcomes, and their predictors. Some of these relationships may be linear and others may be nonlinear, as described below. \( X \) refers to a vector of explanatory variables that include patient characteristics such as demographic variables (e.g., gender, age, region dummies, diagnosis dummies). \( T \) refers to the vector of treatment patients received in the prior period (e.g., number of psychotherapy visits in the prior period) and baseline health conditions. \( Z \) refers to a vector of variables measuring provider characteristics. \( \mathbf{H} \) refers to baseline health characteristics of the person. In this example the structure is assumed to be recursive in nature (i.e., it is sequential). Furthermore, while a recursive relationship is plausible among the major concept areas described in Figure 3, it is also likely that some of these are determined jointly rather than sequentially. This, along with the potential that some of the equations may be nonlinear, presents a variety of interesting estimation challenges in the statistical modeling of these relationships.

In particular, the drug selection or choice of pharmacotherapy occurs first in the sequence of events. After the drug selection decision, patients generate medication adherence patterns that in turn influence the observed outcomes (i.e., probability of relapse). It is also possible, however, that outcomes can feed back on drug use patterns. For example, a patient hospitalized for mental illness is likely to experience a medication change as a result. Moreover, both drug use patterns and outcomes may be influenced by unobserved factors. As discussed earlier, such patterns of time-varying and omitted variables can lead to biased parameter estimates. Finally, drug use patterns and observed outcomes may be correlated with unobserved variables associated with drug choice.

To illustrate the issues involved in modeling outcomes associated with alternative pharmaceutical treatments, consider the outcomes associated with a decision to treat depression-related illness with an selective serotonin reuptake inhibitor (SSRI) antidepressant versus an serotonin/norepinephrine (SNRI). Drug use patterns are considered as an intermediate outcome that may have a significant effect on costs of treatment. In this analysis, antidepressant use patterns will be defined using a dichotomous variable that identifies antidepressant use as stable (4 or more 30-day prescriptions for the initial antidepressant within the first 6 months) or some other pattern of use.

These two relationships may be expressed in the following equations, which are analogous to \( f_0 \) and \( f_1 \) above:

\[
D = X_1 \lambda_1 + \epsilon_1 \\
A = X_1 \lambda_1 + X_2 \lambda_2 + \epsilon_2
\]

(1)

(2)

where D is an indicator of initial SSRI versus SNRI antidepressant selection; U is an indicator of the subsequent antidepressant use pattern that is realized; \( X_1 \) and \( X_2 \), are sets of explanatory variables (not mutually exclusive); \( \lambda_1, \lambda_2, \pi \) are parameters to be estimated.

Equation (1) models the selection of the initial antidepressant as a function of explanatory variables that include patient demographics, baseline health conditions, and provider characteristics. Similar explanatory variables appear in the use patterns equation, Equation (2), which also includes the indicator for the class of drug initially prescribed for the patient.

Suppose the research objective was to estimate rates of relapse for patients using SSRIs versus patients using SNRIs. The outcome models would have the general form:

\[
Y_r = X_r \lambda_r + U \theta + \lambda \Gamma + \epsilon_r
\]

(3a)

(3b)

where, \( Y_r \) and \( Y_r \) are outcome variables (i.e., probability of relapse) for SNRI and SSRI patients respectively; \( X_r \) are sets of explanatory variables; \( \lambda \) and \( \theta \) are parameters to be estimated; \( \Gamma \)
and \( \Gamma \) are inverse mills ratios, \( \lambda' \) and \( \lambda^* \) are the associated parameter estimates, and \( \epsilon_1 \) and \( \epsilon_2 \) are residuals.

Equations (3a) and (3b) specify observed outcomes as a function of patient and provider characteristics and according to the use pattern achieved by the patient with the study antidepressants. Also included in this specification of Equations (3a) and (3b) is an inverse mills ratio, \( \lambda \), to test for the possibility of unobserved variables that may be correlated with both initial drug selection and time to relapse. It is the IVs (sample selection bias) approach described in the previous section.

The outcome models are estimated using a variety of statistical techniques, depending upon the nature of the dependent variable. For example, time to relapse could be estimated using Cox Proportional Hazard models. Separate outcome equations for patients treated with SNRIs and SSRIs allow for structural differences in the relationships between observed outcomes and the observable characteristics of patients receiving each type of drug (i.e., different coefficient signs and/or significance of variables in the outcome models for each drug). Including the sample selection terms accounts for the potential influence of the covariance between the residuals of the antidepressant choice (1) and the outcome Equations (3a) and (3b). However, correlation between the residuals of the drug choice Equation (1) and use pattern Equation (2), as well as the correlation of the residuals of the use pattern and outcome Equations (3a) and (3b) may also affect the standard errors and bias of parameter estimates in the various equations. Moreover, the use of IV methods with non-linear outcomes equations is straightforward only for a small number of specific functional forms. For example, although our conceptual structural equation model calls for the use of IV and Cox Proportional Hazard models in combination, the reader will not find this estimator to be available in any statistical software packages.

The appropriate estimation method for the above system of equations critically depends on the structure of the covariance in the error terms across Equations (1)–(3). If the error terms are uncorrelated across equations, each equation can be estimated independently of the others. Often, however, there is reason to believe that the error covariances across each of the three equations may be nonzero. If so, parameter estimates and standard errors may be biased if the equations are estimated independently. If the equations are interrelated, any bias such as that resulting from unobserved variables may be transferred to the other equations as well.

**Deriving Treatment Effects from Structural Equation Models**

The major challenge with the use of structural equation models is that they do not contain a simple dummy variable providing the magnitude, sign, and statistical significance of the estimated treatment effect. In particular, when separate outcome models are estimated for each treatment cohort, decomposition methods are required to construct the treatment effect estimate. This is done by estimating separate outcome equations for each treatment cohort as in the above example. The coefficients in each equation show the structural relationship between the explanatory variables and the outcome variable within each cohort. In addition to these structural effects, the variables within each treatment cohort may well have different distributions (e.g., different distributions on age, gender, race, medical comorbidities). By substituting the distributions of one treatment cohort through the estimated equation of another cohort, it is possible to estimate the expected value of the outcome holding both the structural and distributional effects constant. Standard errors for the differences in expected values across treatment groups can then be generated using bootstrapping methods.

Regression-based decomposition methods have not been widely used in outcomes research but have been considered useful in labor economics to investigate wage disparities by gender and race [59]. Recently, this approach has been used to examine racial disparities in access to health care.

**Recommendations**

- Include variables that are only weakly related to treatment selection because they may potentially reduce bias more than they increase variance.
- Variables related to outcome should be included in the propensity score despite their strength of association on treatment (exposure) selection.
- All factors that are theoretically related to outcome or treatment selection should be included despite statistical significance at traditional levels of significance.
- In the presence of time-varying confounding, standard statistical methods may be biased, and alternative methods such as marginal structural models or G-estimation should be examined.
- Researchers should always report on the strengths of their instruments to allow the reader to assess the potential strengths and weaknesses of the evidence presented.
- Researchers would be well advised to focus their efforts on reducing the sources of bias (omitted variables, measurement error, etc.), rather than wishing for a “magic bullet” from an IV.
- Residual confounding should be assessed, and approaches to estimating its effect, including sensitivity analyses, should be included.

**Residual Confounding**

Residual confounding refers to confounding that has been incompletely controlled, so that confounding effects of some factors may remain in the observed treatment-outcome effect. Residual confounding is often only discussed qualitatively without trying to quantify its effect. Yet, methods are available to attempt to assess the magnitude of residual confounding after adjusted effects have been obtained [60,61]. Residual confounding should be assessed and approaches to estimating its effect, including sensitivity analyses, should be included.

**Sensitivity Analyses Related to Residual Confounding**

The basic concept of these sensitivity analyses is to make informed assumptions about potential residual confounding and quantify its effect on the relative risk estimate of the drug-outcome association [62]. Several approaches are available to obtain a quantitative estimate in the presence of assumed imbalance of the confounder prevalence in the exposure or outcome groups. The array approach varies the confounder prevalence in the exposed versus the unexposed and the magnitude of the confounder–disease association and obtains different risk estimates over a wide range of parameter constellations [63].

Another approach is directed to the question on how strong a single confounder would have to be to move the observed study findings to the null (rule-out approach). This method allows us to rule out confounders that would not be strong enough to bias our results. A limitation of this method is that it is constrained to one
binary confounder and that it does not address the problem of the effect of several unmeasured confounders.

Approaches to reduce residual confounding from unmeasured factors include:

- case-crossover study designs;
- different time periods with same patient serving as case and control.
- clinical details in a subsample;
- additional clinical information obtained on a subset of patients to adjust main results.
- proxy measures;
- measured confounders may be correlated with unmeasured confounders. High dimension propensity scoring may represent unmeasured covariate matrix.
- other methods.
- IVs.

**Conclusions**

The analysis of well-designed studies of comparative effectiveness is complex. However, the methods reviewed briefly in this section are relatively well established, in the case of stratification and regression, and/or rapidly on their way to becoming so, in the case of propensity scoring and IV analysis. Other methods, such as marginal structural models and structural equation modeling may not be as common yet in pharmaceutical outcomes research, but we expect these to become more so in the near future. Indeed, one may predict that longitudinal data analysis with time-varying measures of exposure will be almost a requirement of good observational research of treatment effects in the near future. Many other techniques such as multinomial or ordered logit or probit modeling, parametric survival analysis, transition modeling, nested models, G-estimation, and many others could not be treated at all in our report. The use of all of these methods requires extensive training, careful implementation, and appropriate balanced interpretation of findings.

Careful framing of the research question with appropriate study design and application of statistical analysis techniques can yield findings with validity, and improve causal inference of comparative treatment effects from nonrandomized studies using secondary databases.

Source of financial support: This work was supported in part by the Department of Veteran Affairs Health Services Research and Development grant HFP20-90. The views expressed are those of the authors and do not necessarily reflect the views of the Department of Veteran Affairs.

**References**

47 Heckman J. The common structure of statistical models of truncation, sample selection, and limited dependent variables and an estimator for such models. Ann Econ Soc Meas 1976;5:475–92.
Establishing Optimal Requirements for Content Validity: A Work in Progress

Sheri Fehnel, PhD
RTI Health Solutions, Research Triangle Park, NC, USA

The paper written by the ISPOR’s Patient-Reported Outcomes (PRO) Task Force, “Use of Existing PRO Instruments and their Modification: Good Research Practices for Evaluating and Documenting Content Validity,” is both an extremely important and timely contribution to the literature. Demonstrating the content validity of a PRO measure to the satisfaction of the Food and Drug Administration’s (FDA’s) Study Endpoint and Label Development (SEALD) team appears to be the most common stumbling block when requesting that such a measure be allowed to support drug approvals or label claims. As such, even instruments that are well accepted by clinicians and researchers in a given therapeutic area, have been used successfully in previous clinical trials with the same or very similar patient populations, and are supported by a wealth of quantitative evidence, have not been deemed acceptable. This perception is based not only on my own experiences, but also on those of colleagues across various pharmaceutical, biotechnology, and consulting companies.

Certainly, the paucity of PRO label claims that have been approved since the release of the draft PRO Guidance [1], as described in a recent ISPOR poster presentation [2], suggests a sizeable disconnect between the FDA/SEALD team and those seeking or attempting to support such claims (e.g., sponsors, instrument developers, and psychometricians).

The current paper, written collaboratively by the members of the ISPOR PRO Task Force and reviewed by numerous other experts in the field, describes a rigorous and practical approach to demonstrating the content validity of a PRO measure. While the substance of this paper is largely consistent with and clarifies important aspects of the draft PRO Guidance, there are also a number of important points of departure between the two documents. Perhaps most striking is the definition of content validity itself. While the draft PRO Guidance recommends and SEALD has commonly required that PRO measures include all concepts important to patients related to the proposed label claim, Rothman and her colleagues are clearly in agreement with standards questionnaire-development procedures and psychometric texts which state that only a representative or “adequate” sampling of these concepts are required for a measure to have content validity. For example, the draft Guidance defines content validity as follows: “being sure the items in the questionnaire cover all important aspects of the concept from the patient perspective” [1]. Content validity is defined elsewhere in the literature, however, as a “sampling from a pool of required content” [3]. Rather than addressing all symptoms or impacts of a condition or treatment relevant to the proposed label claim, as has been required to satisfy the FDA recently, including a representative sample of the potential universe of items in a PRO measure is both sufficiently rigorous and more practical for the purpose of clinical trials. As daily assessment of PROs is becoming commonplace, if not standard in many therapeutic areas, the number of items within a PRO measure must be considered carefully to avoid overburdening patients participating in clinical trials and to minimize missing data. Such pragmatic issues were discussed more thoroughly in an issue panel conducted in May at the ISPOR International Meeting by PRO experts employed by a pharmaceutical company, a consulting firm, and the National Cancer Institute [4].

While the FDA/SEALD team’s intent is clearly to protect consumers, many of us “in the trenches” fear that the hurdle for PROs, particularly with respect to drug approvals, may be delaying the availability of products in development, as well as discouraging sponsors from developing compounds for which PROs will be necessary. Nonetheless, I remain optimistic that scientific contributions such as the paper written by ISPOR’s PRO Task Force, well-respected experts in the field, and substantiated by standard psychometric principles will foster further dialog among all parties involved and ultimately ensure that the hurdle for the approval of future PRO claims is set at an optimal level—one that protects patients from false claims while also facilitating the communication of treatment benefits backed by scientifically appropriate evidence.

Source of financial support: None.

References
1 FDA. Food and Drug Administration draft guidance for industry on patient-reported outcome measures: use in medical product development to support labeling claims. Federal Register, 2006.
Use of Existing Patient-Reported Outcome (PRO) Instruments and Their Modification: The ISPOR Good Research Practices for Evaluating and Documenting Content Validity for the Use of Existing Instruments and Their Modification PRO Task Force Report

Margaret Rothman, PhD,1 Laurie Burke, RPh, MPH,2 Pennifer Erickson, PhD,3 Nancy Kline Leidy, PhD,4 Donald L. Patrick, PhD, MSPH,5 Charles D. Petrie, PhD6

1WW Patient-Reported Outcomes Center of Excellence, Johnson & Johnson Pharmaceutical Services, LLC, Raritan, NJ, USA; 2Study Endpoints & Labeling, Office of New Drugs, CDER, FDA, Silver Spring, MD, USA; 3On-line Guide to Quality-of-Life Assessment (OLGA), State College, PA, USA; 4Scientific Affairs, United BioSource Corporation, Bethesda, MD, USA; 5University of Washington, Seattle, WA, USA; 6Pfizer, Inc, New London, CT, USA

ABSTRACT

Background: Patient-reported outcome (PRO) instruments are used to evaluate the effect of medical products on how patients feel or function. This article presents the results of an ISPOR task force convened to address good clinical research practices for the use of existing or modified PRO instruments to support medical product labeling claims. The focus of the article is on content validity, with specific reference to existing or modified PRO instruments, because of the importance of content validity in selecting or modifying an existing PRO instrument and the lack of consensus in the research community regarding best practices for establishing and documenting this measurement property.

Methods: Topics addressed in the article include: definition and general description of content validity; PRO concept identification as the important first step in establishing content validity; instrument identification and the initial review process; key issues in qualitative methodology; and potential threats to content validity, with three case examples used to illustrate types of threats and how they might be resolved. A table of steps used to identify and evaluate an existing PRO instrument is provided, and figures are used to illustrate the meaning of content validity in relationship to instrument development and evaluation.

Results & Recommendations: Four important threats to content validity are identified: unclear conceptual match between the PRO instrument and the intended claim, lack of direct patient input into PRO item content from the target population in which the claim is desired, no evidence that the most relevant and important item content is contained in the instrument, and lack of documentation to support modifications to the PRO instrument. In some cases, careful review of the threats to content validity in a specific application may be reduced through additional well documented qualitative studies that specifically address the issue of concern.

Conclusion: Published evidence of the content validity of a PRO instrument for an intended application is often limited. Such evidence is, however, important to evaluating the adequacy of a PRO instrument for the intended application. This article provides an overview of key issues involved in assessing and documenting content validity as it relates to using existing instruments in the drug approval process.

Keywords: content validity, instruments, outcomes, patient-reported outcomes, validity.

Background to the Task Force

In January 2007, the ISPOR Health Science Policy Council recommended to the ISPOR Board of Directors that an ISPOR Patient-Reported Outcomes (PRO) Task Force on Use of Existing Instruments and Their Modification be established. The Board of Directors approved this PRO Task Force in March 2007. The PRO Task Force chair and members were chosen based on their experience as scientific leaders in the field, as well as developers and users of PRO instruments. A range of perspectives on PRO instruments was provided by the diversity of their work experiences: research organizations, government, academia, and industry.

The PRO Task Force met every 6 weeks to develop the report outline and discuss issues that arose in the manuscript’s development. The manuscript outline was presented for comment at ISPOR 12th Annual International Meeting, Arlington, VA, USA in 2007. The first draft report was presented for comment at the 13th Annual International Meeting, Toronto, Canada in 2008.

In March 2009, the second draft report was submitted for review and comments to the 360 members of the ISPOR PRO Special Interest Group. Comments were discussed and incorporated. A revised draft report was presented for final comment at the ISPOR 14th Annual International Meeting, Orlando, FL, USA in May 2009. The task force then addressed and incorporated comments elicited during this final review process as appropriate. All written comments on the PRO Existing Instruments and Their Modification Task Force Report are available on the ISPOR Web site at: http://www.ispor.org/TaskForces/PROInstrumentsUse.asp.

Introduction

During early deliberations, the task force discussed the importance of content validity in selecting or developing a modified...
version of an existing PRO instrument and the lack of consensus in the research community regarding best practices for establishing and documenting this important measurement property. The task force decided to focus their review and discussion on content validity in PRO evaluation with specific reference to existing or modified instruments. It should be noted that this article is meant to represent current best practices in assessing the content validity of existing PRO instruments for purposes of making a regulatory claim rather than a repetition of the Food and Drug Administration (FDA) Draft PRO Guidance [1]. We feel that the recommendations in this article are consistent with the guidance; however, terminology and interpretation may vary slightly. A working knowledge of instrument development and qualitative methods is assumed; those readers who desire to know more about this topic are encouraged to read one of the many basic textbooks, including those referenced throughout the document. The article begins with a definition and general description of content validity. This is followed by a discussion of PRO concept identification as the important first step in establishing content validity in the context of a regulated claim (section III). A description of identification of an instrument and the initial review process, summarized in Table 1 (section IV), is followed by a discussion of key issues in qualitative methodology, the foundation of content validity (section V). The article concludes with a discussion of potential threats to content validity (section VI), with three case examples used to illustrate types of threats and how they might be resolved, as presented in Tables 2 and 3. References of existing work in each of the areas discussed are provided throughout the article to guide the reader interested in further information. Whereas an overview of key issues in qualitative research methodology as it relates to either selecting or developing a modified version of an existing instrument is presented, it is not the intent of this article to provide an in-depth review of these methods which is available in more detail elsewhere.

**Content Validity: Basic Principles**

Definition: The goal of measurement is to quantify a concept. The Standards for Educational and Psychological Testing, a key reference describing development and evaluation of instruments used to evaluate an individual's behavior, defines validity as an overall assessment of the degree to which evidence and theory support the interpretation of scores entailed by proposed uses of the instrument [1]. (The Standards for Educational and Psychological Testing uses the term “tests” rather than “instruments.”) The term “test” may be confusing when used in the context of PRO research; therefore, the term PRO “instruments” is used in this document.) One type of validity is based on content. Evidence of content validity is obtained from an analysis of the relationship between an instrument’s content and the construct it intends to measure. The Standards for Educational and Psychological Testing uses the term “construct” as “the concept or characteristic that a test is designed to measure.” The term “concept” will be used in this article to be consistent with the

<table>
<thead>
<tr>
<th>Table 1</th>
<th>Steps in identifying and evaluating an existing PRO measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>✓ 1. Name and define the concept.</td>
<td></td>
</tr>
<tr>
<td>✓ 2. Draft the claim, target population, target product profile, and end point model.</td>
<td></td>
</tr>
<tr>
<td>✓ 3. Identify candidate measures.</td>
<td></td>
</tr>
<tr>
<td>✓ 4. Identify or formulate a conceptual framework for the instrument(s).</td>
<td></td>
</tr>
<tr>
<td>✓ 5. Assemble and evaluate information on development methods.</td>
<td></td>
</tr>
<tr>
<td>✓ a. Elaboration focus groups and interviews; sample size and characteristics relative to intended use; analytical approach; results, including evidence of saturation</td>
<td></td>
</tr>
<tr>
<td>✓ b. Cognitive interviews; sample size and characteristics relative to intended use; methods (mode of administration); results</td>
<td></td>
</tr>
<tr>
<td>✓ c. Transcripts—for independent review and stratified analyses; evidence of saturation</td>
<td></td>
</tr>
<tr>
<td>✓ 6. Conduct any needed qualitative research.</td>
<td></td>
</tr>
<tr>
<td>✓ a. Prepare documents for decision-making and regulatory submission—including protocol and study report, saturation tables, transcripts.</td>
<td></td>
</tr>
<tr>
<td>✓ b. Map qualitative data to existing instrument items.</td>
<td></td>
</tr>
<tr>
<td>✓ 7. Assess adequacy of content validity for purpose.</td>
<td></td>
</tr>
<tr>
<td>✓ a. Identification of any concepts, domains, or items relevant to patients in the target population not included in the existing instrument</td>
<td></td>
</tr>
<tr>
<td>✓ b. Assess the relevance and importance of this content</td>
<td></td>
</tr>
<tr>
<td>✓ 8. Determine the need for modifications or new instrument development.</td>
<td></td>
</tr>
<tr>
<td>✓ a. Evaluate cost benefit in terms of need, timelines, and resource allocation.</td>
<td></td>
</tr>
<tr>
<td>✓ b. Discuss alternatives—change the claim/concept; alter PRO positioning.</td>
<td></td>
</tr>
</tbody>
</table>

PRO, patient-reported outcome.

<table>
<thead>
<tr>
<th>Table 2</th>
<th>Framework (model) for evaluating content validity of an existing instrument within the context of a specific claim</th>
</tr>
</thead>
<tbody>
<tr>
<td>Element</td>
<td>Acceptable level of agreement†</td>
</tr>
<tr>
<td>1. Conceptual match‡</td>
<td></td>
</tr>
<tr>
<td>2. Input from the target population§</td>
<td></td>
</tr>
<tr>
<td>3. Item content, including saturation¶</td>
<td></td>
</tr>
<tr>
<td>4. Modification, e.g., mode of administration, translation*§</td>
<td></td>
</tr>
</tbody>
</table>

†The greater the absence of evidence to substantiate content validity, the more comprehensive the evaluation and the greater the degree of remediation required. Definitions of the elements follow:

‡Level of evidence refers to the basis on which the elements are evaluated. The following four levels of evidence are proposed: completely met, mostly met, partially met, not met.

§Conceptual match: The concepts, as defined by the developer and represented by the items in the instrument, match the intended claim.

⊕Input from the target population: Patient concerns were obtained using appropriate qualitative methods.

¶Item content: Each concept in an item reflects patient concerns across the range of patients appropriate for the intended claim.

*Modification: Any change in the instrument from the original version needs to be identified and evaluated for its impact on content validity in terms of the above elements, in addition to the psychometric performance of the modified version relative to the original form.
FDA draft PRO guidance. Content refers to the themes or subjects addressed in the instrument; the wording and format of items, tasks, or questions on an instrument; as well as the guidelines for procedures for administration and scoring. In the context of content validity, the appropriateness of the content is related to the specific inferences to be made from the instrument scores. An evaluation of content validity is therefore essential in the selection of instruments to evaluate PROs to be used in making labeling claims.

The classic text, *Psychometric Theory*, by Nunnally and Bernstein [2], notes that there are two major standards for ensuring content validity. One standard is the representative nature of the collection of items comprising the instrument. Because “random sampling” is impractical, the method used to identify and select the items to represent the concept must be explicit. The second, related standard is based on the methods used in instrument construction, i.e., the rigor with which the instrument is constructed or formulated, including item and response option wording, scaling method (e.g., dichotomous, Likert, visual analog), and organization [1]. The appropriateness of a given content domain is related to the specific inferences to be made from the instrument scores [1].

PRO instruments are designed to capture data related to the health experiences of individuals, specifically the way patients feel or function in relation to their condition or disease and/or their treatment. Figure 1 depicts the relationship between disease attributes, including observed signs and laboratory values, and patient-reported manifestations of the condition, e.g., symptoms, and patient experiences, including their descriptions of the disease and human experiences unrelated to the disease. For any disease, there are a host of characteristics, including observed signs, laboratory values, and patient-reported manifestations of the disease. Patients with the disease have a wide range of experiences, including those directly related to the disease itself, and other experiences that may be important to the patient’s life experience, but are not characteristic of the specific disease of interest. For PRO instruments, content validity begins with the intersection between disease characteristics and patient experience, as shown in Figure 1, and is evaluated in terms of the adequacy with which this intersection “universe” of content is sampled in a given instrument to accurately capture patient-reported manifestations of the disease. Although the focus of this article is on PRO instruments used to support evidence of treatment efficacy, the adequacy of content validity for instruments to assess adverse treatment impact would follow the same principles.

In practice, content validity is determined by the relationship between the intended measurement concept and the methods used: 1) to develop and select items; 2) to evaluate the content; and 3) to formulate the instrument. A detailed description of these methods and their results provides evidence that the pro-
posed use and interpretation of scores from the instrument represent the intended concept and therefore possess content validity [1–5]. Documentation of content validity describes the following three processes: identifying and defining the concept of interest, identifying the patient experience related to the targeted concept, and using appropriate methodology to develop the instrument to capture and quantify the concept.

Concept Identification within the Labeling Context

Selecting or developing a modified version of an existing PRO instrument for evaluating treatment efficacy begins with the identification of the concept to be measured and its positioning in relationship to other trial end points to be used in the development program, as described in the United States FDA PRO Draft Guidance [6]. Outcomes essential for testing product efficacy and informing appropriate use in clinical practice are considered highest priority and serve as primary end points in clinical trials. Outcomes considered important for providing additional understanding of treatment benefit or for communicating product attributes of value to the patient are also identified early in the development program and can serve as key secondary end points. Exploratory end points are useful for product development planning, including gathering data to inform future trial design decisions, but are not used to support labeling or promotional claims. PRO instruments can serve as primary, secondary, or exploratory end points.

Targeting the desired claim using the target product profile (TPP) approach can be useful when selecting and positioning the PRO, and linking it to the desired claim [7,8]. An end point model that describes each study end point, including PRO end points and their relationship to other end points, may be developed consistent with the goals identified in the TPP. End point models are useful for product development planning, including gathering data to inform future trial design decisions, but are not used to support labeling or promotional claims. PRO instruments can serve as primary, secondary, or exploratory end points.

Targeting the desired claim using the target product profile (TPP) approach can be useful when selecting and positioning the PRO, and linking it to the desired claim [7,8]. An end point model that describes each study end point, including PRO end points and their relationship to other end points, may be developed consistent with the goals identified in the TPP. An end point model should clearly delineate the relative importance of each end point in terms of labeling priorities to inform the subsequent drafting of study objectives, study protocol, and statistical analysis plan. Each concept to be measured is included, as well as how each concept will correspond to the ultimate labeling goals (i.e., the claims).

Selecting and defining the concept, specifying the intended claim, identifying the target population, and drafting the TPP are the initial steps in assuring and documenting content validity. Existing instruments are identified and reviewed with this information in mind.

Instrument Identification and Initial Review Process

Existing PRO instruments can be identified through literature searches, Web searches, and dedicated instrument databases. A review of candidate instruments includes a close examination of the items (stem and response options), recall period, mode of administration and instructions in relationship to the targeted concept in keeping with good scientific practice [9], and the Draft PRO Guidance [6].

If a conceptual framework of the instrument is available, it is examined for consistency with the concept. A conceptual framework is a detailed description or diagram of the relationship among the concepts, domains, and items comprising the instrument [10,11]. If such a framework does not exist, one should be developed, showing how the items, subscales, and total scales are related to one another and to the underlying concept and claim. The names used to describe the concept and subscales should be critically evaluated in light of the content and structure of the items and the targeted PRO claim. Adjustments in the name or concept referenced in the PRO instrument may be desirable to more accurately reflect the content and link to the claim. These adjustments, however, should be made after consulting with the instrument developer, keeping in mind that any name changes can have an adverse effect on subsequent interpretation and attempts to replicate findings. Strong and clear links between item content, subscale names, concept names, study objectives, and target claims are desirable and enable ease of understanding, interpretation, and communication.

A complete understanding of the methods used to develop a PRO instrument is essential to evaluate the suitability of an existing PRO instrument for any purpose. These methods are generally available in published literature and documentation; however, some may need to be obtained directly from the developer. An instrument may be best evaluated for potential use to support claims if: 1) patient-derived qualitative data forming the basis of the instrument are available; 2) a careful critique shows

Figure 1  Content validity: the intersection of disease characteristics and patient experience.
the data generation methodology to be sound, and a sample similar to the development program’s target population was used; and 3) the results of this qualitative work are comprehensive and relevant. Documentation of methods and results should demonstrate that the instrument is adequate for decision-making, and appropriate for regulatory submission as part of a PRO evidence dossier [12].

It is not uncommon for existing instruments to have little to no published or available information on their development history, to have little to no patient input in the development process, or to have minimal qualitative data from the patient population specifically enrolled in the trial. In these cases, there is little empirical evidence for the sponsor or regulatory reviewer to make an informed decision regarding the content validity of a PRO instrument, and suitability for the drug development program. If clinical and research experts conclude that the items in the PRO instrument may be an adequate representation of the concept in the intended claim and that there is quantitative evidence of reliability and validity, then qualitative research methods can be used to elicit information from patients to inform instrument development. That is, these methods are used to obtain descriptive words and phrases, as shown in Figure 1. Cognitive debriefing interviews, generally conducted 1:1, are used to evaluate patient understanding of an instrument, as a draft item pool or as an existing measure being evaluated for possible use. Items are developed and evaluated with the target patient population in mind with wording designed to maximize ease of reading, translatability, and content coverage, with the final instrument reflecting all of these considerations. As shown in Figure 2, a PRO instrument is based on the concepts, words, and phrases generated during elicitation focus groups and interviews, with adjustments made based on interpretation and meaning provided during cognitive interviews with a new set of patients. Although patients are experts in their personal experience with a disease, they are not experts in the disease pathophysiology or instrument design techniques. Instrument developers play a key role in the development of consensus wording of items and response options, the selection of recall period, and the instructions to respondents, based on their knowledge of the content area and instrument development techniques.

**Sample Selection**

For product labeling in the context of a regulated claim, the FDA Draft PRO Guidance [6] indicates that PRO instruments assessing treatment benefit should show evidence of content validity using methods that include the elicitation of input from the patient population for which the claim is intended, and results that demonstrate the relevance and importance of item content to this group. Because these instruments are designed to capture patient experiences, the Draft PRO Guidance suggests this input be elicited directly from the patients targeted for the clinical trials. Clinicians or other experts or literature may be useful in

---

**Figure 2 Content validity: content consensus through qualitative research.**

---

**Content Validity of Existing PRO Instruments**

“Qualitative research methods involve the systematic collection, organization and interpretation of textual materials from talk or observation. It is used in the exploration of meanings of social phenomena as experienced by individuals themselves, in their natural context” [13]. These empirical methods, specifically focus groups and 1:1 interviews, are used to elicit information from patients to inform instrument development. That is, these methods are used to obtain descriptive words and phrases, as shown in Figure 1. Cognitive debriefing interviews, generally conducted 1:1, are used to evaluate patient understanding of an instrument, as a draft item pool or as an existing measure being evaluated for possible use. Items are developed and evaluated with the target patient population in mind with wording designed to maximize ease of reading, translatability, and content coverage, with the final instrument reflecting all of these considerations. As shown in Figure 2, a PRO instrument is based on the concepts, words, and phrases generated during elicitation focus groups and interviews, with adjustments made based on interpretation and meaning provided during cognitive interviews with a new set of patients. Although patients are experts in their personal experience with a disease, they are not experts in the disease pathophysiology or instrument design techniques. Instrument developers play a key role in the development of consensus wording of items and response options, the selection of recall period, and the instructions to respondents, based on their knowledge of the content area and instrument development techniques.
preparing interview guides and in defining enrollment criteria in the qualitative studies, but without patient input, PRO instrument development is incomplete.

Patients should represent variations in severity of the target condition, as well as in population characteristics such as age, sex, ethnicity, and language groups in accordance with the anticipated characteristics of the patients to be enrolled in the clinical trials. For example, the content of an existing instrument may initially have been generated on a broad population of adults ranging from age 18 to 65 years. Subsequent use of the instrument outside of this age range, e.g., adolescents or older adults, would require additional qualitative research for eliciting or confirming age-relevant content, testing modifications of the existing item content, and determining adequate comprehension of the instrument items and response format.

**Data Collection Methods**

Data from 1:1 interviews and/or focus groups with patients form the basis of PRO instrument content. When evaluating candidate instruments, users should examine the data collection methods used to generate the instrument to understand the content validity of the instrument. Gathering qualitative data through 1:1 interviews or focus groups is both a rigorous scientific method requiring a well-defined protocol, and an art requiring trained and experienced interviewers or focus group moderators. Interviews are audio recorded for later transcription and analyses. To assure representation from all group participants and to assist in data analyses, focus group interviews/moderators are trained to encourage rapport and elicit comments from all participants. Focus groups can be videotaped and/or audio recorded. An assistant moderator can take notes with participant initials, and key words or quotations to facilitate data transcription and analyses. This individual can also map the discussion, marking the frequency with which various participants contribute comments to the discussion, and alerting the moderator for the need to query certain participants who have been less active in the discussion.

One-on-one interviews require a skilled interviewer and are particularly effective for sensitive topics unsuited to a group discussion, e.g., urinary incontinence, or for patient populations unable to participate or uncomfortable in a group setting, e.g., men with erectile dysfunction. One-on-one interviews are also used for cognitive interviews in which patients review an existing instrument or item pool, and can provide the developer with both insight into the extent to which their interpretations match the intent of the items and with any critical content that has been omitted from the instrument.

Cognitive debriefing interviews with patients from the target population are used to evaluate patient understanding of the items relative to the concept of interest. Also known as complement elicitation focus groups or interviews, cognitive interviews provide additional evidence of content validity. Cognitive interviews also provide an opportunity to query patients about the comprehensiveness of the instrument content relative to their experiences with the concept of interest, serving as a "pilot test," to make certain the instrument selected is, in fact, interpreted correctly, no additional instructions or special training is required, and that all of the appropriate concepts are covered. Specifically, at the end of the interview, patients may be asked if there were any aspects of the concept, e.g., experiences, symptoms, or sensations that were not addressed in the instrument, and if so, how important these are to the concept. If missing themes emerge across multiple interviews, and these themes are clearly related to the underlying concept, it is likely the instrument is missing important content and should be modified before use in a development program. This finding is referred to as "construct underrepresentation" [1].

Qualitative data from elicitation or evaluative methods not only provide information on the content validity of an existing measure, but also offer insight into concept names used to represent scales or subscales. As discussed previously, an assessment of the formulation of the conceptual framework of an instrument may suggest that scale or subscale names created by developers are unclear or inaccurate, particularly for regulated claims. Qualitative data can inform the evaluation of instrument naming conventions, and suggest alternatives more suited to the concept and proposed claim. Although it is usually inappropriate to rename existing instruments and domains, an accurate description of the content may facilitate communication between researcher and reviewer.

**Data Analysis**

Analyses of qualitative data use a carefully constructed methodology that includes independent coding with inter-rater analyses and reconciliation. When reviewing instrument development methods, attention should be paid to how the data were analyzed. Coding transcripts by participant, using initials or other coding system to protect anonymity, allows the researchers and reviewers to evaluate the representativeness of content across participants, and provides assurance of saturation (discussed in greater detail below). Coding by a specific patient characteristic, such as gender or disease severity, permits stratified analyses with an assessment of the consistency in experiences, and the words and phrases to describe these experiences by patient subgroups. Reference to this type of analysis can reassure sponsors and regulators that the instrument development methods were rigorous and that the instrument captures the content most relevant for the concept across the full range of the target population.

Analyses of focus group and interview data to evaluate and document the content validity of an existing measure are similar to those used in instrument development, identifying themes that emerge from the data in relationship to the concept of interest. These themes are used as analytical codes that are then mapped to the existing instrument content, with words and phrases compared with the wording used in the instrument. Figure 3 shows three examples of the relationship between the content of an existing measure relative to the universe of content derived through qualitative research methods. Example A shows a strong match between the content in the instrument and the information provided by patients. Note that the instrument content is not all inclusive, but represents the vast majority of the potential item content. Example B shows a poor match, with the content capturing less than 30% of the possible concept content. Example C shows a mismatch, with some of the instruments covering the content of interest, and coverage of content external to the concept of interest. It is unlikely that the instruments in examples B and C would be suitable for use as a primary or key secondary end point in registration trials because of the inadequate coverage of item content in the instrument with relevant information provided by patients. In such cases, a decision must be made either to adopt the existing instrument or to develop a new one, or to re-evaluate the intended claim and relevant PRO concept to support it.

**Saturation**

Qualitative data should be gathered to the point of saturation to ensure that the items in an instrument appropriately represent the relevant "universe of content" for the concept when conducting focus groups or 1:1 interviews. In instrument development, satu-
ration refers to the point in the data collection process when no new concept-relevant information is being elicited from individual interviews or focus groups, or no new information is elicited and deemed missing during cognitive interviewing. There is no fixed rule on either the sample size or the number of iterations needed to reach saturation. During the development process, evidence is based on empirical observation, where no new concepts or codes emerge after the nth interview or focus group [10,14]. Saturation can be evaluated and documented through a saturation table structured to show the elicitation of information by successive focus group or interview (individual or by set), organized by concept code. For practical purposes of budgeting projects, it is not uncommon to set a sample size of 20–30 interviews, even though saturation may occur earlier in the interview process. Saturation is then documented for where it occurs in the process, often during the interviewing process or sometimes at the end of all interviews. The risk of waiting to the end of the project before applying a coding frame and analyzing qualitative data is that saturation may not be reached and additional interviews will have to be conducted, or that saturation was evident at some point before the end of the designated stopping point, and the additional interviews were unnecessary.

It is not at all uncommon for participants in focus groups and interviews to stray from the intent of the interview, discussing issues or concepts important to them as individuals or as patients, but unrelated to the concept of interest (see Fig. 1, unrelated descriptions). These data are clearly not part of the core concept, are inconsistent with the intent of the instrument, and should be excluded from a description of the concept and saturation tables. This is a particularly important issue when analyzing qualitative data to evaluate the content validity of an existing instrument. For example, data from focus groups eliciting information on patient perception of anginal pain as part of an evaluation of an anginal pain scale may include comments from one or more patients describing their knee pain. These descriptions should be coded as part of the patient’s comorbid experience, but excluded from the thematic summary and content item mapping.

Identifying and Resolving Threats to Content Validity

The fundamental question in the evaluation of a PRO instrument in the context of a labeling claim, whether an instrument is new, existing, adapted, or modified, is the adequacy of empirical evidence to support content validity for the desired claim. Four important threats to PRO content validity, in relative order of importance, are shown in Table 2. This table also provides an organizing heuristic for evaluating each threat; a discussion of this approach to resolving these threats follows:

**Absent or Unclear Conceptual Match Evident between the PRO Instrument and the Intended Claim**

As discussed earlier in the article, the conceptual match is the primary task in identifying the PRO instrument’s conceptual framework, and specifying its linkage to the intended claim. If there is no clear match, then probably the most effective strategy is to identify another instrument to measure the claim with targeted concepts relevant for supporting the claim.

**Lack of Direct Patient Input into PRO Item Content from the Target Population in which the Claim Is Desired**

The patient population in which the PRO instrument was developed should be compared to the patient population targeted for enrollment in the clinical trials, to determine whether the instrument is appropriate for that population. This requires sponsors to carefully consider the inclusion and exclusion criteria of the clinical trial to identify important patient characteristics that may indicate a need for additional validation work.

**Lack of Evidence Regarding Saturation—No Evidence that the Most Relevant and Important Item Content Is Contained in the Instrument**

This threat addresses limitations in empirical data to confirm that the PRO item content associated with each concept captures the full range of important and relevant patient experiences across a representative sample of the targeted patient population. Evaluative interviews, focus groups, and cognitive debriefing interviews can be used to address this threat.

**Modification of the Original PRO Instrument**

Modifications to an instrument may include: changes in wording or content, changes in mode of administration, translation and cultural adaptation, and application to a different patient popu-
Making, and/or insufficient for regulatory submission, additional information is inadequate for informed decision-making, case in scenario A.

In scenario A, a PRO measure was developed by clinicians to assess dyspnea in chronic obstructive pulmonary disease in clinical practice. The question here is whether the instrument could be used to evaluate treatment efficacy in clinical trials involving a new patient population, i.e., asthma. In this example, the first threat is that there may be aspects of the concept of “dyspnea” that are experienced uniquely by asthma patients and are not addressed in the instrument. A review of the literature and discussions with the developer may uncover qualitative study reports in which data from patients with asthma are presented as part of the instrument development process or by others interested in using the instrument in asthma. The review might also uncover independent qualitative studies examining the concept of dyspnea as experienced by patients with asthma with characteristics similar to those to be enrolled in the trials.

The second threat in this example is the potential that the instrument was developed without patient input. This is addressed by reviewing all of the documentation from the developer(s) for evidence of direct patient input. The material should provide a detailed description of all of the steps that were taken to identify instrument content, and to generate specific items based on patient data. Instruments developed before 2000 often lack this information entirely or in the detail needed to support labeling claims following the FDA Draft PRO Guidance [6]. Remediation for this problem is discussed below.

A third threat involves the change in intended use of the instrument, from clinical practice to clinical trial. From a content validity standpoint, the most important element of this threat is the method used to inform the overall design of the instrument, including recall period and item content. It is not uncommon for clinical instruments to be developed based solely on clinician expertise and experience, with content that addresses the specific information needs of the practice setting. For example, if the instrument was developed using qualitative research methods with direct input from patients with asthma similar to those to be enrolled in the clinical trial, and if results of this work are available, the magnitude of the threat declines. This is not the case in scenario A.

If documentation provided by the developer or available in the published literature is inadequate for informed decision-making, and/or insufficient for regulatory submission, additional work will be needed. In this scenario, one approach that could be used is to conduct focus groups and/or cognitive interviews with asthma patients who meet the inclusion/exclusion criteria for the product development trials, mapping the results to the instrument. This would enable the user to evaluate concept coverage and, if adequate, provide data for documenting the relationship between patient data and instrument content in the target population. If the content is found to be inadequate, this process would provide the sponsor with an opportunity to modify the instrument, with permission, and perhaps participation, of the instrument developer, with the potential for increasing the sensitivity of the instrument to detect treatment effects in clinical trials involving this new target population.

Scenario B illustrates a second example of a potential threat to content validity. In this example, an existing, multi-item, multiscale measure has been reduced to a shorter, more efficient assessment form using psychometric procedures, including cross-validation in the relevant patient samples, and has been shown to have nearly identical psychometric properties, including factor structure, to the original long form of the instrument. Here, the threat is that there may be aspects of the concept of “dyspnea” that are experienced uniquely by asthma patients and are not addressed in the instrument. A review of the literature and discussions with the developer may uncover qualitative study reports in which data from patients with asthma are presented as part of the instrument development process or by others interested in using the instrument in asthma. The review might also uncover independent qualitative studies examining the concept of dyspnea as experienced by patients with asthma with characteristics similar to those to be enrolled in the trials.

The second threat in this example is the potential that the instrument was developed without patient input. This is addressed by reviewing all of the documentation from the developer(s) for evidence of direct patient input. The material should provide a detailed description of all of the steps that were taken to identify instrument content, and to generate specific items based on patient data. Instruments developed before 2000 often lack this information entirely or in the detail needed to support labeling claims following the FDA Draft PRO Guidance [6]. Remediation for this problem is discussed below.

A third threat involves the change in intended use of the instrument, from clinical practice to clinical trial. From a content validity standpoint, the most important element of this threat is the method used to inform the overall design of the instrument, including recall period and item content. It is not uncommon for clinical instruments to be developed based solely on clinician expertise and experience, with content that addresses the specific information needs of the practice setting. For example, if the instrument was developed using qualitative research methods with direct input from patients with asthma similar to those to be enrolled in the clinical trial, and if results of this work are available, the magnitude of the threat declines. This is not the case in scenario A.

If documentation provided by the developer or available in the published literature is inadequate for informed decision-making, and/or insufficient for regulatory submission, additional
with the intent. If this is not the case, the original instrument may need modification.

Conclusions

Content validity refers to the extent to which an instrument contains the relevant and important aspects of the concept(s) it intends to measure. This article discussed the key issues involved in assessing and documenting the content validity of an existing instrument, including concept clarification, instrument identification, and initial review, as well as qualitative methods as they might be used to evaluate the suitability of one or more existing instruments. Case examples illustrate threats to content validity and various approaches for remediating these threats. Several tools were identified to aid in the evaluation of content validity, including end point models that describe the correspondence between concepts, measures, and labeling goals; the conceptual framework of the PRO instrument to evaluate and communicate the extent of the match between item content and targeted concepts; and qualitative research methods that form the empirical basis for evaluating and documenting content validity.

Source of financial support: The views expressed herein represent those of the authors and not those of Johnson & Johnson, the FDA, OLGA, United BioSource Corporation, University of Washington, or Pfizer.

References


6 FDA. Food and Drug Administration draft guidance for industry on patient-reported outcome measures: use in medical product development to support labeling claims. Federal Register 2006.
8 Delasko J, Coccheto DM, Burke LB. Target product profile: beginning drug development with the end in mind. UPDATE 2005;January/February.
17 Wild D, Eremenco S, Mear I, et al. Multinational trials—recommendations on the translations required, approaches to using the same language in different countries, and the approaches to support pooling the data: the ISPOR patient reported outcomes translation & linguistic validation good research practices task force report. Value Health 2009;12:430–40.
The Imperative to Improve

Chris L. Pashos, PhD

HERQuLES: Health Economic Research and Quality of Life Evaluation Services. Abt Bio-Pharma Solutions, Inc., Lexington, MA, USA

Keywords: ISPOR Task Forces, cost-effectiveness research, quality improvement, evidence-based medicine, decision-makers.

Providing access to appropriate health care given economic constraints is a challenge that confronts all the world’s nations. Pharmacoeconomics, and more broadly health economics, and outcomes research can offer critical input to help our countries worldwide to understand the burden of disease and give us the tools to help allocate, or reallocate, resources to optimize, or at least improve, access to health care, and improve public health. Cost-effectiveness research has the potential to contribute much in this regard.

Yet, this year, as public policy makers—around the world from the United States to Greece to China, and many other nations—debate issues of health care financing and delivery, the role in resource allocation of health economics and outcomes research in general, and cost effectiveness research specifically, is relatively modest.

In policy debates in the world’s capitals and even in plenary sessions at ISPOR’s international congresses, one can see the skepticism that often greets published cost-effectiveness research. To the extent that skepticism may be at least in part attributed to the use of inappropriate or immature methods, it has been and remains incumbent on ISPOR’s members to continuously promote, if not ensure, the rigor of the scientific methods we use, and to continuously promote advancements in those scientific methods and their interpretation to all who review and use them or findings that result from their use.

In fact, as its central mission since its inception, ISPOR has sought to promote the science of pharmacoeconomics (health economics) and outcomes research and facilitate the translation of this research into useful information for health care decision-makers to ensure that society allocates scarce health-care resources wisely, fairly and efficiently. Many of the core activities of the Society have had as their central objective the continuous improvement of the quality of the science and the quality of the presentation of that science to core constituencies, including physicians, pharmacists and other health-care providers, as well as policy makers, payers, patients and others. These core activities have included the international meetings, distance learning initiatives, and this journal.

Early in her service as Editor-in-Chief for Value in Health, Dr. Josephine Mauskopf set forth her vision for the journal. She deemed “bridging the gap” to be the journal’s central role—in effect building bridges between researchers of the multiple disciplines comprising pharmacoeconomics and outcomes research to develop and improve the field by cross-fertilizing these multiple disciplines, and bringing an ever-improving science to practitioners and policy decision-makers. The ultimate goal has been for the journal to help ensure that scientific advancements leads to better research that is successfully incorporated into both patient-level and societal-level health care to ultimately improve the public health of societies worldwide [1].

Within this context, another key ISPOR activity has been the creation of Task Forces to establish Best Practices. The Task Force whose report appears in this issue of Value in Health has had the responsibility of reviewing and commenting on quality improvement in cost effectiveness research, a component of our field that itself incorporates many of the methods used across our field [2].

In this report, McGhan and Task Force colleagues present a masterful, comprehensive view of quality improvement for cost-effectiveness research that sets the standard for future such reports. They efficiently review historical development (including recent advancements), current status, present limitations and gaps, and appropriate improvements to be made with respect to:

- Guidelines
- Statistics and scientific methodologies
- Journals and publication quality
- Decision-makers, practitioners and evidence-based medicine

Appropriately, nearly half the Report, aside from the Task Force’s Recommendations, focuses on statistical and scientific issues. The authors have been attentive to the armamentarium of methods including new developments—as well as their strengths and weaknesses. Identifying both the inappropriate use of methods and the use of inappropriate methods, the Task Force presents its recommendations.

In this regard, reference is frequently made to the recommendations made by other ISPOR Task Forces (TF) that have developed Good Research Practices. For example, when describing the issues of concern related to the representativeness and value of data from multinational clinical trials, and commonly used, but unsatisfactory, solutions, this Report refers readers to the recommendations of the TF on Economic Data Transferability. In particular, it highlights three proposed statistical methods for consideration [3].

Indeed, one of the main contributions of this Report may be the attention to developments in methods considered across the field of health economics and outcomes research, and providing them some visibility for use in cost effectiveness research. The work of the Task Force on Good Research Practices in Modeling Studies is also cited with respect to specific methods improvements to be made—encouraging cross-fertilization across the field [4].

For the Future

The need for continuous quality improvement across our field, and specifically in cost effectiveness research, will never end. This applies to the science, and its publication and presentation at scientific meetings as well as to its use by policy decision-makers.
Although ISPOR typically closes Task Forces when their work is completed, it is safe to say that the work of this particular group should not end. This Task Force should be maintained, or at least resurrected every couple years, to reevaluate: Where is the field? How are we doing? What are our main challenges? Given where we are, what paths should we follow next?

Only by continuous methodological improvement will we make cost effectiveness research specifically, and pharmacoconomics and outcomes research in general, as valuable as they can be to researchers, practitioners and policy decision-makers. As long as we continue to improve the field, we build our ability to confront effectively and efficiently the evolving challenges our countries face, and make it harder for the skeptics to minimize its importance and its value.

Source of financial support: None.

References


William F. McGhan, PharmD, PhD, 1 Maiwenn Al, PhD, 2 Jalpa A. Doshi, PhD, 3 Isao Kamae, MD, DrPH, 4 Steven E. Marx, PharmD, MS, 5 Donna Rindress, PhD 6

1 University of the Sciences, Philadelphia, PA, USA; 2 Erasmus University, Rotterdam, The Netherlands; 3 University of Pennsylvania, Philadelphia, PA, USA; 4 Keio University, Fujisawa, Japan; 5 Abbott Laboratories, Chicago, IL, USA; 6 BioMedCom Consultants Inc., Montréal, Québec, Canada

ABSTRACT

Objectives: The International Society for Pharmacoeconomics and Outcomes Research (ISPOR) Health Science Policy Council recommended and the ISPOR Board of Directors approved the formation of a Task Force to critically examine the major issues related to Quality Improvement in Cost-effectiveness Research (QICER). The Council’s primary recommendation for this Task Force was that it should report on the quality of cost-effectiveness research and make recommendations to facilitate the improvement of pharmacoeconomics and health outcomes research and its use in stimulating better health care and policy. Task force members were knowledgeable and experienced in medicine, pharmacy, biostatistics, health policy and health-care decision-making, biomedical knowledge transfer, health economics, and pharmacoeconomics. They were drawn from industry, academia, consulting organizations, and advisors to governments and came from Japan, the Netherlands, Canada and the United States.

Methods: Face-to-face meetings of the Task Force were held at ISPOR North American and European meetings and teleconferences occurred every few months. Literature reviews and surveys were conducted and the first preliminary findings presented at an open forum at the May 2008 ISPOR meeting in Toronto. The final draft report was circulated to the expert reviewer group and then to the entire membership for comment. The draft report was posted on the ISPOR Web site in April 2009. All formal comments received were posted to the association Web site and presented for discussion at the Task Force forum during the ISPOR 14th Annual International Meeting in May 2009. Comments and feedback from the forums, reviewers and membership were considered in the final report. Once Task Force consensus was reached, the article was submitted to Value in Health.

Conclusions: The QICER Task Force recommends that ISPOR implement the following:

• With respect to CER guidelines, that ISPOR promote harmonization of guidelines, allowing for differences in application, regional needs and politics; evaluate available instruments or promote development of a new one that will allow standardized quantification of the impact of CER guidelines on the quality of CER studies; report periodically on those countries or regions that have developed guidelines; periodically evaluate the quality of published studies (those journals with CER guidelines) or those submitted to decision-making bodies (as public transparency increases).

• With respect to methodologies, that ISPOR promote publication of methodological guidelines in more applied journals in more easily understandable format to transfer knowledge to researchers who need to apply more rigorous methods; promote full availability of models in electronic format to combat space restrictions in hardcopy publications; promote consistency of methodological review for all CER studies; promote adoption of explicit best practices guidelines among regulatory and reimbursement authorities; periodically update all ISPOR Task Force reports; periodically review use of ISPOR Task Force guidelines; periodically report on statistical and methodological challenges in HE; evaluate periodically whether ISPOR’s methodological guidelines lead to improved quality; and support training and knowledge transfer of rigorous CER methodologies to researchers and health care decision-makers.

• With respect to publications, that ISPOR develop standard CER guidelines to which journals will be able to refer their authors and their reviewers; lobby to establish these guidelines within the International Committee for Medical Journal Editors (ICMJE) Requirements to which most journals refer in their Author Instructions; provide support in terms of additional reviewer expertise to those journals lacking appropriate reviewers; periodically report on journals publishing CER research; periodically report on the quality of CER publications; and support training and knowledge transfer of the use of these guidelines to researchers and reviewers.

• With respect to evidence-based health-care decision-making, that ISPOR recognize at its annual meetings those countries/agencies/private companies/researchers using CER well, and those practitioners and researchers supporting good patient use of CER in decision-making; and promote public presentation of case studies of applied use of CER concepts or guidelines.

Keywords: cost-effectiveness, guidelines, health economics, quality improvement.

Background to the Task Force

The International Society for Pharmacoeconomics and Outcomes Research (ISPOR) Health Science Policy Council recommended that the ISPOR Board of Directors establish a Task Force to critically examine the major issues related to Quality Improvement in Cost-effectiveness Research (QICER) in July 2005. The Council’s primary recommendation for this new Task Force was that it should report on the quality of cost-effectiveness research and make recommendations to facilitate the improvement of pharmacoeconomics and health outcomes research and its use in stimulating better health care and policy. The ISPOR Board of Directors approved creation of the Task Force in December 2005. An email was sent to all ISPOR members in March 2006 seeking candidates interested in serving on the leadership group or the expert reviewer group. Task Force leadership and reviewer groups were finalized by October 2006. Task Force members were knowledgeable and experienced in medicine, pharmacy, biostatistics, health policy and health care decision-making,
Quality Improvement in Cost Effectiveness Research

biomedical knowledge transfer, health economics and pharmaco-economics. They were drawn from industry, academia, consulting organizations, and advisors to governments and came from Japan, The Netherlands, Canada and the United States.

Face-to-face meetings of the Task Force were held at ISPOR North American and European meetings and teleconferences occurred every few months. Literature reviews and surveys were conducted and the first preliminary findings presented at an open forum at the May 2008 ISPOR meeting in Toronto. The final draft report was circulated to the expert reviewer group in March 2009 and then to the entire membership for comment in April 2009. The draft report was posted on the ISPOR web site in April 2009; 26 responses were received. All formal comments received were posted to the ISPOR Web site and presented for discussion at the Task Force forum during the ISPOR 14th Annual International Meeting in May 2009. Comments and feedback from the forums, reviewers and membership were considered in the final report. Once Task Force consensus was reached, the article was submitted to Value in Health.

Introduction

Quality assessment and continuous quality improvement has long been recognized as a vital process in all societal systems and organizations. In health care, critical review of interventions and reports on the quality of outcomes can help correct deficiencies and further advance efficiency and quality. Continuous quality improvement is integral to our global efforts to improve the economics and quality of life in all health care sectors and all patient populations. There is an important role for ISPOR in macro review and examination of quality and trends in pharmaco-economics, health care economics research, and their resulting impact on global policies and practice.

Mission

The mission of the ISPOR Task Force on Quality Improvement in Cost-effectiveness Research (QICER) is to generate periodic quality reports and make recommendations to facilitate the improvement of pharmacoconomics and health outcomes research, and its use in stimulating better health care and policy. This will be accomplished through periodic systematic reviews and surveys. The results and findings will be made available to ISPOR membership for comments and published as white papers and reports, including recommendations for future ISPOR initiatives, educational programs, and member services.

In this first report, the task force has focused primarily on cost-effectiveness research (CER). While broader topics in health economics and outcomes research (HEOR), such as patient reported outcomes, health-related quality of life, training, software, etc., are beyond the range of this first report, they are envisioned as targets for future work. The HEOR scope was, however, considered in our discussion of journals to capture a more holistic view of the current state of peer-reviewed publication and to position CER within a broader perspective.

As summarized in Figure 1, organizationally and individually, we embrace patients, providers, researchers, regulators, and payers to collectively advocate that scarce health care resources are allocated wisely, fairly, and efficiently. These health sector linkages are promoted through organizational services that facilitate education, communications, research, and international cooperation. Continuous quality improvement efforts are vital for improving activities and outcomes from international policies to individual patient care. The outside ring in the diagram depicts the classic phases for quality improvement that include: developing guidelines, designing guideline implementation, conducting interventions, measuring impact, analyzing outcomes, and the feedback for improved guidelines.

The Role of Guidelines in Quality Improvement

What role do guidelines play in promoting the quality and improvement of CER? It is usually assumed that the presence of guidelines leads to quality improvement, assuming that established guidelines increase credibility and usefulness by defining generally accepted standards and the requirements of specific users. However, in this field, there is not much evidence to support or disprove this assumption. A number of studies have evaluated the quality of research, but few have examined the relationship between the presence of guidelines and the quality of research. Two topics were examined: the availability of HEOR guidelines and the impact of guidelines on the quality of CER.

A few authors have reviewed available guidelines, comparing and contrasting them [1–3], and other resources are also available at: http://www.ISPOR.org; http://www.biomedcom.org/en/resources-BMC-databases.html. Most guidelines have similar content with minor variations, but some significant differences do exist among them, most generally because of their intended purpose, the audience to be addressed, regional, cultural or political variation, or author or sponsor preferences.

How can the impact of guidelines on CER quality be measured? Most journals neither have specific guidelines or requirements for HEOR (see further discussion), nor do authors normally reveal which guidelines, if any, they observe. So despite easy accessibility of published articles, their quality, and the improvement of their quality, is not easily linked to specific guidelines. Those who measure the quality of HEOR publications [1,4–8] choose their own quality measures from those guidelines currently available and generally accepted standards.

Several formulary evaluation bodies (such as NICE in the UK [9], Canadian Agency for Drugs and Technology in Health (CADTH) in Canada [10], and PBS in Australia [11–13]) have developed specific guidelines and requirements for HEOR studies submitted, but these studies are often not publicly available for evaluation in the short term. Some of these bodies have performed, or allowed, quality evaluation of studies submitted to them, which have been presented publicly [14] or published [15,16]. These often use small samples, are qualitative, and more importantly, not easily comparable across jurisdictions.

The Evolution of Guidelines

Guideline development began in Australia in 1992, followed closely by Canada and a few academic groups in the United States [17–19]. Their form has tended to reflect their purpose, e.g., those intended for reimbursement decision-making tend to be more prescriptive, while those with more academic purposes more descriptive (discussion of the appropriateness of each will be left to a future report focused specifically on guidelines).

Over the last decade, many countries have produced their own guidances and others are in development. There are currently about 39 CER guidances from 34 countries (with multiples from some countries). These have been produced by government bodies, by academic groups, and health-care insurers, and fre-
quently as collaborations among all three groups. Over half of these were prepared as part of formulary submission guidances or requirements. Table 1 summarizes those guidelines of which we are currently aware. The task force welcomes all feedback on publicly available new guidelines or updates not included here.

Measuring and Improving Guideline Quality

No instrument has been found that permits quantitative measurement of guideline quality, or comparison of guidelines. What publicly available information there is has been rather qualitative; for example, in the early days of guideline development, Regence BlueShield found that guidelines were practical from a logistics perspective, improving the relevance and timeliness of information available for decision-makers [54].

Although guidelines have not been quantitatively evaluated, there are some studies published evaluating the quality of studies submitted to the guideline-producing bodies. In one of these, Hill et al. evaluated 326 Australian studies and found significant issues in the interpretation of guidelines and in the conduct of studies [16]. In another, Colmenero et al. analyzed 53 economic studies submitted to US-managed care organizations and found low levels of compliance with accepted standards of CER research [55]. One pilot study did measure the relationship between specific guidelines and the quality of studies submitted [56]. Unfortunately, among published studies of the quality of studies, the measures used are not easily comparable.

If the impact of guidelines on quality of studies, and improvement of quality, is to be assessed rigorously, a tool is required that is quantifiable, anchored to guidelines, and to generally accepted practices, and comparable across guidelines, studies, and time. There does not appear to be an instrument that yet fits this objective, but possibly, one might be developed incorporating the most relevant aspects of existing tools. Some already in use that might be considered include: 1) Neumann et al. measured the quality of economic analyses in several studies over the last decade [5,6]; 2) Chiou et al. developed a grading system to measure the value and quality of CER analyses using the QHES instrument [1]; and 3) Goetghebeur et al. evaluated quality of 10 submissions to the Canadian Common Drug Review, assessing the studies submitted with respect to the CADTH requirements [56,57].

Future Work

Based on this preliminary review, there are a number of promising steps that could be taken with regard to guidelines as instruments in measuring and improving quality in CER:

1. Perform a formal evaluation of currently available instruments that might be used to quantify and compare the value of CER guidelines. If none are found that meet this need, adapt, develop, or promote the development of one.
2. Once an instrument is available to assess guideline quality, promote the harmonization of guidelines, allowing for differences in application, regional needs, and politics.
Joint comparison of costs and effects and estimation of sampling uncertainty. A joint comparison of costs and effects using the incremental cost-effectiveness ratio (ICER) or the incremental net monetary (health) benefit is a useful decision tool to help determine whether the new therapy offers good value relative to the alternative. The use of this tool is particularly important when there is a trade-off between costs and effects; that is, one therapy is both significantly more effective and more costly compared with the other therapy. If there is no trade-off between costs and effects, that is, when one therapy is significantly more effective and less costly when compared with the other therapy, this decision tool may not be necessary because therapy is unambiguously dominant over its alternative. A third possibility occurs when the two treatments have the same effect. In this case, some authors have interpreted textbooks and guidelines on the use of ICER or net monetary benefit as sufficient (i.e., the lowest cost treatment is the treatment of choice) and there is no need to perform a joint comparison of costs and effects. Nevertheless, as our understanding of sampling uncertainty for the comparison of costs and effects has grown, the cases where this interpretation is appropriate is not as common as previously thought. Because cost-effectiveness ratios and net monetary benefit estimated from trial data are the result of samples drawn from the population, one should report the uncertainty in this outcome that derives from such sampling. Identification of methods such as confidence intervals for cost-effectiveness ratios [65–68], acceptability curves [69], and confidence intervals for net monetary benefit [70] for the measurement of this uncertainty have been important methodologic developments in the economic evaluation of medical therapies [71,72]. As a result of uncertainty, the cost-minimization approach has been shown to be rarely appropriate as a method of analysis and the need for a joint comparison still remains under most circumstances [73]. Alternatively, observing no significant difference in costs and effects alone need not rule out that one can still be confident that one of the two therapies is good value because it is possible to have more confidence in the combined outcome of differences in costs and effects than in either outcome alone. In these cases, one should jointly compare costs and effects, and one should report on their sampling uncertainty.

**Analysis of cost data.** For all economic analysis calculations, costs and cost differences between treatment groups should be expressed by the use of the arithmetic mean, and not medians, because this summary measure permits a budgetary assessment of treatment (N × arithmetic mean = total cost) and is the statistic of interest for health-care policy decisions [58]. Because of the often highly skewed distribution of cost data, the normality assumption underlying the parametric t test is often called into question and standard nonparametric tests (e.g., Mann-Whitney U-test or Wilcoxon rank-sum test), or parametric tests on normalizing transformations (e.g., log transformation) are often

3. Concurrently, evaluate available instruments or promote development of one to quantify the impact of CER guidelines on the quality of CER studies. The outcomes of such an instrument should allow comparison across guidelines as well as comparison over time, to allow rigorous longitudinal evaluation of quality improvement.

4. Encourage adoption of and adherence to CER guidelines through training and knowledge transfer.

**Statistics and Science**

**Introduction**

There are a number of statistical issues in CER, which could benefit from standard approaches to address these issues. Recommendations for improving the statistical methods applied in the cost-effectiveness literature are presented. The statistical issues and their solutions are discussed separately for the two most common approaches to cost-effectiveness analyses, namely, clinical trial-based economic evaluations and decision modeling-based studies. Economic evaluations based on registry and administrative data sets have started gaining popularity in recent years, and several of the issues discussed herein also apply to such analyses. However, a comprehensive discussion of all the statistical issues and approaches to address such issues that arise in analyses of nonrandomized observational data are beyond the scope of this first article.

**Statistical Issues in Clinical Trial-Based Economic Evaluations**

A critical source of the evidence on costs and cost-effectiveness of new medical treatments comes from analyses of patient-level data on cost and effect collected as part of randomized clinical trials. The number of clinical trial-based economic evaluations has increased considerably over the last decade. In the same time frame, the field has matured substantially, including the advance-ment of, and a growing consensus about, appropriate statistical methods for analysis of costs and cost-effectiveness alongside clinical trials [58]. Systematic reviews suggest that published studies on clinical trial based economic evaluations have begun to use some of these new statistical techniques [59,60]. Nevertheless, there are still a substantial number of studies using statistical methods of poor quality. In addition, there still remain areas needing further research.

<table>
<thead>
<tr>
<th>Table 1 Countries, regions, or groups with publicly available health economic research guidances</th>
</tr>
</thead>
<tbody>
<tr>
<td>Source</td>
</tr>
<tr>
<td><strong>Australia</strong></td>
</tr>
<tr>
<td>The Baltics (Latvia, Estonia &amp; Lithuania)</td>
</tr>
<tr>
<td>Austria</td>
</tr>
<tr>
<td>Belgium</td>
</tr>
<tr>
<td>Brazil</td>
</tr>
<tr>
<td>Canada (CADTH, Ontario)</td>
</tr>
<tr>
<td>China</td>
</tr>
<tr>
<td>Finland</td>
</tr>
<tr>
<td>France</td>
</tr>
<tr>
<td>Germany</td>
</tr>
<tr>
<td>Hungary</td>
</tr>
<tr>
<td>Ireland</td>
</tr>
<tr>
<td>Israel</td>
</tr>
<tr>
<td>Italy</td>
</tr>
<tr>
<td>Mexico</td>
</tr>
<tr>
<td>The Netherlands</td>
</tr>
<tr>
<td>New Zealand</td>
</tr>
<tr>
<td>Norway</td>
</tr>
<tr>
<td>Poland</td>
</tr>
<tr>
<td>Portugal</td>
</tr>
<tr>
<td>Russian Federation</td>
</tr>
<tr>
<td>Scotland</td>
</tr>
<tr>
<td>Singapore</td>
</tr>
<tr>
<td>Slovak Republic</td>
</tr>
<tr>
<td>South Korea</td>
</tr>
<tr>
<td>Spain</td>
</tr>
<tr>
<td>Sweden</td>
</tr>
<tr>
<td>Switzerland</td>
</tr>
<tr>
<td>Taiwan</td>
</tr>
<tr>
<td>UK (England and Wales)</td>
</tr>
</tbody>
</table>
used as a substitute. Yet these popular alternatives are not appropriate for drawing statistical inferences on differences in arithmetic mean costs [74–76]. For example, when one uses a t test to evaluate the log of costs, the resulting P value has direct applicability to the difference in the log of costs and to the difference in the geometric mean of costs. It may or may not be directly applicable to the arithmetic mean costs. A Mann-Whitney U test tests differences in the median of costs. Thus, statistical inferences about these other statistics may not be representative of inferences about the differences in arithmetic mean, which is the statistic of interest. If one does not want to adopt a parametric t test to directly test for differences in arithmetic mean costs, one can compare the arithmetic means by using a nonparametric bootstrap. This procedure has the added advantage of avoiding a parametric assumption about the distribution of costs. As a result, the nonparametric bootstrap has increasingly been recommended either as a check on the robustness of standard parametric t tests, or as the primary statistical test for making inferences about arithmetic means for moderately sized samples of highly skewed cost data [76–78].

Many clinical trial-based economic evaluations are limited to univariate analyses of costs. Even if treatment is assigned in a randomized setting, there are advantages to using multivariable techniques to analyze costs. Multivariable analysis of costs may be superior to univariate analysis because it improves the power for tests of differences between groups (by explaining variation due to other causes). It also facilitates subgroup analyses for cost-effectiveness, for example, more and less severe; different countries/centers, etc. Finally, it accounts for potentially large and influential variations in economic conditions and practice patterns by provider, center, or country that may not be balanced by randomization. Adoption of multivariable analysis does not, however, avoid the issues that arise in the univariate analysis of cost. For example, regressions on the logarithmic transformation of costs were previously considered an ideal remedy to the violation of the assumption of normally distributed error term that underlies ordinary least squares regression. Nevertheless, as the shortcomings of multiple regression models of log transformed costs became more widely publicized [75], the use of the generalized linear models have become the more acceptable alternative [79–81].

Handling of censored cost data. Incomplete or censored cost data occur in most randomized trials that follow participants for clinically meaningful lengths of time; yet they are often not addressed in the analysis. The ISPOR RCT-CEA taskforce recommended that “ignoring small amounts of missing data is acceptable if a reasonable case can be made that doing so is unlikely to bias treatment group comparisons” [58]. However, no clear guidance exists for how much censoring is too much. Hence, whether or not cost data were incomplete, the amount of incomplete data and the statistical method adopted to address any problems posed by incomplete data should routinely be reported in trial-based analyses [58] Many studies in the literature have adopted naive approaches wherein censored observations are either excluded from analysis (i.e., complete-case analysis) or included as though they were complete observations, (i.e., full sample analysis). In the first naive approach, only the uncensored cases are used in the estimation of mean cost and this method is biased toward the costs of the patients with shorter survival times because patients with larger survival times are more likely to be censored [82,83]. Also, completely discarding patients with censored data can lead to the loss of information and statistical power, which can be problematic if the percentage of censored cases is high. The second naive approach that uses all cases without differentiating between censored and uncensored observations is always biased downward, because the costs incurred after censoring times are not accounted for [83]. Although there exists a mix of approaches to impute the cost data, recent statistical interest in addressing censored cost data has led to the proposal of several methods of estimation that explicitly account for incomplete cost data due to loss-to-follow-up [82,84–95]. It is well established that these methods are prone to less bias and return a better estimate of sampling variance than other naive estimation methods [82,83,87,89,96–98].

Sample size and power. Prior to the development of methods for assessing sampling uncertainty for the joint comparison of cost and effect, health economists commonly attempted to estimate sample size based on the larger of the sample sizes needed for estimating prespecified cost and effect differences — i.e., what sample size was required to identify a $1000 difference in costs, and what was required to identify a 10% reduction in mortality. With the development of methods for assessing uncertainty, sample size calculations should now be based on the sample size needed to rule out the net monetary benefits of the intervention are less than zero [99–102]. Often economic evaluations are piggy-backed on clinical trials with a prespecified sample size. In such instances, researchers should estimate and report the power available to rule out cost-effectiveness ratios that exceed the maximum willingness to pay.

Evaluating transferability (generalizability) of trial results. Multinational clinical trials are the norm for the evaluation of new medical therapies. However, the presence of between-country heterogeneity in trials has led to a growing concern that the pooled or average economic results from multinational trials may not be reflective of the results that would be observed in individual countries that participated in the trial [103]. Common sources for concern about the representativeness of data from multinational trials include transnational differences in morbidity/mortality patterns, practice patterns (i.e., medical service use); and absolute and relative prices for medical service use (i.e., price weights). The use of trial-wide clinical results, trial-wide medical service use, and price weights from a single country has been one of the commonly proposed, potentially inadequate solutions to the problem of transferability (e.g., to tailor the results to the UK, simply use UK price weights, and conduct the analysis as if all participants were treated in the UK). A second potentially inadequate solution has been to use trial-wide clinical results, and country-specific medical service use and price weights. Both approaches have the failing that they ignore the fact that clinical and economic outcomes may influence one another. That is, differences in cost may affect practice patterns, which in turn may affect outcome; differences in practice pattern may affect outcome, which in turn may affect cost. The ISPOR Good Research Practices task force on Economic Data Transferability has recently recommended good research practices for dealing with aspects of transferability including three proposed statistical methods that use patient-level data to address transferability, detection of heterogeneity [104,105], fixed effects models [106,107] and multilevel, or hierarchical models [108–115].

Decision Models

The estimation of the full economic effects of health technologies generally requires the extrapolation of clinical trial evidence beyond the follow-up period through the use of decision modeling techniques to synthesize data from various sources. The aim
of the modeling study is to aid decision-makers in making decisions under uncertainty. Obviously, the results of modeling studies will only be helpful to decision-makers if the study is performed according to current standards. While the quality of cost-effectiveness analyses has improved over time, still current studies do not address all issues appropriately [5,113]. It is clear that guidance is important for those performing modeling studies. In several countries, authorities have formulated guidelines, and ISPOR has also published guidelines through the task force on Good Research Practice in Modeling Studies [116]. Some of the most important issues where quality might be improved, and some new methodological topics that have emerged in the past few years, are discussed.

**Methods for evidence synthesis.** The ISPOR task force on Good Research Practice in Modeling Studies suggested in their report that systematic review should be conducted on key model inputs. There are various ways of synthesizing the evidence found in various studies (e.g., fixed or random effects meta-analysis, either frequentist or Bayesian), but there is not one optimal method of synthesizing data currently available [117], and the typical meta-analyses cannot straightforwardly be applied to synthesize data for cost-effectiveness models.

One reason is that meta-analysis has been developed to combine quantitative results of several similar studies into a pooled estimate of the treatment effect (e.g., odds ratio, relative risk, difference in change from baseline). It uses the magnitude of the treatment effect and its uncertainty from each individual study to produce a weighted mean of the treatment effect [118,119]. However, in modeling studies, the parameter to be estimated is not only a treatment effect like the odds ratio of having an event. Typically, models contain parameters like transition probabilities between disease states, event probabilities, rate ratios of treatment effects, quality of life or utility values, and costs. These parameters have different distributions which need to be combined. Moreover, the comparator needs to be modeled too, meaning that we are dealing with more heterogeneity than usually remains after the variance in treatment effect has been corrected for the variance in comparator-effect.

Second, meta-analyses have traditionally been performed on studies that compare the same intervention with the same comparator. However, comprehensive decision analytic models aim to identify the most cost-effective treatment among the entire spectrum of all relevant treatment options. This issue may be dealt with through so-called mixed treatment comparisons, which combine multiple different pairwise comparisons across a range of different interventions [120–123]. In mixed treatment comparisons, the relative effect of a treatment compared to a range of alternatives is estimated by including indirect comparisons of two interventions through a common comparator. Evidence from direct and indirect comparisons are analyzed simultaneously, which allows estimates of treatment effects in the absence of head-to-head comparisons. Such mixed treatment comparisons are inevitable in modeling cost-effectiveness.

The fact that the choice between fixed and random effects model and between a Bayesian and a frequentist approach can have a large impact on the outcome of the model underlines the need for complete transparency in the reporting of a modeling study.

**Probabilistic sensitivity analysis.** Sensitivity analysis should always be an integral part of a modeling study: input parameters of the model are varied to see if and how the outcome changes. For a long time, the common type of sensitivity analysis was a deterministic one-way or multi-way analysis, in which one or more (usually not more than three) parameters are varied between certain limits. In the last few years, however, more and more studies include a probabilistic sensitivity analysis, in which (ideally) all input variables are varied simultaneously, according to probability distributions [124]. Such an analysis presents information on all possible outcomes, as well as on the likelihood of these outcomes.

In their report, the ISPOR task force on Good Research Practice in Modeling Studies stresses the need to include a sensitivity analysis as part of the modeling study. While those task force guidelines mentions that deterministic and probabilistic analyses are equally appropriate, other guidelines such as the UK and Dutch guidelines strongly prefer a probabilistic sensitivity analysis as a way to correctly represent parameter uncertainty. However, there are still good reasons to also include deterministic sensitivity analysis in a modeling study, such as to account for other types of uncertainty such as uncertainty relating to the structure and assumption of decision models [125]. Series of sensitivity analyses may be done to look at the consequence of changing different assumptions and scenarios in a model. Recently, however, an article was published exploring the option of explicitly incorporating structural uncertainties into the model [126].

**Value-of-information.** In the last few years, much attention has been given to Value-Of-Information (VOI) analysis [127,128]. This type of analysis addresses the question of what the value is of collecting additional information to eliminate or reduce uncertainty, since making the wrong decision comes with a cost that is equal to the benefits forgone because of the wrong decision. These expected costs of uncertainty can be determined by 1) the probability that a decision based on the current ICER is wrong; and 2) the size of the opportunity loss if the wrong decision is made.

The first step in a VOI analysis is the estimation of the expected value of perfect information, which is the maximum amount the decision-maker should be willing to pay to eliminate all uncertainty in the decision. The next step is to calculate the expected value of partial perfect information, which is the maximum amount the decision-maker should be willing to pay to eliminate all uncertainty on one parameter or subset of parameters [129]. Based on the latter analyses, priorities may be set for further research. The final part of the VOI is to calculate the expected value of sampling information, which is the maximum amount the decision-maker should be willing to pay to reduce uncertainty through a sample of a certain size and to set this against the costs of obtaining that sample [130].

VOI analysis is important in situations where a decision-maker might prefer to postpone making a decision to collect more information. Basing that decision merely on the outcomes of sensitivity analyses would lead to suboptimal decision-making. As suggested by the ISPOR task force on Good Research Practice in Modeling Studies "The decision to obtain additional data to inform a model should be based on a balance between the expected value of the additional information and the cost of the information."

**Model validation.** The final part of any model development should concern validation. Several types of validation may be distinguished, among which are: face validity, internal validation, between-model validation, predictive (or prospective) validation, and external validation [116,131].

Face validity means that the results produced by a model look valid at first inspection. If for example a model for pneumococcal vaccination results in so many cases of otitis media that every
child under four would have at least four episodes per year, one might question this result. Such unexpected results are always reason to thoroughly check the model and the inputs.

Internal validation concerns the comparison of model outputs with data used in the model development. While this type of validation is straightforward if the model is based on one source of data, it becomes more complicated if the model is based on a synthesis from various sources of data. It is possible that no model input can be chosen so that the model validates well against each separate data source. However, good effort should at least be made to describe the deviations from the data sources and possible explanations.

Between-model validation involves comparison of the current model and published or publicly available models. If the results differ, an attempt should be made to clarify whether these discrepancies are because of difference in model structure or model input.

Predictive validation aims to compare model results to newly available data from the same data source that was used as model input. On the other hand, external validation concerns the comparison of model results to data from studies not used in the model development. These types of validation are not always possible; if a model contains all data currently available, there is no data source for external validation. As the ISPOR task force on Good Research Practice in Modeling Studies remarks, “...it is not necessary that every data estimate or structural assumption be tested in prospective studies, in advance of model use.” However, they also stress that models should never be regarded as immutable. They should be updated and possible abandoned as new evidence becomes available to inform structure or input values. If models are inconsistent with the new evidence but have not been amended to calibrate against this evidence, the model should be abandoned until such recalibration has been finished.

The validation phase of model development should be seen as important as all other phases. Researchers should document their validation procedure, and report it in their publications.

How Can We Make the Science Better?

The International Society for Pharmacoeconomics and Outcomes Research has published best practices document for the design, conduct, and reporting of economic analyses alongside clinical trials as well as decision modeling studies [58,116]. Whether explicit guidelines alone will foster improvements in the quality of future studies remains a question, given evidence that such guidelines alone will foster improvements in the quality of subsequent studies [3,64,132,133].

Part of this problem may be that most of the advances in the statistical techniques for analyzing cost data have been published in highly technical economic or biostatistics journals. Although some applied researchers may not be reading such literature, many may have difficulty understanding the rationale for and implementation of these technical methods. There is clear need for publication in more applied journals that focus on explaining these technical advances in an easily understandable format to support knowledge transfer to researchers who need to apply these methods. In addition, there is an equally clear need for better training and education so that researchers can understand more sophisticated techniques. Another problem in raising the quality of studies is space. Transparency of a model is of utmost importance for modeling studies to be taken seriously, both within the health economics community and among clinicians and decision-makers. To achieve full transparency, the model or detailed model description should be made available in electronic format, since it is often not possible to describe every detail of the model sufficiently in an article because of space restrictions.

Additional efforts to improve the quality of future studies may involve providing tools to peer reviewers for both funding agencies and journals so as to identify studies that fail to apply best practices. For example, peer reviewers might be provided with a clear check list of all requirements. Thus, all studies are reviewed consistently, hopefully leading to an increased quality. Regulatory and reimbursement authorities should also explicitly adopt best practices guidelines and uphold all economic data submissions to these high standards while making reimbursement decisions.

Journals and Publication Quality

Journal publication plays a critical role in quality improvement of CER research. This can be done by establishing requirements and guidelines for the conduct and reporting of the various types of studies that comprise the field, through the peer-review process, by dissemination of studies, by peer feedback, and as an ongoing learning process for researchers. Although published work may include abstracts, posters and podium presentations, newsletters and other non peer-reviewed publications (such as educational texts, patient information and marketing materials), for the purposes of this first report the focus has been on peer-reviewed journal publications because these are most accessible and easiest to track.

There are a great many journals globally. To determine how many of these routinely accept and publish articles relevant to HEOR, and how many of these provide or require guidelines for the conduct and reporting of these studies, a survey of the World Association of Medical Editors (WAME; http://www.wame.org) was carried out. This organization represents more than 965 biomedical journals, from more than 91 countries, from all geographic regions of the world. As such, WAME was an ideal source of information relevant to quality improvement in HEOR publication. In the survey which all WAME members were invited to participate, they were asked about journal type, location, scope, circulation, whether they accepted HEOR articles and which types of studies, and how they found reviewers for this type of work. They were also asked whether any HEOR or CER guidelines were recommended or mandatory for authors or reviewers, and if so, which ones.

Of the 965 journals represented in WAME, 55 (6%) responded to the survey. These came from 29 countries and all continents, with 45% representation for North America and Europe. Almost all (98%) were peer reviewed, and the majority (72%) international in readership. Most respondents (83%) were high-level editorial staff. Journal readerships encompassed clinical and academic health-care researchers (76% of respondents), health-care decision-makers, health service researchers, and medical generalists, and specialists (50–67%), health-care policymakers (40%), and other types of readers (37%) (students, patients, or the general public, the paramedical professions, other areas of academia).

Part of this problem may be that most of the advances in the statistical techniques for analyzing cost data have been published in highly technical economic or biostatistics journals. Although some applied researchers may not be reading such literature, many may have difficulty understanding the rationale for and implementation of these technical methods. There is clear need for publication in more applied journals that focus on explaining these technical advances in an easily understandable format to support knowledge transfer to researchers who need to apply these methods. In addition, there is an equally clear need for better training and education so that researchers can understand more sophisticated techniques. Another problem in raising the quality of studies is space. Transparency of a model is of utmost importance for modeling studies to be taken seriously, both within the health economics community and among clinicians and decision-makers. To achieve full transparency, the model or detailed model description should be made available in electronic format, since it is often not possible to describe every detail of the model sufficiently in an article because of space restrictions.

Additional efforts to improve the quality of future studies may involve providing tools to peer reviewers for both funding agencies and journals so as to identify studies that fail to apply best practices. For example, peer reviewers might be provided with a clear check list of all requirements. Thus, all studies are reviewed consistently, hopefully leading to an increased quality. Regulatory and reimbursement authorities should also explicitly adopt best practices guidelines and uphold all economic data submissions to these high standards while making reimbursement decisions.

Additional efforts to improve the quality of future studies may involve providing tools to peer reviewers for both funding agencies and journals so as to identify studies that fail to apply best practices. For example, peer reviewers might be provided with a clear check list of all requirements. Thus, all studies are reviewed consistently, hopefully leading to an increased quality. Regulatory and reimbursement authorities should also explicitly adopt best practices guidelines and uphold all economic data submissions to these high standards while making reimbursement decisions.

The vast majority (92%) of journals accepted all or some types of HEOR work. Of the 10 categories of HEOR research published by respondents (Fig 2), epidemiological burden of illness studies, database analyses and systematic reviews or meta-analyses were most commonly reported (77–79%), registry studies, clinical trials with economic or resource utilization data, economic burden of illness studies and epidemiological modeling studies were reported by about half the respondents (46–62%), and economic modeling, naturalistic clinical trials were least frequently reported (37–40%).
Although most journals recommended directly in their Author Instructions the International Committee of Medical Journal Editors (ICMJE) requirements (which includes references to no economic guidelines but some HEOR-relevant guidances [MOOSE (meta-analysis of observational studies in epidemiology), CONSORT (RCTs), STARD (studies of diagnostic accuracy), QUORUM (systematic reviews and meta-analyses) and STROBE (observational epidemiology studies)] or other specialized guidances [QUORUM, STROBE, CONSORT, STARD, and two references for basic statistics), none of the journals provided their own HEOR guidance, and only 4 of the 54 responding journals recommended the BMJ health economic study guidelines [64]. The Cochrane Web site did, however, have links to some PRO and HEOR guidances.

About 58% of journals did not provide their reviewers with any guidelines for evaluating HEOR studies. For the 42% who said they did, in all cases, these were the same as author guidelines (e.g., instructions to authors, ICMJE) and only rarely specific to HEOR.

Journals were asked if they would consider using a standard set of HEOR guidances from a recognized professional body to enhance the quality of published HEOR research in their journal; 91% said they would if these were made available.

When asked about the ease with which they found reviewers for HEOR articles, 27% of journals had great difficulty, 60% said it was difficult for some types of articles, and 6% had no difficulty. Almost 90% of respondents felt it would be useful if they had a pool of expertise available to perform reviews of HEOR for their journal. The areas of expertise specifically mentioned covered the spectrum of HEOR research: policy analysis, economic outcomes, resource utilization, clinical epidemiology, public health, preventive medicine, mental health (and other specialties), statistics, and methodologies.

Although this survey sample size was small, a fairly representative range of journals responded and several clear messages were heard. 1) Many biomedical journals accept and publish HEOR research; 2) Almost all do so without giving clear guidance to either authors or reviewers about quality standards for this type of research; so discussion of quality control at either the article production stage or the peer-review stage is currently moot. HEOR quality is currently left entirely up to the skill, knowledge, and experience of each author and reviewer; 3) Most respondents expressed interest in clear guidelines to which they could refer; and 4) many of the respondents reported difficulty finding expert reviewers for HEOR, and almost all were interested in having a larger pool of reviewers.

Future Plans
First steps in improving the quality of HEOR research in publications would be to: 1) develop standard guidelines to which journals are able to refer their authors and their reviewers; 2) lobby to establish these guidelines within the ICMJE Requirements to which most journals refer in their Author Instructions; 3) provide some form of support in terms of additional expertise to those journals without appropriate reviewers. Finally, 4) it would be worthwhile to resurvey WAME members to gauge change over time.

There are also a few research groups who from time to time perform evaluations of HEOR research quality in the published literature [4,6,8,134]. It would perhaps be more useful for these evaluations to use quantitative quality measures that might be compared across evaluations, and follow the evolution of these over time. This might be an ISPOR sponsored initiative, or one undertaken by one or more of the groups currently involved in such evaluations. Rather than reinventing the wheel, it would be worthwhile to examine possible quantitative measures already established, such as the QHES (Quality of Health Economic Studies) instrument [1,7]. A search for other instruments already in development is recommended.

Finally, once standardized, quantitative quality measures have been established, we recommend ongoing assessment of the quality of published HEOR and its reporting through longitudinal sampling of the literature and other publications, perhaps with annual or biannual reports.

![Figure 2](https://example.com/figure2.png)

**Figure 2** Types of Health Economic and Outcomes Research (HEOR) articles accepted for publication by journals responding to the QICER survey.
Decision-Makers, Practitioners and Evidence-Based Medicine

The Simplest Scheme Model (SSM)

The process of health care decision-making in the era of evidence-based medicine (EBM) can be described by the simplest scheme model (SSM) with three steps:

1. generating evidence by researchers and compiling this evidence in databases;
2. extracting and interpreting evidence from these databases by decision-makers or EBM practitioners; and
3. applying this evidence in health care settings at local levels.

The three-step scheme (SSM) is a clue to the barriers to use and application of cost-effectiveness results faced by decision-makers and clinical practitioners. The second step might result in some information loss or inappropriate interpretation, an evidence gap. Since decision-makers rarely access the full range of evidence from databases, there could be gaps between stored evidence and that extracted in forms such as partial evidence, abstracts, conclusions, executive summaries, commentaries, translations, etc. In general, practitioners of EBM and decision-makers in health care tend to regard CER as of limited use, even though it must be considered for rational resource allocation in health care. Some obstacles result from insufficient knowledge or skills in cost-effectiveness analysis. Such insufficiencies are related to the “evidence gap” mentioned above, and may lead to skepticism about cost-effectiveness research, often bringing decision-makers to doubt the quality of the studies and the data.

One classic finding on limitations comes from a study [63] which surveyed almost 800 UK decision-makers. The authors concluded that the use of health economic evidence at the local level was not extensive. The major reasons for such a limited use were the inflexibility of budgets, limiting movement of resources between primary and secondary care, and the inability to free resources to adopt new interventions. These issues are related to the third step in the SSM. It was also reported that decision-makers were concerned about the credibility of the evidence presented, specifically with respect to large numbers of assumptions and industry funding. These are validity or credibility issues associated with the first step in the SSM.

Another limitation was reported by a survey [135] which conducted interviews with 17 pairs of UK NHS decision-makers. It raised the issue of technical terms employed in pharmaco-economics. Some of these words are not in everyday language, and knowledge of these terms is essential. The same research group [135] pointed out the general assumption has been that if decision-makers do not find economic evaluations useful, then the way the evaluations are conducted or presented

The Evidence and Value: Impact on Decision-Making (EVIDEM) Model

The EVIDEM [56,57,138–142] model is somewhat more detailed than SSM. The basic framework from evidence to decision-maker is similar to that of SSM, but EVIDEM deals with more sophisticated subprocesses in the decision-making scheme, which includes:

1. Providing relevant evidence;
2. Assessing evidence for HTA report;
   a. Synthesizing evidence with weights; and
   b. Quantifying quality of evidence with quality matrix resulting in scores;
3. Evaluating intrinsic values based on Multicriteria Decision Analysis with weights and scores; and
4. Evaluating extrinsic values leading to decisions.

Discussion

As for economic evaluations and decision-making framework, one study [143] was conducted by drawing on decision-makers from two UK health authorities, Leicestershire and North Yorkshire, employing the National Health Service Economic Evaluation Database (NHS EED) [144] as a research vehicle. It confirmed that decision-makers generally recognize the usefulness and necessity of published cost-effective evidence in informing their decision-making processes. However, they often regarded the value of studies limited because of poor generalizability, narrowness of research questions, and lack of methodological rigor, all of which are common seen in published articles. As reported previously [136], using trusted sources to appraise studies encouraged decision-makers to use CER studies, as well as having a quality-scoring system for published studies and not just the critical summaries from NHS EED.

The same research group [143,145] argued that there could be two approaches to addressing the problem: changes to the performance of economic evaluation and changes to the process of decision-making in the NHS. However, they continue to concentrate on one side of the problem, exploring ways to make economic studies more accessible without losing the key elements of critical appraisal. Nevertheless, they also state that: 1) it would be productive to examine some of the potential flaws in health care decision-making; 2) they share concern about requests for quality scores which might lead to even less critical assessment of findings; and 3) cost-effectiveness analysis makes the shortcomings of the clinical data much more apparent. Regarding the quality of clinical effectiveness data in economic studies, we need to find out whether the problem is: the lack of good-quality data for economic evaluations or the lack of available good clinical data for economic studies performed in time of need. There is often a problem of a long and inexplicable lag between the publication of the first clinical data and the subsequent publication of the first cost-effectiveness study [146]. It was suggested that there should be better strategic planning connecting clinical and economic research plans.

The other study [145] pointed out the general assumption has been that if decision-makers do not find economic evaluations useful, then the way the evaluations are conducted or presented
must be changed. Also if an economic decision framework is not satisfactory for decision-makers, it must be assumed there is a different and superior model for decision-making. In addition, they questioned how decisions are made by health authorities for service planning and resource allocation without substantive economic input. To promote better use of cost-effective evidence among NHS decision-makers, it was suggested:

- changing the process for NHS decision-making;
- changing policies on funding local economic studies and expertise; and
- using decision-makers with a better grasp of economics.

They insisted that the changes above might be more appropriate than modifying economic methods to inform unclear decision process.

In an effort to better understand HEOR researchers and decision-makers, a Web-based survey was conducted of all ISPOR members, to which 122 responded. The aim of the survey was to understand the degree of incorporation of HEOR in decision-making [147].

Thirty-four percent (n = 42) of respondents identified themselves as a drug, medical device or diagnostic treatment decision-makers. Years of experience as a decision-maker were categorized as: less than one year, 1–5 years, 6–10 years, 16–20 years, or greater than 20 years. The majority of respondents had between one and five years of experience. Approximately a quarter of the respondents were members of a Pharmacy & Therapeutics Committees, and a fifth were practicing pharmacists. The majority resided in the US, followed by the United Kingdom, Germany, Canada, and Japan.

All respondents had read articles containing HE information; the majority had read articles related to other common terms in HEOR. Overall, 95% of reported receiving formal training; 90% had some training in pharmacoeconomics, and only 40% formal training regarding burden of illness and patient-reported outcomes. The term training here referred to "formal schooling, attended courses, worked with an expert or read extensively." Ninety-five percent of decision-makers believed the training helped them in selecting appropriate interventions, 86% used some form of HE analysis to make decisions. Cost-effectiveness and cost-minimization analyses were the most common analyses used. Seventy-nine percent of respondents had conducted some form of HE analysis to make decisions. Cost-effectiveness and cost-minimization analyses were the most common analyses used. Seventy-nine percent of respondents had conducted some form of HE analysis to make decisions.

In reaching out to decision-makers and practitioners, a sophisticated decision-making model such as EVIDEM should be considered. An advantage of the EVIDEM model is its inner structure that separates intrinsic and extrinsic values.

**Recommendations**

To recognize true value of CER and to establish evidence-based decisions, the QICER task force recommends that ISPOR implement the following:

- Provide the right information at the right time in health care delivery;
- Translate clinical research into clinical decisions and actions; and
- Promote communications with practitioners and decision-makers who are faced with the challenges as how to better understand complex socioeconomic evaluations, how to improve the decisions, and how to seek the rationality of reasoning.

**QICER Final Report Recommendations**

The QICER Task Force recommends that ISPOR implement the following:

**Guidelines**

- Promote harmonization of CER guidelines, allowing for differences in application, regional needs and politics;
- Evaluate available instruments or promote development of one to quantify the impact of CER guidelines on the quality of CER studies;
- Report periodically on available guidelines;
- Evaluate periodically the quality of studies submitted to decision-making bodies (as public transparency increases) and journals (preferably with standardized CER guidelines); and
- Develop trainings for researchers and others using CER guidelines, and promote effective mechanisms of continuous knowledge transfer.

**Methodologies**

- Promote publication of methodological guidelines in more applied journals in more easily understandable format to transfer knowledge to researchers who need to apply more rigorous methods;
- Promote full availability of models in electronic format to combat space restrictions in hardcopy publications;
- Promote consistency of methodological review for all CER studies;
- Promote adoption of explicit best practices guidelines among regulatory and reimbursement authorities;
- Periodically update all ISPOR Task Force reports;
- Periodically review use of ISPOR Task Force guidelines and their impact on CER quality; and
- Periodically report on statistical and methodological challenges in CER.

**Publications**

- Develop standard HEOR conduct and reporting guidance (beginning with CER) to which journals may refer their authors and their reviewers;
- Lobby to establish these guidelines within the ICMJE Requirements to which most journals refer in their Author Instructions;
- Provide support in terms of additional reviewer expertise to those journals without appropriate reviewers;
- Periodically report on journals publishing CER; and
- Periodically report on the quality of CER publications.

**Decision-making**

- Recognize those countries/agencies/practitioners/ private companies using CER well at least annually;
- Recognize those practitioners/researchers supporting patient use of CER in decision-making at least annually; and
- Promote frequent presentation of case studies of the applied use of CER concepts or guidelines.
Sources of financial support: No financial support was received by any of the authors to work on this report.

References


36. McGBan et al.
Quality Improvement in Cost Effectiveness Research


131 Kim LG, Thompson SG. Uncertainty and validation of health economic decision models. Health Econ 2009; February 10 [Epub ahead of print].


The Cost-Effectiveness of Bosentan in the United Kingdom for Patients with Pulmonary Arterial Hypertension of WHO Functional Class III

Matt D. Stevenson, PhD,1 Frances C. Macdonald, PhD,2 Jonathan Langley, BSc,2 Elke Hunsche, PhD,2 Ron Akehurst, BSc1

1School of Health and Related Research, University of Sheffield, Sheffield, UK; 2Actelion Pharmaceuticals UK Ltd., London, UK

ABSTRACT

Objectives: To assess whether bosentan or no active intervention, in addition to palliative care, is the more cost-effective first-line treatment option for patients with idiopathic pulmonary arterial hypertension (iPAH) or PAH associated with connective tissue disease (PAH-CTD) of WHO functional classification (FC) III in the United Kingdom.

Methods: A cost-utility model simulated the treatment of patients with PAH of FC III. Patients remained on the selected intervention until death or clinical deterioration to FC IV, which would trigger initiation of epoprostenol treatment. The initial first-line treatment choice was assumed to not affect survival, but to affect the time until clinical deterioration, with this assumption being relaxed in sensitivity analyses. The distribution of time to clinical deterioration was estimated from long-term clinical trial databases of bosentan and from published literature. Utility associated with FC was taken from published literature.

Conclusions: Bosentan is likely to be a more cost-effective first-line therapy for patients with PAH FC III in the UK than no active intervention.

Keywords: cost-effectiveness, mathematical modeling, pulmonary arterial hypertension, simulation.

Introduction

Disease Overview

Pulmonary arterial hypertension (PAH) is a rare but devastating disease that is characterized by elevated pulmonary vascular resistance eventually leading to increased right heart pressures, right heart failure, and untimely death [1,2]. The disease can either be classified as idiopathic (iPAH) or associated with other conditions, such as connective tissue disease (CTD), congenital heart disease, or human immunodeficiency virus. Patients are classified according to the degree of functional impairment based on exercise tolerance, using the New York Heart Association/World Health Organization (NYHA/WHO) functional classification (FC), which ranges from least severe, FC I, to most severe, FC IV. The prevalence of the more severe classification of PAH (FC III and IV) is estimated to be from 30 to 50 cases per million [3], which equates to approximately 1800 to 3000 patients with FC III or FCIV in the United Kingdom.

PAH has a significant impact on patients’ quality of life [4,5] and mortality [6]. Median survival after diagnosis of iPAH in an untreated population has been estimated at 2.5 years for those in FC III [6], which is comparable with prognoses in some advanced cancers [7].

Improving survival, stabilizing disease and improving quality of life are key aims in any treatment strategy for patients with PAH. Because there is no cure for PAH, patients will ultimately deteriorate, resulting in additional treatments, including continuous intravenous treatment with prostaglandins (PGs), lung transplantation (in very few cases), or death [8].

Treatment Options

Historically, standard care in the United Kingdom for patients with PAH FC III was palliative care (diuretics, warfarin, and calcium antagonists, as required), with the option for continuous i.v. PGs, typically epoprostenol, once they reached FC IV [9]. Nevertheless, the treatment armoury was extended in 2002 with the launch of bosentan. Bosentan is an oral dual endothelin receptor antagonist licensed for the treatment of PAH. Bosentan competitively blocks both ETA and ETB, and therefore decreases both pulmonary and systemic vascular resistance, resulting in increased cardiac output without increasing heart rate [10]. In the United Kingdom, bosentan (Tracleer Actelion Registration Ltd., London, UK) is indicated for the treatment of PAH to improve exercise capacity and symptoms in patients in FCIII. In addition, improvements have also been shown in patients with FC II; bosentan is also indicated to reduce the number of new digital ulcers in patients with systemic sclerosis and ongoing digital ulcer disease. Since the launch of bosentan, the treatment armory has been further extended with the introduction of additional pharmaceutical interventions (sildenafil, sitaxentan, vena- tavis, and ambrisentan).

Objectives

We aim to assess in the context of England and Wales which of bosentan or no active intervention, each used in addition to
palliative care, represents the more cost-effective first-line therapy for FC III patients within the two largest subgroups of PAH, i.e., patients with iPAH, or PAH related to CTD (PAH-CTD). Emphasis was placed on using long-term data (with follow up for up to 3 years) as the basis for the cost-effectiveness model because these would provide a more robust measure of disease progression than modeled extrapolations of randomized controlled trials (RCT) that are typically restricted to a maximum of 16 weeks and use surrogate end points.

Ideally, sildenafil sitaxentan and ambrisentan should be modeled as well to determine the most cost-effective first-line treatment from all currently licensed treatment options; however, necessary data were not available to incorporate these interventions: For sildenafil published efficacy data for the licensed dose of 20 mg thrice daily stems from a small sample RCT (n = 40 in FC III) with a short follow-up period (12 weeks) [11]; Limited long term data exist for this dose to date. For sitaxentan and ambrisentan, only 1-year long-term data from open-label extension studies of RCTs have been published in peer-reviewed journals to date [12,13]; however, these did not provide the detailed level of information required to model clinical deterioration as defined here. No data beyond 1 year have been published.

Intravenous PGs are rarely used in FC III patients because of the associated expense and the route of administration, which carries the risk of line infections. Therefore, i.v. PGs have been included as the treatment provided to those who fail oral treatment by deteriorating to FC IV.

The Efficacy of Bosentan

The efficacy and safety of bosentan has been demonstrated in a range of studies, including two pivotal, randomized, double-blind, placebo controlled trials [14,15]. These trials showed statistically significant (P < 0.05) improvements versus placebo at 12 and 16 weeks in the 6-minute walk test (6MWT) [9], a standardized test that measures the distance a patient can walk quickly on a flat, hard surface in a period of 6 minutes. Bosentan was also shown to significantly reduce the time to clinical worsening, a composite end point, which included death, lung transplantation, PAH-related hospitalization, PAH worsening leading to discontinuation, need for PGs, or atrial septostomy [15,16].

Longer-term studies have been undertaken, where patients who completed the pivotal RCTs had the option to continue or switch to bosentan treatment, with the addition of other treatments where necessary. From these long-term studies, the survival of patients with iPAH, who began on bosentan treatment in addition to palliative care, was estimated at 96, 89, 86% at 1, 2, and 3 years, compared with a predicted survival of 69, 57, 48%, respectively, for iPAH patients treated with palliative care only [16]. In the long-term study in iPAH, the majority of patients remained on bosentan monotherapy (85% and 70% at 1 and 2 years, respectively). For patients with PAH associated with CTD (a patient cohort with an even poorer prognosis), survival with bosentan was estimated at 86% and 73% at 1 and 2 years, which is significantly higher than the 55% 1-year survival of untreated patients [17].

All long-term data to date has been collected within open-label extension studies. Conducting long-term RCTs in a rare and life-threatening condition such as PAH poses considerable challenges because of the difficulty in recruitment of adequate numbers of patients and the ethical issues of carrying out placebo-controlled studies on a long-term basis in patients with poor prognoses.

Methods

Description of the Cost-Effectiveness Model

A mathematical model was constructed to simulate the changes in FC and the time until death of hypothetical FC III patients with either iPAH or PAH-CTD. The time horizon was that of patient lifetime. The cost perspective was that of the NHS with only direct costs considered. Ten thousand patients were simulated and replicated to form two cohorts, one that received bosentan and palliative care and one that received palliative care only. There was a possibility of a change in an individual’s FC in the first 12 weeks of treatment, the probability of which was taken from RCTs of patients with iPAH (Table 1). These were also assumed applicable to patients with PAH-CTD. Patients deteriorating to FC IV at 12 weeks were assumed to change treatments to epoprostenol and palliative care.

For each patient, predicted life expectancy and the “time to clinical deterioration” (TTCD), a metric that estimates the long-term progression of PAH to FC IV and which is described later, were compared. If the life expectancy of the patient was lower than the TTCD, it was assumed that the patient remained on their intervention and within their FC at 12 weeks until death. If the TTCD was lower than the life expectancy, it was assumed that the patient would remain in their FC at 12 weeks until TTCD, at which time the patient was assumed to deteriorate to FC IV and receive epoprostenol and palliative care until death. For each patient, a discrete event simulation approach was used where the simulated time moved directly to the time of the next scheduled event (either TTCD or death). This approach was used because it was more computationally efficient and removed the need for establishing arbitrary durations for the time cycles required by Markov models.

In the base case, the life expectancy of a patient was assumed to be independent of the initial treatment prescribed; assuming both that clinicians would be vigilant in observing deterioration and that there would be some stabilization of disease once epoprostenol is prescribed. The outcome measure that would thus vary between interventions would be the proportion of time spent by the patient in the most severe disease state, with the least efficacious intervention (palliative care alone) being associated with a longer duration in FC IV. The proportion of time a patient resides in FC IV affects cost-effectiveness because the utility associated with being in this class is lower than in FC III [5], and because the relatively expensive epoprostenol treatment, the cheapest i.v. PG at the time of the study (Table 2) would be initiated. The assumption of constant and treatment-independent

<table>
<thead>
<tr>
<th>Table 1</th>
<th>The number of patients, initially in functional class III, in each functional class following 12 weeks of treatment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intervention</td>
<td>Functional class after 12 weeks of treatment</td>
</tr>
<tr>
<td></td>
<td>n</td>
</tr>
<tr>
<td>Bosentan</td>
<td>139</td>
</tr>
<tr>
<td>No active intervention</td>
<td>65</td>
</tr>
</tbody>
</table>
life expectancy may be counter-intuitive because it would seem reasonable to assume that bosentan provides a survival advantage compared with no active intervention. Nevertheless, we conservatively assumed no impact on survival, because there are no data on the incremental survival benefit by bosentan, and then tested this assumption within sensitivity analyses.

Beyond the TTCD for bosentan, all patients, regardless of initial therapy would either be in FC IV and be receiving epoprostenol and palliative care, or would be dead. In either circumstance, the costs and quality-adjusted life-years (QALYs) accrued would be identical regardless of the initial treatment. This allowed the lower of the TTCD for bosentan or life expectancy to be used as the time horizon for each patient, thereby reducing computational time of the model. Figure 1 provides a schematic of the care pathway for a patient with iPAH and an average TTCD, which is assumed shorter than the patient’s life expectancy.

The model calculates the costs and QALYs associated with each treatment strategy with analyses undertaken separately for iPAH patients and for those with PAH-CTD. Both benefits and costs have been discounted at a rate of 3.5% per annum as recommended by the National Institute for Health and Clinical Excellence [18].

**Primary Outcome Measure**

Although the primary end point of RCTs has been 6MWT, which measures symptom severity and physical functioning, we have not modelled on this parameter. Although the baseline value of the 6MWT has been shown to be predictive of survival, the change in the distance from pre- to posttreatment has consistently not been found to be of prognostic value [13,19–23; Sitbon O, Channick R, McLaughlin W, unpubl. data]. The CE model has instead used “time to clinical deterioration” (TTCD), an outcome parameter reflecting progression of disease over longer time periods than seen in the RCT. For bosentan, the data for TTCD was collated from the combined data set of the two bosentan pivotal RCTs [11,12], the long-term follow-up of these studies [Actelion, unpubl. data], and a database associated with the article by Williams et al. [24], a longitudinal study recruiting patients with CTD-PAH. These sources, which provide data for periods of up to 3 years, are hereafter referred to as the full bosentan dataset. The TTCD was defined as the time at which a patient discontinued bosentan monotherapy, either through the addition (or substitution) of another intervention or through death. In the full bosentan database, 46 of 169 patients with iPAH had a TTCD, with corresponding values for PAH-CTD being 30 and 66, highlighting the poorer prognosis of PAH-CTD patients. TTCD was used within the model as a proxy for the time at which the patient would deteriorate to FC IV and receive epoprostenol and palliative care. It should be noted therefore that TTCD is not equivalent to the “time to clinical worsening” endpoint from the randomized controlled trials; TTCD should be a more robust measure, reflecting clinically and economically relevant outcomes.

**Estimating the TTCD**

Kaplan–Meier analysis was conducted in STATA (STATA Corporation, College Station, TX) using all data from the bosentan full data set in order that informative data were not excluded. PAH type and FC were used as covariates in the model to avoid bias in estimating the distribution for patients with iPAH of FC III and CTD-PAH of FC III. Patients who were alive and remained on bosentan or palliative care were censored. The best fitting distribution of TTCD for bosentan was a Weibull model with a shape parameter of 1.11 (95% confidence interval 0.93–1.33), with the scale parameters calculated separately for patients with iPAH and PAH-CTD. The mean duration of TTCD for patients in FC III treated with bosentan was 6.54 years for patients with iPAH and 2.46 years for patients with PAH-CTD.

The distribution of TTCD for patients with iPAH, who received palliative care alone, was estimated from the median time to death reported by D’Alonzo et al. [6]. Epoprostenol treatment was not readily available to patients in the D’Alonzo study but is now regularly used in severe cases; we have therefore

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Acquisition cost*</th>
<th>Home delivery costs</th>
<th>Additional palliative care costs</th>
<th>Total cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bosentan</td>
<td>£20,102</td>
<td>£1,608</td>
<td>£1,773</td>
<td>£23,483</td>
</tr>
<tr>
<td>Epoprostenol</td>
<td>£47,508</td>
<td>£3,801</td>
<td>£9,036</td>
<td>£60,344</td>
</tr>
<tr>
<td>No active intervention</td>
<td>£–</td>
<td>£–</td>
<td>£1,303</td>
<td>£1,303</td>
</tr>
</tbody>
</table>


Figure 1 A schematic of the care pathway for an average patient with iPAH of FC III. FC, functional classification; iPAH, idiopathic pulmonary arterial hypertension; PG, prostaglandin; QALY, quality-adjusted life-years.
assumed that epoprostenol treatment would now be initiated (with the patient progressing to FC IV) at the time of death estimated by D’Alonzo et al. [6]. If epoprostenol would be initiated before this time to prevent mortality, then this assumption will underestimate the costs of first-line palliative care alone. It was assumed that the TTCD for iPAH initially treated with palliative care alone would be a Weibull distribution with a shape equal to that for bosentan treatment and the scale calculated so that the median time to death was equal to the 2.5 years reported by D’Alonzo et al. [6]. This resulted in a TTCD for patients with iPAH treated with palliative care alone of 3.51 years. No data were found on the survival of patients with PAH related to CTD treated with palliative care alone. It was assumed that the ratio in survival times between iPAH and PAH-CTD calculated for patients treated with bosentan (2.46/6.54) was applicable for patients treated with palliative care alone, resulting in a predicted TTCD for patients with PAH-CTD treated with palliative care alone of 1.32 years.

TTCDs were sampled independently for bosentan with palliative care and for palliative care alone. It was deemed implausible that a patient receiving palliative care alone would have a TTCD that was larger than that of a corresponding patient receiving bosentan and palliative care. Where the estimated TTCD values for an individual patient did not fit this constraint, the TTCD associated with palliative care treatment was reduced to that of bosentan plus palliative care.

Patient Characteristics
The distribution of patient age for those with iPAH or PAH related to CTD was estimated from the full bosentan dataset using STATFIT (Geer Mountain Software Corp., Rolla, MO), assuming no correlation between age at presentation and FC. This information was used in conjunction with UK life tables to determine the life expectancy [25], assuming that the average of male and female values was applicable. The distribution of age (in years) for patients with iPAH was a triangular distribution with a minimum value of 10.4, a maximum value of 82.4 and a most likely value of 50.9. The age of patients with PAH-CTD was represented as a beta distribution (alpha = 2.39, beta = 1.88) scaled to fit between 26 and 80 years.

Utility associated with functional class. The utility associated with each FC has been taken from Keogh and colleagues [5], who recorded SF-36 data for patients receiving bosentan treatment for a period of between 3 and 6 months. The utility values (0.73, 0.67, 0.60, and 0.52 for patients in FC I, II, III and IV, respectively) were assumed to apply to both patients with iPAH and those with PAH related to CTD. No differentiation in utility was made between treatments.

Costs used within the model. The acquisition costs, in December 2007, of bosentan (Tracleer 62.5 mg and 125 mg twice daily), epoprostenol (Flolan [Glaxo Wellcome UK Ltd., Uxbridge, Middlesex, UK] 1.5 mg, diluted to 20 ml per day) and palliative care are given in Table 2. We have included home delivery costs for bosentan and epoprostenol, which, on the advice of a PAH specialist center, were assumed to be 8% of drug acquisition costs. The costs of palliative care have been estimated from data provided by two specialist PAH centers. These specialist centers also provided data on the one-off costs of initiating therapy, which were estimated at £1347 for both bosentan and palliative care and £11,016 for an i.v. PG. All costs refer to the 2006 to 2007 financial year.

Calculation of Results
Deterministic analyses were undertaken using the midpoint value for each parameter. Probabilistic sensitivity analyses (PSA) were conducted varying the shape and scale of the TTCD for each treatment, the ratio of TTCD between patients with iPAH, and those with PAH-CTD, and the utility associated with each FC. The distributions of values used for parameters in the PSA are given in Table 3. Further sensitivity analyses relaxed the assumption that all patients have the same survival duration irrespective of first-line treatment. This was undertaken to examine the effects on the cost-effectiveness ratio if it were assumed that patients who received palliative care only died more quickly than those who started with bosentan. Sensitivity analyses on survival benefit used the midpoint value for each parameter.

Results
The results were similar in both the deterministic and probabilistic analyses. The discounted costs and QALYs for each treatment strategy produced by the PSA analyses are provided in Table 4. In all of the 1000 parameter configurations, bosentan with palliative care as a first-line treatment provided more QALYs and was cost-saving compared with palliative care only. This is reflective of the faster progression of patients receiving palliative care alone to FC IV, which is associated with the initiation of the relatively expensive epoprostenol and a reduction in utility. The results from the analyses that relaxed the assumption that life expectancy was independent of initial

<table>
<thead>
<tr>
<th>Table 3</th>
<th>The distribution of parameters used in the probabilistic sensitivity analyses</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Midpoint value</td>
</tr>
<tr>
<td>Shape for Weibull distribution</td>
<td>1.113</td>
</tr>
<tr>
<td>Scale of Weibull distribution for FC III patients associated with CTD compared with patients with iPAH</td>
<td>Treatment specific 0.376</td>
</tr>
<tr>
<td>Utility FC I</td>
<td>0.73</td>
</tr>
<tr>
<td>Utility FC II</td>
<td>0.67</td>
</tr>
<tr>
<td>Utility FC III</td>
<td>0.60</td>
</tr>
<tr>
<td>Utility FC IV</td>
<td>0.52</td>
</tr>
</tbody>
</table>

Note that these references are [6] (D’Alonzo) and [5] (Keogh). CI, confidence interval; CTD, connective tissue disease; FC, functional classification; iPAH, idiopathic pulmonary arterial hypertension.
treatment are given in Figure 2. The cost per QALY of bosentan compared with palliative care alone became £30,000 (a potential cost-effectiveness threshold used in the United Kingdom) [19] or more when the survival benefit of bosentan treatment increased beyond 2 years for patients with iPAH, and when the survival benefit of bosentan approached 1 year for patients with PAH-CTD.

Discussion

Our results show that added to palliative care, bosentan is a more cost-effective first-line therapy for FC III patients with iPAH or PAH-CTD than no active intervention. There are insufficient data to analyze further etiology subgroups, but it is likely that these results can be extrapolated to other types of PAH FC III. Our results are not dissimilar to previously published models. Wlodarczyk et al. concluded that bosentan is a potentially cost-effective in an Australian context [26], while Highland et al. concluded that bosentan was more cost-effective than either epoprostenol or treprostinol [27]. Our analysis, however, has the strength of a longer and more populated data set, from which TTCD were predicted.

Our methodology has some limitations: The base-case assumed that survival was independent of initial treatment, which is likely to be incorrect, because bosentan has been shown to provide a survival advantage of currently undetermined length [17,18]. Similar life expectancies would rely on clinicians carefully monitoring patients on palliative care alone and providing prompt epoprostenol treatment as the patient deteriorates to FC IV. Sensitivity analyses have shown that any survival advantage of bosentan has a marked effect on the cost-effectiveness, which reflects the frequent issue that it can be more cost-effective to let patients die rather than to use relatively costly treatments. Data on the additional survival associated with initial bosentan treatment rather than no active intervention are not available and will be difficult to generate, because long-term comparative studies using placebo are ethically not justifiable.

Our results are dependent on i.v. PGs being a treatment option for patients in FC IV. Were this withdrawn because of high acquisition costs, then bosentan would likely not be cost-effective compared with no active intervention. We have not explicitly modeled this scenario assuming that it would be unethical to remove a widely used last resort treatment from the physicians’ armory. We have excluded the possibility that patients, who initiate epoprostenol in FC IV, could improve to FC III, and have also not modeled the disutility associated with receiving continuous intracardiac infusion. Nevertheless, because these effects are independent of the initial treatment (bosentan or no active intervention) and may counterbalance themselves, these omissions are unlikely to bias the results.

It was assumed that the rate of hospitalization was independent of treatment or functional class, allowing these costs to be excluded. This is likely to be favorable to palliative care because patients deteriorate to FC IV most quickly on this treatment, which is likely to be associated with a greater number of hospitalizations.

Not all treatments for PAH have been evaluated as the data needed to determine TTCD at licensed doses for interventions other than bosentan, and no active intervention were not available to the authors. Thus, only conclusions regarding the cost-effectiveness of bosentan compared with no active treatment can be made until comparable long-term data on TTCD become available for the licensed comparators, allowing for an indirect comparison of treatment options after appropriately controlling for differences in study designs and study populations.

Because of the absence of clinical data, the scenarios modeled do not consider combination therapy options, such as initiating two active interventions while in FC III or using bosentan and epoprostenol concurrently once a patient has deteriorated to FC...

### Table 4 Cost per QALY results, per patient, from the probabilistic sensitivity analyses

<table>
<thead>
<tr>
<th>Type of PAH</th>
<th>First-line treatment</th>
<th>Discounted costs (£000)</th>
<th>Discounted QALYs</th>
<th>Cost per QALY compared with palliative care (£000)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Idiopathic</td>
<td>Bosentan</td>
<td>134</td>
<td>3.32</td>
<td>Dominating</td>
</tr>
<tr>
<td></td>
<td>Palliative care</td>
<td>203</td>
<td>2.95</td>
<td>n/a</td>
</tr>
<tr>
<td>Related to CTD</td>
<td>Bosentan</td>
<td>62</td>
<td>1.36</td>
<td>Dominating</td>
</tr>
<tr>
<td></td>
<td>Palliative care</td>
<td>94</td>
<td>1.21</td>
<td>n/a</td>
</tr>
</tbody>
</table>

Dominating denotes producing more QALYs at a lower cost.

CTD, connective tissue disease; n/a, not applicable; PAH, pulmonary arterial hypertension; QALY, quality-adjusted life-year.
The Cost-Effectiveness of Bosentan

IV. Once data become available, the relative cost-effectiveness of such strategies can be assessed.

Conclusion

Compared with no active intervention, bosentan is a potential cost-effective first-line treatment for UK patients with iPAH or PAH-CTD within FC III. It is estimated to be more cost-effective than palliative care, followed by epoprostenol, providing that the survival benefit of bosentan followed by epoprostenol is not greater than 2 years for patients with iPAH and 1 year for patients with PAH-CTD.

References

Cost-Effectiveness of Using Clinical Risk Factors with and without DXA for Osteoporosis Screening in Postmenopausal Women

Dirk Mueller, Health Econ,1 Afshin Gandjour, MD, PhD, MA2

1Department of Health Economics and Health Care Management, University of Cologne, Cologne, Germany; 2The James A. Baker III Institute for Public Policy, Rice University, Houston, TX, USA

ABSTRACT

Background: According to several guidelines, the assessment of postmenopausal fracture risk should be based on clinical risk factors (CRFs) and bone density. Because measurement of bone density by dual x-ray absorptiometry (DXA) is quite expensive, there has been increasing interest to estimate fracture risk by CRFs.

Objective: The aim of this study was to determine the cost-effectiveness of osteoporosis screening of CRFs with and without DXA compared with no screening in postmenopausal women in Germany.

Methods: A cost-utility analysis and a budget-impact analysis were performed from the perspective of the statutory health insurance. A Markov model simulated costs and benefits discounted at 3% over lifetime.

Results: Cost-effectiveness of CRFs compared with no screening is €4607, €21,181, and €10,171 per quality-adjusted life-year (QALY) for 60-, 70-, and 80-year-old women, respectively. Cost-effectiveness of DXA plus CRFs compared with CRFs alone is €20,235 for 60-year-old women. In women above the age of 70, DXA plus CRFs dominates CRFs alone. DXA plus CRFs results in annual costs of €175 million, or 0.4% of the statutory health insurance’s annual budget.

Conclusion: Funders should be careful in adopting a strategy based on CRFs alone instead of DXA plus CRFs. Only if DXA is not available, assessing CRFs only is an acceptable option in predicting a woman’s risk of fracture.

Keywords: cost-utility analysis, modeling, osteoporosis, secondary prevention, women’s health.

Introduction

Osteoporosis, a multifactorial disorder resulting in increased bone fragility, occurs in women after menopause and is one of the most important disorders affecting the elderly [1]. Population aging is expected to increase the number of osteoporosis-related fractures such as hip fractures and, hence, the economic burden for society.

Bone mineral density (BMD) is considered an important predictive factor for osteoporotic fractures and is measured by densitometry. Densitometry results are usually reported as a t-score, which is the number of standard deviations between the value of an individual and the mean value of a group of young adults of the same sex [2]. According to the criteria of the World Health Organization (WHO) osteoporosis is defined by a t-score of ≤−2.5 [3]. If bone density is measured by dual x-ray absorptiometry (DXA), the risk of hip fracture (other fractures) increases by a factor between 3.7 (1.2) at age 50 and 1.7 (1.6) at age 90 for each standard deviation decrease in BMD at the femoral neck [4]. This increase in fracture risk for each standard deviation change is called the gradient of risk (GR/SD) [4].

DXA is expensive, not available everywhere, and to a certain degree unreliable because BMD can vary by up to 20% to 50% around an individual’s true BMD [5]. Furthermore, different scanners for bone density used in the same patients vary considerably in the proportion of those who receive a diagnosis of osteoporosis [6]. While, in Germany, the prevalence of osteoporosis in women varies between 7% (age: 55) and 19% (age: 80) [7], the majority of fractures occur in women who do not have osteoporosis [7]. Thus, the use of DXA in primary or secondary prevention is disputed [8].

Recently, several risk factors for fracture have received attention [9]. These include prior fragility fractures, a family history of hip fracture, low body mass index (BMI), smoking, alcohol intake, and the use of oral corticosteroids. Combinations of these risk factors were used to develop decision rules for BMD referrals [10]. Although a case finding based on DXA alone (compared with fractures that will occur in the following 10 years) has a specificity (proportion of true-negatives) of approximately 90%, its sensitivity (proportion of true-positives) is only 34% [11]. Different strategies based on CRFs alone, in turn, have shown to exceed a sensitivity of 80% although their specificity is only about 50% (compared with the reference standard low BMD measured by DXA) [12]. Thus, predicting fracture risk based on CRFs in addition to BMD increases the GR for the prediction of hip and other fractures [4]. As a result, the sensitivity and the positive predictive values (proportion of women with positive test results who will have a fracture in the following 10 years) increase [13,14].

To guide treatment based on a combined use of CRFs and BMD, several organizations recently have recommended using CRFs and BMD to assess an individual’s absolute 10-year risk of fracture [9,15] or annual incidence of fracture [16]. According to the German osteology umbrella organization, Dachverband Osteologie (DVO) guideline, DXA should be provided for women when there is a 10-year risk of combined vertebral (clinical and morphometrical) and hip fractures of ≥20%. Drug treatment should be provided for women with a combined risk of ≥30% for vertebral and hip fractures [7]. The t-score required to reach this risk threshold varies by age. A 55-year-old woman, for example,
receives treatment for a t-score of −4, whereas a 67-year-old woman receives it based on a t-score of −3. Additional risk factors further increase the t-score required to reach the threshold [7].

While this and other recently developed screen-and-treat strategies agree that CRFs should be given more attention, treatment recommendations still tend to center on DXA [7,15,16]. Although the combined use of DXA and CRFs improves the GR/SD [4], for women aged >65 years, its sensitivity is only 60% even if a risk threshold of 30% is chosen [13]. This is an increase of 80% compared with DXA alone in women aged 70 to 79 years, but the price to pay for this improvement is a decrease in specificity by 16% [11,13]. The usage of CRFs alone, however, may be of diagnostic value for predicting fracture risk because age-specific GRs are similar to those of BMD alone [4]. Thus, a strategy where fracture risk is calculated by CRFs alone may improve the cost-effectiveness compared with an expensive DXA-centered strategy. The National Institute for Health and Clinical Excellence, for example, recommends bisphosphonates in postmenopausal women aged 75 years and older even without the need for DXA if the clinician considers DXA to be clinically inappropriate or unfeasible [16].

In the vast majority of cost-effectiveness analyses on postmenopausal osteoporosis treatment, women at increased risk were selected by low BMD [17]. Nevertheless, the use of CRFs as a prescreening tool for DXA (i.e., DXA only for those women with elevated CRFs) has been shown to be cost-effective when compared with mass screening with DXA alone [18]. In addition, there are two modeling studies that analyzed the cost-effectiveness of treatment in postmenopausal women, based on long-term fracture risk rather than on BMD alone [19,20]. In contrast to our analysis, these studies did not consider treatment costs of false positives, selected women at increased risk based on additional risk factors (e.g., BMI or the use of oral glucocorticoids), used a 10-year modeling horizon, and assumed that beyond 10 years, women would have a mortality rate equal to that of an age- and sex-matched population [19,20,21].

The present study investigated the cost-effectiveness of the following strategies: 1) screening based on CRFs alone (without information about BMD) and treatment with alendronate in case of risk of ≥30% (age groups 60–70 and 70–80), or treatment with alendronate for all women (age group: 80–90); 2) screening with DXA plus CRFs (plus alendronate); and 3) no screening (Fig. 1). While different medical treatment options are recommended, we chose alendronate, an antiresorptive bisphosphonate, as the sole drug because, in our previous cost-utility analysis, it has shown to be most cost-effective [22].

**Methods**

The analysis was performed from the perspective of the German statutory health insurance (SHI). For the base case we considered a cohort of 10,000 women aged 60, 70, and 80 years. Because patients with osteoporosis face fracture risk that is continuous over time, we developed a Markov model in Microsoft Excel.
The health benefit was estimated in terms of quality-adjusted life-year (QALYs), and incremental cost-effectiveness ratios (ICERs) were calculated. Preference weights are expressed by values between 1 for perfect health and 0 for the death [23]. A budget impact analysis (BIA) was performed by multiplying incremental costs per woman by the number of women who are insured by the SHI [24]. The BIA included future costs for unrelated health conditions during added years of life. The annual expected resource use for the SHI was estimated based on public databases [24,25]. Lost productivity was not included in the BIA because it is not relevant to the expenditure side of the SHI. As recommended for chronic diseases, a lifetime horizon was chosen [26].

**Overview and Model Design**

The Markov model uses age-specific fracture rates. The cycle length is 1 year because transition probabilities obtained from the literature refer to periods of at least 1 year. The model starts with a cohort of high-risk women who are identified on the basis of CRFs with and without DXA and who have a combined 10-year risk of ≥30% for vertebral and hip fractures. It stops at the age of 100 because, for Germany, there are no survival data beyond the age of 100 [27]. There are eight health states (Fig. 2): no fracture (≥30% probability of fracture in the 10 years from the start of the model), three-fracture states (hip, vertebral, and forearm), the corresponding postfracture states, and death. All women start in a state with a long-term risk of ≥30%. For each cycle, there is a defined probability of staying in the no-fracture state, experiencing a fracture, or dying. A woman who is in a fracture state can have another fracture at the same or another site. Women, who have a fracture change to the postfracture state, stay in the same fracture state if they have a refracture, change to another fracture state, or die. We conducted a half-cycle correction. Costs and benefits were discounted at an annual rate of 3% [28]. All costs are presented in euros and year 2006 values and were adjusted for inflation based on the German Consumer Price Index [29].

The structure of this Markov cohort model is similar to that of an established reference model [30]. As recommended there, a lifelong time horizon with a cycle length of 1 year was used, effectiveness was assumed to decrease linearly for a given “offset time,” and fracture states for hip, vertebral, and forearm were modeled [30]. In contrast to the reference model, in our model, a woman may suffer a vertebral or forearm fracture after a hip fracture. Excess mortality after hip or vertebral fracture was modeled as well as the conservative assumption that costs of vertebral and wrist fractures only incur during the first year after the fracture. Although a state for other osteoporotic fractures was not modeled, for all fracture sites included in the analysis, postfracture states were added to reflect a persistent decrease in health-related quality of life (QoL) (hip, vertebral fractures) and the sustained risk increase for subsequent fractures.

**Data**

*Efficacy/Effectiveness.* Data on effectiveness of alendronate were taken from a meta-analysis that was based on large randomized controlled trials [31]. While effectiveness data on high-risk women selected by DXA were restricted to patients with osteoporosis or severe osteoporosis, effectiveness data on high-risk women selected by CRFs were also based on osteopenic women [31], resulting in lower effectiveness in women selected by CRFs.

Alendronate was offered for 4 years, which is in the range of recommended treatment duration in Germany (3–5 years) [7]. Effectiveness was assumed to decrease in linearly over a period of 4 years after the last intake [32]. Basing on the effectiveness data used for our analysis, we assumed that alendronate has no relevant side effects [31].

Medication compliance (or adherence) for individuals with chronic diseases such as osteoporosis is poor. Compliance is defined as the “extent to which a patient acts in accordance with the prescribed interval and dose of a dosing regimen” [33]. To adjust drug effectiveness and costs for the rate of noncompliance in the real world, we multiplied effectiveness (at full compliance) and costs of alendronate and checkups by the “medication possession ratio” (the number of doses dispensed in relation to the dispensing period [33]) for daily intake of oral bisphosphonates (more than 12 months) using German prescription data [34]. Persistence, which is defined as “the duration of time from initiation to discontinuation of therapy” [33], was not modeled.
**Clinical Risk Factors for Osteoporosis Screening**

Absolute long-term risk and incidence. The number of women at risk was taken from the DVO guideline, which modeled an absolute 10-year fracture risk for the German population, based on published data on population age, prior vertebral fractures, and bone density [35,36].

In women below the age of 80, incidence rates of the general population [37–39] were increased to be at the level of the average fracture risk in women above the treatment threshold. The average risk in a group of women selected by CRFs depends on the GR and the ratio of the chosen risk threshold to the population risk. If treatment is provided on the basis of CRFs alone, the GR varies from 1.4 to 2.1, depending on age and fracture site [4]. An absolute risk of 30% in the age group of 70 to 80, e.g., means that this threshold equals 1.3 times the population risk, although the average risk, which is higher than the threshold, equals 1.9 times the population risk for hip fracture and 1.7 times for vertebral or forearm fractures, respectively [14] (details of formulae used for calculations of incidence can be found at: http://www.ispor.org/Publications/value/VIHsupplementary/VH1218_Mueller.asp). In women above the age of 80, incidence rates of the general population were used because all women receive treatment.

For women with a prior fracture, we assumed that the risk increases for all subsequent fractures. The risk of a woman who suffered fractures in two or more different sites was assumed to be determined only by the last fracture. The magnitude of the risk increase depends on the location of the prior fracture [40].

Mortality and health-related QoL. Mortality data were obtained from a public database of the Federal Statistical Office (Wiesbaden, Germany) [41] and were adjusted for women at increased risk [42–44]. Long-term mortality and life expectancy associated with hip fractures were modeled in a previous article [45]. In this model, impatient mortality was based on a high-quality study analyzing the volume-outcome relationship of hip fracture surgery in German hospitals [46]. Mortality from revision surgeries was also considered in this model. QoL data were based on the EuroQol, a preference-based questionnaire [47].

For the state "Long-term risk ≥30%," we used QoL data of the general population, that is, utility values of <1 for all ages, estimated by a time trade-off questionnaire [48]. The reason is that individuals with osteoporosis may suffer from additional diseases causing disutility. QoL of women aged 60 to 70 years was further reduced because, in this group, a long-term risk of ≥30% can only be reached if a woman had a prior vertebral fracture [7]. QoL for forearm and clinical vertebral fractures was also based on surveys that used the EuroQol [49]. Because a value of 1 was assigned to the no-fracture state, data were multiplied by health-related QoL (HR-QoL) data from the general population [48]. For the postfracture states, we assumed an improvement of QoL, though not to the prior level. These values were not changed except when additional fractures occurred or patients died. In women with a forearm fracture and a prior vertebral fracture, we did not use values of forearm fractures because these were higher than values for the status post vertebral fracture. For women with a forearm or vertebral fracture and a prior hip fracture, the lower preference weight of the prior hip fracture was maintained.

Validity of the diagnosis. According to the DVO the combined 10-year risk of vertebral and hip fracture in the age group 60 to 70 is 10%. The threshold of 30% can only be reached if a woman has had a vertebral fracture because other risk factors do not increase the risk sufficiently [7]. A pretest x-ray is recommended in Germany for women with severe low-back pain or moderate low-back pain for >4 weeks [50]. In Germany, the 1-year prevalence of back pain in women aged 60 to 70 years is estimated to be 23% [51], and the proportion of these women undergoing an x-ray is 35% [52]. Hence, only 8% of all women in the model undergo a routine x-ray, provided that all women with lasting back pain attend a physician. Nevertheless, this group includes only 35% of all vertebral fractures because 65% are not recognized clinically [53]. Furthermore, in women with clinically apparent fractures who undergo the pretest, a high prevalence of unspecific back pain, morphological changes, and misinterpreted x-rays confounds the validity of the diagnosis [54]. As a result, 2.4% of all women screened receive treatment with alendronate (90% of these women are true-positive).

Because the 10-year risk of vertebral and hip fracture for 70- to 80-year-old women is already 23%, the German DVO considers this risk to be ≥30% if a woman has a prior fracture, has a parental history of a femur fracture, is a smoker, is immobile, or has a tendency to fall [7]. This decision rule is consistent with the case-finding strategy of the National Osteoporosis Foundation, which is carried out to identify individuals with osteoporosis compared with low BMD as a reference standard [55,56]. Based on a t-score of −2.5, which in this age group corresponds to the threshold of 30%, 85% of all women undergo treatment because they are suspected as being at risk as a result of one or more CRFs (26% of these women are true-positive) [10].

All women aged >80 years receive drug treatment because the therapeutic threshold of 30% is assumed to be reached by age alone (Table 1).

The validity of DXA plus CRFs is extensively described in our prior analysis [22]. Briefly, for women aged 60 to 70 years, 24% of all women screened receive treatment (23% of these women are true-positive). In women the age of more than 70 and 80, the proportion of women who undergo treatment is 33% (42% of these are true-positive) and 44% (73%), respectively.

Costs
Our analysis considers the costs of screening as well as treatment costs of false- and true-positives. General health-care costs in added years of life were also taken into consideration. As a cost of screening, we applied the costs of routine checkups, which include a face-to-face interview, an evaluation of the medical history, and a physical examination. Therefore, we assumed that risk factors such as a prior fracture, a parental history of a femur fracture, smoking, immobility, and tendency to fall are assessed during routine visits and do not require specific equipment. The costs of initial x-rays were not included because patients with severe or persistent back pain in Germany routinely undergo an x-ray [50].

Although costs of treating vertebral and forearm fractures were calculated for the year of the fracture, they were not applied to subsequent years [30], which is a conservative assumption. Costs of treating hip fractures were taken from a recent German analysis modeling long-term cost and effects of hip fracture surgeries, which includes costs of initial hospitalization for hip implants, readmissions, transportation services, outpatient treatment, rehabilitation, long-term care, and costs in added years of life [45]. All long-term costs of hip fractures were considered, including those of refractures. For double counting of refracture costs to be avoided, a hip fracture can only occur once in the Markov model. If a woman with a prior hip fracture suffered a forearm or a vertebral fracture, the costs of this additional fracture would be added to the costs of the hip fracture. Costs were
Treatment is recommended if NMB was greater than 0.

The ICERs of a screen-and-treat strategy based on CRFs alone; that is, it is more effective but less expensive. Although, in women aged 60 to 70 years, CRFs alone are considerably more cost-effective than DXA plus CRF, less than one quarter of women at increased risk are detected. In contrast, DXA plus CRFs detects 58% of women at increased risk.

If a screen-and-treat strategy based on CRFs alone was implemented in Germany, costs would total €560 million, or 0.4% of the SHI’s total annual budget [67] (Table 5). The major cost driver is treatment with alendronate for 5.4 million women across all age groups including both high- and low-risk women. Nevertheless, 60% of these women are not at risk (false-positives). When providing DXA plus CRFs in women above the age of 70 years, costs of DXA are offset by lower costs of treatment for false-positives. Only in women aged 60 to 70 years is DXA plus CRFs more expensive than CRFs alone. Thus, over all age groups, a screening strategy based on DXA plus CRFs decreases annual costs by €385 compared with a strategy with CRFs alone (Table 5).

Sensitivity Analysis

We carried out a one-way sensitivity analysis for all model variables. In addition, a threshold analysis was performed to determine the level of risk at different cost-effectiveness thresholds. The combined risk for vertebral and hip fractures was calculated based on cost-effectiveness thresholds between €5000 and €35,000. When varying the risk of vertebral and hip fractures simultaneously, an equal relative risk increase or decrease was assumed; that is, the relative risks of hip and vertebral fractures were varied by the same factor.

To assess how a simultaneous change of several variables affected the cost-effectiveness ratio, we performed a Monte Carlo simulation of model variables listed in Tables 2 and 3 (except for the discount rate as well as costs of checkups, alendronate, and hospitalizations caused by vertebral and forearm fractures).

For variables on a scale between 0 and 1, we assumed a beta distribution (0 ≤ θ ≤ 1, a > 0, b > 0). For cost data, we assumed a gamma distribution (a > 0 and b > 0), with the mean ab/ and the variance ab[b^2] [65]. We conducted 1000 iterations. Given that the interpretation of negative cost-effectiveness ratios is ambiguous, we transformed cost-effectiveness ratios into net monetary benefits (NMBs) using the following equation [66]:

\[ \text{NMB} = \lambda \times \Delta E - \Delta C \]

where \( \lambda \) = maximal willingness to pay, \( \Delta E \) = incremental benefit (QALYs), and \( \Delta C \) = incremental costs.

The decision rule we used was to adopt the screen-and-treat strategies in question if NMB was greater than 0.

Results

Base-Case Analysis

The ICERs of a screen-and-treat strategy based on CRFs alone versus no screening are below €22,000 in all age groups (Table 4). Compared with women aged 70 to 80 years, the increased absolute fracture risk in women aged 60 to 70 years makes immediate treatment of all women more cost-effective. Using DXA plus CRFs compared with CRFs alone (i.e., immediate treatment in women aged 80–90 years), there is an increase in QALYs in all age groups (Fig. 3). In women above the age of 70 years, a strategy based on DXA plus CRFs dominates a strategy based on CRFs alone; that is, it is more effective but less expensive. Although, in women aged 60 to 70 years, CRFs alone are considerably more cost-effective than DXA plus CRF, less than one quarter of women at increased risk are detected. In contrast, DXA plus CRFs detects 58% of women at increased risk.

Table 1: Diagnostic performance of screening for osteoporosis based on clinical risk factors

<table>
<thead>
<tr>
<th>Age group (10-year risk of fracture in the general population)</th>
<th>60–70 (9.5%)</th>
<th>70–80 (23.2%)</th>
<th>80–90 (53.2%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>X-Ray vertebral fracture</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fracture in T0–T10</td>
<td>220</td>
<td>2,232</td>
<td></td>
</tr>
<tr>
<td>No fracture in T0–T10</td>
<td>24</td>
<td>6,313</td>
<td></td>
</tr>
<tr>
<td>Sensitivity = 0.660</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Specificity = 0.950</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sensitivity = 0.962</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Specificity = 0.178</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(t-score = 2.5 or less)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Figures presented are numbers of patients. The calculations of women at increased risk for the different age groups are based on an absolute 10-year fracture risk and a cohort of 10,000 women.

†Period of 10 years.

NPV, negative predictive value; PPV, positive predictive value.
because the average risk decreases more in women identified by DXA plus CRFs than in women identified by CRFs alone. Assuming a higher GR improves the ICER of DXA plus CRFs by 25% to 50% compared with CRFs alone. Using less conservative assumptions such as a higher risk increase by prior fractures improves the cost-effectiveness ratio of DXA plus CRFs compared with CRFs alone. Further details on this can be found at: http://www.ispor.org/Publications/value/ViHsupplementary/ViH12i8_Mueller.asp.

**Threshold analysis.** Using a threshold cost-effectiveness ratio of €30,000, CRFs become cost-effective in women aged >60 years with a combined average fracture risk of 10.3% (hip: 1.3%). For a ratio of €10,000, this risk increases to 27.2% (3.7) (Fig. 3a). Assuming all women identified for being at increased risk would have a combined average fracture risk of exactly 30% (hip, age 60 years: 3.8%; age 70 years: 6.2%; age 80 years: 9.9%), the ICERs are €9575 (age 60 years), €47,351 (age 70 years), and €18,496 (age 80 years), respectively.

Using a threshold ratio of €30,000, DXA plus CRFs compared with no screening becomes cost-effective in women aged >60 years with a combined average fracture risk of 25.3% (hip: 2.6%), in women aged >70 years with an average fracture risk of 30.6% (6.3), and in those aged >80 years with a risk of 11.5% (5.0). Using a threshold ratio of €10,000, the risks increases to 80.1% (14.4), 69.2% (15.5), and 29.4% (13.6), respectively (Fig. 3). Assuming all women identified for being at increased risk would have a combined average fracture risk of exactly 30% (hip, age 60 years: 3.3%, age 70 years: 5.8%, age 80 years: 12.8%), the ICERs are

<table>
<thead>
<tr>
<th>Table 2</th>
<th>Cost data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parameter</td>
<td>Price per unit (€)</td>
</tr>
<tr>
<td><strong>1. Diagnosis</strong></td>
<td></td>
</tr>
<tr>
<td>Routine visit (initial)</td>
<td>25.56</td>
</tr>
<tr>
<td>Check-ups (physician)</td>
<td>76.68</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>2. Medication</strong></td>
<td></td>
</tr>
<tr>
<td>Alendronate, 5–10 mg daily</td>
<td>1.20</td>
</tr>
<tr>
<td><strong>3. Vertebral fractures</strong></td>
<td></td>
</tr>
<tr>
<td>Outpatient treatment</td>
<td></td>
</tr>
<tr>
<td>Consultation for osteoporosis with pathological fracture (MB0)</td>
<td>286.16</td>
</tr>
<tr>
<td><strong>Physical therapy</strong></td>
<td></td>
</tr>
<tr>
<td>Physiotherapy (WS2)</td>
<td>234.90</td>
</tr>
<tr>
<td>Heat (WS2)</td>
<td>116.82</td>
</tr>
<tr>
<td>Ergotherapy (SB1)</td>
<td>428.40</td>
</tr>
<tr>
<td>Total of outpatient and physical therapy</td>
<td>1,066.28</td>
</tr>
<tr>
<td>Inpatient treatment including rehabilitation</td>
<td>3,222.00</td>
</tr>
<tr>
<td><strong>4. Forearm fractures</strong></td>
<td></td>
</tr>
<tr>
<td>Outpatient treatment</td>
<td></td>
</tr>
<tr>
<td>Consultation for forearm fracture (S52, 69% of all forearm fractures)</td>
<td>342.26</td>
</tr>
<tr>
<td>Consultation for wrist fracture (S52, 31% of all forearm fractures)</td>
<td>313.00</td>
</tr>
<tr>
<td><strong>Physical therapy</strong></td>
<td></td>
</tr>
<tr>
<td>S52 and S62 weighted</td>
<td>140.88</td>
</tr>
<tr>
<td>Physiotherapy (EX2/3)</td>
<td>1,066.28</td>
</tr>
<tr>
<td>Cryotherapy (EX2/3)</td>
<td>363.42</td>
</tr>
<tr>
<td>Lymphatic drainage (LY1)</td>
<td>295.15</td>
</tr>
<tr>
<td>Total of outpatient and physical therapy</td>
<td>3,174.00</td>
</tr>
<tr>
<td>Inpatient treatment including rehabilitation</td>
<td>939.42</td>
</tr>
<tr>
<td><strong>5. Hip fractures</strong></td>
<td></td>
</tr>
<tr>
<td>(total costs)</td>
<td>50 years: 69,231</td>
</tr>
<tr>
<td></td>
<td>[67,289–71,443]</td>
</tr>
<tr>
<td></td>
<td>55 years: 62,591</td>
</tr>
<tr>
<td></td>
<td>60 years: 54,264</td>
</tr>
<tr>
<td></td>
<td>65 years: 47,837</td>
</tr>
<tr>
<td></td>
<td>70 years: 42,432</td>
</tr>
<tr>
<td></td>
<td>75 years: 39,586</td>
</tr>
<tr>
<td></td>
<td>80 years: 32,957</td>
</tr>
<tr>
<td></td>
<td>85 years: 29,417</td>
</tr>
<tr>
<td></td>
<td>90 years +: 26,187</td>
</tr>
<tr>
<td></td>
<td>60–65 years: 3,150</td>
</tr>
<tr>
<td></td>
<td>65–85 years: 6,150</td>
</tr>
<tr>
<td></td>
<td>&gt;85 years: 12,405</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Price per unit (€)</th>
<th>Annual frequency</th>
<th>Mean annual cost† per patient (€)</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Diagnosis and treatment codes are shown in parentheses (for physical therapy, we used the German Heilmittelkatalog, for vertebral and forearm fractures ICD-10 numbers are cited).</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>95% confidence intervals are shown in parentheses. For outpatient treatment and physical therapy, we varied mean values by 20%. All costs are given in 2006 Euros and were adjusted for inflation based on the German Consumer Price Index [29].</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Lifetime costs discounted at 3% (this discounting applies only to hip fractures).</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>These costs included costs of inpatient care in hospitals, hip replacements, revisions, readmissions, transportation services, outpatient treatment, rehabilitation, and long-term care.</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>For hip fractures, costs in added years of life are included in the total costs of hip fractures.</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Proportion of x-rays in women above 60–70 years, hip: 4.15 (3.72–4.67), vertebral/forearm: 3.54 (3.43–3.66) [39,40].

Table 3 Input variables for the model (95% confidence intervals or estimates thereof are shown in brackets)

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Range (95% CI or estimates thereof)</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>Incidence (%)</td>
<td>Vertebral fracture 60 years 0.95 (0.62–1.39), 65 years 1.23 (0.82–1.77), 70 years 1.79 (1.20–2.57), 75 years 2.93 (1.80–4.36)</td>
<td>[38]</td>
</tr>
<tr>
<td></td>
<td>Hip fracture 60 years 0.09, 65 years 0.16, 70 years 0.34, 75 years 0.70, 80 years 1.38, 85 years 2.94, 90 years 3.82, 95 years 3.06</td>
<td>[37]</td>
</tr>
<tr>
<td></td>
<td>Forearm fracture 60 years 0.49 (0.40–0.60), 65 years 0.58 (0.50–0.70), 70 years 0.73 (0.61–0.84), 75 years 0.71 (0.60–0.82), 80 years 0.77 (0.62–0.91), 85 years 1.04 (0.86–1.12), 90 years 0.94 (0.83–1.02)</td>
<td>[39]</td>
</tr>
<tr>
<td>10-year risk of fracture (%)</td>
<td>Proportion of inpatient cases Vertebral fracture 0.17 (0.11–0.24), forearm fracture 0.44 (0.19–0.51)</td>
<td>[37,64]</td>
</tr>
<tr>
<td></td>
<td>Relative mortality risk Osteoporosis 1.19 (1.04–1.36), Vertebral fracture 1.66 (1.51–1.80) for clinical fractures</td>
<td>[42]</td>
</tr>
<tr>
<td></td>
<td>Relative risk of secondary fractures 60–90 years Initial fracture hip 2.79 (2.06–3.77), Initial fracture forearm 1.69 (1.35–2.12), Initial fracture vertebral 2.52 (1.99–3.19)</td>
<td>[40]</td>
</tr>
<tr>
<td></td>
<td>Health-related quality of life No fracture (general population data) 60 years 0.83 (0.78–0.88), 65 years 0.81 (0.75–0.86), 70 years 0.75 (0.69–0.80), 75 years 0.73 (0.67–0.79), 80 years 0.70 (0.61–0.78), 85 years 0.68 (0.59–0.76)</td>
<td>[48]</td>
</tr>
<tr>
<td></td>
<td>Fracture states Clinical vertebral, first years 0.63 (0.50–0.75), Subsequent years 0.91 (0.83–0.98), Morphometrical vertebral, first years 0.82 (0.72–0.92), Subsequent years 0.91 (0.84–0.99), Forearm first years 0.98 (0.96–1.00), Subsequent years 1.00 (0.99–1.00)</td>
<td>[49]</td>
</tr>
<tr>
<td></td>
<td>Quality-adjusted life expectancy Hip fracture 60 years 10.98 (8.97–12.34), 65 years 8.17 (6.4–10.17), 70 years 6.13 (4.79–7.69), 75 years 4.67 (3.65–5.97), 80 years 3.39 (2.41–4.58), 85 years 2.55 (1.78–3.54), 90 years 1.64 (1.32–2.21)</td>
<td>[45]</td>
</tr>
<tr>
<td></td>
<td>Relative fracture risk (alendronate) CRFs Forearm: 0.64 (0.30–1.35), hip: 0.62 (0.40–0.98), vertebral: 0.56 (0.46–0.68)</td>
<td>[31]</td>
</tr>
<tr>
<td></td>
<td>DXA plus CRFs Forearm: 0.48 (0.31–0.75), hip: 0.46 (0.23–0.91), vertebral: 0.53 (0.42–0.67)</td>
<td>[31]</td>
</tr>
<tr>
<td></td>
<td>Proportion of inpatient cases Vertebral fracture 0.17 (0.11–0.24), forearm fracture 0.44 (0.19–0.51)</td>
<td>[37,64]</td>
</tr>
<tr>
<td></td>
<td>Discount rate 3% (0–7)</td>
<td>[28]</td>
</tr>
<tr>
<td></td>
<td>Compliance 0.38 (0.30–0.46)</td>
<td>[34]</td>
</tr>
<tr>
<td></td>
<td>Validity of X-ray of vertebral fracture Decision rule P(T</td>
<td>D=) = 0.66 (0.64–0.68), P(T'</td>
</tr>
<tr>
<td></td>
<td>Gradient of risk 60–70 years, hip: 1.95 (1.63–2.33), vertebral/forearm: 1.48 (1.39–1.58)</td>
<td>[4]</td>
</tr>
<tr>
<td></td>
<td>Fracture risk increase in women above the risk threshold compared with general population 60–70 years, hip: 4.15 (3.72–4.67), vertebral/forearm: 3.54 (3.43–3.66)</td>
<td>Calculated from [4,14]</td>
</tr>
<tr>
<td></td>
<td>Fracture risk increase in women above the risk threshold compared with general population 70–80 years, hip: 1.94 (1.78–2.11), vertebral/forearm: 1.70 (1.64–1.76)</td>
<td>[31]</td>
</tr>
<tr>
<td></td>
<td>Proportion of clinical vertebral fractures 0.35 (0.29–0.41)</td>
<td>[53]</td>
</tr>
<tr>
<td></td>
<td>Proportion of x-rays in women with back pain 60–70 years 0.35 (0.30–0.40)</td>
<td>[52]</td>
</tr>
</tbody>
</table>

*The incidence of hip fractures was not varied because the whole German population was sampled.

**Quality-adjusted-life years were discounted at 3%.

CI, confidence interval; CRF, clinical risk factor; DXA, dual x-ray absorptiometry; RG, gradient of risk; P, probability; T, test; D, disease.

Discussion

This analysis presents data on the cost-effectiveness of two screen-and-treat strategies in which the treatment threshold is defined by 10-year risk of fracture, based on CRFs alone or DXA plus CRFs. Compared with screening with DXA plus CRF, a strategy based on CRFs alone is more expensive and less effective in women above the age of 70 years. Note again that using CRFs implies treatment of all women above the age of 80 years. Compared with no screening, the cost-effectiveness of CRFs alone at an intervention threshold of ≥30% is fairly moderate in all age groups and similar to those of other generally accepted medical interventions such as hypertension treatment with thiazide diuretics in patients with systolic blood pressure of 140 mm Hg [68].

Conclusions

Treatment should be based on bone density as long as DXA is available for screening women above the age of 70 years. There-
fore, health policy should aim to increase the availability of DXA. Only in cases where DXA is not available, the usage of CRFs alone is justified. In women aged 60 to 70 years, where CRFs are superior, there is substantial uncertainty in the results. Because cost-effectiveness acceptability curves (CEACs) do not indicate the costs of making a wrong decision, funders should be careful in adopting CRFs alone instead of DXA plus CRFs.

In this analysis, ICERs were calculated based on an intervention threshold of ≥30%. Lowering this threshold would decrease specificity, and, as a result, treatment costs for false positives would largely increase. Thus, we would expect a lower threshold to be less cost-effective. In contrast, raising the threshold above 30% may improve the ICER in women aged ≥70 years because the number of false positives would decrease.

Table 4 Results of the base-case analysis (costs are presented in €, year 2006)*

<table>
<thead>
<tr>
<th>Intervention</th>
<th>Costs †</th>
<th>QALYs</th>
<th>Control</th>
<th>Incremental costs</th>
<th>Incremental QALYs</th>
<th>ICER (costs/QALY)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Women aged 60–70 years</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No screening</td>
<td>114,040</td>
<td>20.208</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CRFs alone</td>
<td>114,318</td>
<td>20.268</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DXA plus CRFs</td>
<td>118,431</td>
<td>20.471</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>i) No screening</td>
<td>4,391</td>
<td>0.263</td>
<td>16,696</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ii) CRFs alone</td>
<td>4,113</td>
<td>0.203</td>
<td>20,235</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Women aged 70–80 years</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No screening</td>
<td>103,144</td>
<td>13.276</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CRFs alone</td>
<td>107,683</td>
<td>13.490</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DXA plus CRFs</td>
<td>105,037</td>
<td>13.562</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>i) No screening</td>
<td>4,539</td>
<td>0.214</td>
<td>21,181</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ii) CRFs alone</td>
<td>1,893</td>
<td>0.286</td>
<td>6,611</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Women aged 80–90 years</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No screening</td>
<td>74,137</td>
<td>7.452</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Immediate treatment</td>
<td>76,512</td>
<td>7.686</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DXA plus CRFs</td>
<td>74,967</td>
<td>7.806</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>i) No screening</td>
<td>2,375</td>
<td>0.234</td>
<td>10,171</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ii) Immediate treatment</td>
<td>Dominates</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Incremental costs and QALYs are presented per high-risk woman treated over lifetime and were discounted at 3%.
†Costs in added years of life are included.
CRF, clinical risk factor; DXA, dual x-ray absorptiometry; ICER, incremental cost-effectiveness ratio; QALY, quality-adjusted life years.

Figure 3 (a) Cost (€) per quality-adjusted life year gained by combined 10-year risk for hip and vertebral fractures in women aged 60–70 (Δ = DXA plus CRFs, ▲ = CRFs alone). (b) Cost (€) per quality-adjusted life year gained by combined 10-year risk for hip and vertebral fractures in women aged 70–80 (DXA plus CRFs). (c) Cost (€) per quality-adjusted life year gained by combined 10-year risk for hip and vertebral fractures in women aged 80–90 (DXA plus CRFs). CRF, clinical risk factors; DXA, dual x-ray absorptiometry; QALY, quality-adjusted life years.
Nevertheless, these assumptions could not be tested because of a lack of data.

Our analysis has several strengths. It considers long-term follow-up costs of hip fractures in Germany including costs of hip implants, revision surgery, transportation services, rehabilitation care, and long-term care [45]. Moreover, in contrast to many economic analyses in the field of osteoporosis prevention, which did not consider the costs of case finding [69], our analysis includes all screening costs including costs of women being diagnosed incorrectly (false-positives). Furthermore, in contrast to the model by Zethraeus and colleagues [30], our model is able to consider the occurrence of vertebral and forearm fracture after hip fracture. The no-memory assumption of Markov models was avoided by creating health states that correspond to combined health states [70].

An important limitation is that the number of women at increased risk in the age group of 60 to 70 years is based solely on prior vertebral fractures. Additional risk factors to increase the number of women at high risk could not be used because their distribution in the German population has not been sufficiently evaluated yet [7]. This results in a lower ICER than that of screening in women aged 70 and 80 years although fracture risk increases with age and the ICERs of screening strategies usually decrease with age. The reason is that the cost-effectiveness ratio is very sensitive to specificity that is, based on this strategy, 95% for screening in women aged <70 years, but only 20% in women aged 70 years, and 0% in women aged 80 years. If it were possible to calculate the risk increase by combining several risk factors, the detection of more women at risk would be inevitably attenuated by a significant decrease of specificity in the identification of nonosteoporotic women, as shown in the evaluation of many decision rules for CRFs in osteoporosis screening [10,12].

There are a number of reasons why the results of our study are rather conservative. First, when calculating the costs of fracture treatment, analgesics were not considered. A recent cost analysis showed that osteoporotic patients receive three times more prescriptions for analgesics than nonosteoporotic patients do [71] and that those who have received nonsteroidal anti-inflammatory drugs were significantly more often hospitalized for peptic ulcer disease than those who have not received nonsteroidal anti-inflammatory drugs. Including these and other drug costs in the calculation of treatment costs for vertebral and forearm fractures would improve cost-effectiveness of CRFs with and without DXA.

Second, our analysis incorporated an increased risk of subsequent fractures for women who have suffered a prior fracture. This risk increase was determined only by the last fracture because there were no data available on the relationship between risk increase and the number of prior fractures. Consideration of

<table>
<thead>
<tr>
<th>Age group (years)</th>
<th>CRFs/no screening</th>
<th>DXA + CRFs/ no screening</th>
<th>DXA + CRFs/ CRFs alone</th>
</tr>
</thead>
<tbody>
<tr>
<td>60–70</td>
<td>1</td>
<td>46</td>
<td>45</td>
</tr>
<tr>
<td>70–80</td>
<td>220</td>
<td>57</td>
<td>-163</td>
</tr>
<tr>
<td>80–90</td>
<td>339</td>
<td>72</td>
<td>-267</td>
</tr>
<tr>
<td>Total</td>
<td>560</td>
<td>175</td>
<td>-385</td>
</tr>
</tbody>
</table>

CRF, clinical risk factor; DXA, dual x-ray absorptiometry.

![Figure 4](image-url)  
(a) Cost-effectiveness acceptability curves for screening and treatment in women aged 60–70 (b) Cost-effectiveness acceptability curves for screening and treatment in women aged 70–80 (c) Cost-effectiveness acceptability curves for screening and treatment in women aged 80–90 (CRFs, --- DXA plus CRFs, — no screening). CRF, clinical risk factors; DXA, dual x-ray absorptiometry.
two or more prior fractures would probably further increase the risk of a subsequent fracture and improve the cost-effectiveness ratio of CRFs with and without DXA. If these data become available, individual patient-level models could be more accurate in simulating the relationship between the number of prior fractures and subsequent fractures than between cohort-based approaches [72]. Moreover, by considering only the last fracture, in women with a prior hip or vertebral fracture who have a forearm fracture, the risk of subsequent fractures decreases. Continuing the higher risk increase of the prior fracture would improve the ICERs of CRFs and DXA plus CRFs.

Finally, the fracture states in this model are the same as in the reference model [30] except for fractures at “other” sites that were not modeled here. By considering only hip, vertebral, and forearm fractures, our analysis may have underestimated potential cost savings from the prevention of fractures at other skeletal sites. Although studies on alendronate did not show a significant reduction of other fractures [31], a study on risedronate, which also belongs to the class of bisphosphonates, reported a significant reduction of nonvertebral fractures by 39% (defined as fractures of the clavicle, humerus, wrist, pelvis, hip, or leg) [73].

On the other hand, there are also several reasons why a screen-and-treat strategy based on CRFs alone may be less cost-effective than that calculated by the base-case analysis. First, there is contradictory evidence whether individuals selected for treatment based on CRFs alone benefit from treatment or not. To date, efficacy data have to be taken from a population selected by the WHO criterion of BMD including individuals with osteopenia [21]. It remains unclear whether these data reflect the true efficacy of a population with a 30% long-term risk based on CRFs. In some efficacy trials, however, pharmacological interventions with bisphosphonates have been shown effective in patients not selected on the basis of low BMD [74–77].

Third, alendronate was assumed to be efficacious in all age groups although there have not been studies in women above the age of 80 [31].

Finally, for the state “long-term risk of ≥30%,” disutility of women aged 70 to 80 was underestimated. The reason is that an unknown number of women in this group suffer from prior fractures causing disutility. If a lower QoL was assumed, the incremental gain in QoL using CRFs with and without DXA compared with no intervention would be reduced, and, therefore, both strategies would become less cost-effective. Nevertheless, the increase of the ICER would be negligible, as shown in the sensitivity analysis.

The structure of this Markov cohort model is similar to that of an established reference model [30] although there are also some discrepancies: in our model, a state for other osteoporotic fractures was not modeled, and a woman may suffer a vertebral or forearm fracture after a hip fracture. As recommended there, a lifelong time horizon with a cycle length of 1 year was adopted, effectiveness was assumed to decrease linearly for a given “offset time,” and increased mortality after hip or vertebral fracture was modeled. In addition, for all fracture states, a postfracture state was included to reflect the sustained risk increase for subsequent fractures.

To compare the results of this analysis to those of other models, we used two cost-effectiveness studies for treatment with bisphosphonates in postmenopausal women [19,20], which are also based on long-term fracture risk. In these analyses, intervention thresholds for cost-effectiveness of bisphosphonates in women at different T-scores with or without prior fracture were calculated. Although the 5-year baseline risks of hip fracture were similar (e.g., 7.1% and 12.3% for women with a fracture history aged 70 and 80 years, respectively, compared with 6.7% and 12.1%, respectively, in our analysis) [19], the results of these analyses are quite different from our results. But again, there are important methodological differences compared with our study: costs of hip fracture are much lower given that costs beyond the first year after fracture were not included, treatment costs of false-positives were not considered, the discount rates used for costs (6%) and benefits (1.5%) differed from our study, and compliance was not considered in the base case. In addition, the selection of women at increased risk was based on several other risk factors such as BMI, history of peripheral fractures, use of oral glucocorticoids, and history of rheumatoid arthritis [19,20]. Consideration of several risk factors is likely to have decreased the specificity of the identification of nonosteoporotic women and thus increased the cost-effectiveness ratio.

If we assumed a societal perspective for our analysis, we would expect similar results. Copayments for drugs may be partly outweighed by savings for copayments for fracture treatment. Costs through loss of productivity are of minor importance because the proportion of women aged ≥60 years being employed in Germany is below 15% [78].

We do not know whether our findings are transferable to other countries. The main reason is that, for costs and epidemiological data, German sources were preferred as inputs to the model, so differences in resource consumption and prices may exist.

There are several important areas for future research. First, predicting fracture risk based on CRFs is still not accurate enough. Whereas risk factors such as a prior vertebral fracture and low BMD can be used for precise measurements of the risk increase [7], other risk factors such as low BMI and immobility only are known to be indicators of low BMD [7]. Thus, their interrelationship has yet to be formalized with more precision.

Second, effectiveness of alendronate has to be shown in clinical trials with patients selected on the basis of CRFs instead of low BMD [31].

Finally, this analysis compared different screen-and-treat strategies, based on an arbitrary risk threshold of 30% for hip or vertebral fractures. It would be of general interest to determine ICERs for different intervention thresholds of screen-and-treat strategies, using different ratios of hip to vertebral fracture risk or hip to nonhip–fracture risk. In our analysis, the ICERs of a combined risk of exact 30% were calculated for a specific ratio of hip to vertebral fracture risk. Nevertheless, different risk factors have different impacts on fracture sites and, thus, different impacts on QoL, costs, and mortality [20]. For example, if the absolute fracture risk threshold is reached by risk factors that are assumed to have a larger impact on the risk of hip fractures than on vertebral fractures (e.g., use of corticosteroids), treatment is more cost-effective because hip fractures have a higher impact on costs and mortality than fractures at other sites do [20].

In summary, CRFs are of considerable value for decision-making regarding the treatment of postmenopausal osteoporosis. Nevertheless, until the interrelationships between CRFs have been evaluated more extensively and until treatment with bisphosphonates in women selected by risk factors has shown to be as effective as in women selected by BMD, their usage should be combined with DXA. As recommended also by National Institute for Health and Clinical Excellence, providing treatment without DXA should be limited to older women only if the responsible clinician considers it to be clinically appropriate or unfeasible [20]. As long as DXA is available, the implementation of any approach based on CRFs alone will result in an uncontrolled increase in health expenditures.

Source of financial support: None
References


Clinical Risk Factors for Osteoporosis Screening


The Economic Value of Innovative Treatments over the Product Life Cycle: The Case of Targeted Trastuzumab Therapy for Breast Cancer

Louis P. Garrison Jr, PhD, David L. Veenstra, PharmD, PhD
University of Washington, Seattle, WA, USA

ABSTRACT

Objective: Pharmacoeconomic analyses typically project the expected cost-effectiveness of a new product for a specific indication. This analysis develops a dynamic life-cycle model to conduct a multindication evaluation using the case of trastuzumab licensed in the United States for both early-stage and metastatic (or late-stage) human epidermal growth factor receptor 2 (HER2)-positive breast cancer therapy (early breast cancer [EBC]; metastatic breast cancer [MBC]), approved in 2006 and 1998, respectively.

Methods: This dynamic model combined information on expected incremental cost-utility ratios for specific indications with an epidemiologically based projection of utilization by indication over the product life cycle—from 1998 to 2016. Net economic value was estimated as the cumulative quality-adjusted life years (QALYs) gained over the life cycle multiplied by a societal valuation of health gains ($/QALY) minus cumulative net direct treatment costs. Sensitivity analyses were performed under a range of assumptions.

Results: We projected that the annual number of EBC patients receiving trastuzumab will be more than three times that of MBC by 2016, in part because adjuvant treatment reduces the future incidence of MBC. Over this life cycle, the estimated overall incremental cost-effectiveness ratio (ICER) was $35,590/QALY with a total of 432,547 discounted QALYs gained. Under sensitivity analyses, the overall ICER varied from $21,000 to $53,000/QALY, and the projected net economic value resulting from trastuzumab treatment ranged from $6.2 billion to $49.5 billion.

Conclusions: Average ICERs for multiindication compounds can increase or decrease over the product life cycle. In this example, the projected overall life-cycle ICER for trastuzumab was less than one half of that in the initial indication. This dynamic perspective versus the usual static one—highlights the interdependence of drug development decisions and investment incentives, raising important reimbursement policy issues.

Keywords: cost-utility analysis, economics, modeling, pharmaceutical pricing, product life cycle.

Introduction

Most economic evaluations involving cost-effectiveness analyses of new pharmaceutical products are performed shortly following product launch for a specific indication covered in the license. They are based on mathematical models that project clinical and economic outcomes from results of phase III randomized clinical trials for a typical patient for the expected horizon of clinical impact. Indeed, such models are often a key element in dossiers submitted to public and private payers for purposes of coverage and reimbursement.

In the past decade, we have witnessed the licensing of many new, innovative advances in biologic agents that work in multiple cancer types and indications. When these products are launched for the initial indication, manufacturers will set the price by which cost-effectiveness is judged. From a payer’s perspective, this evaluation is logical and useful as they see each indication as a separate “purchase” on behalf of their beneficiaries. However, this process has some limitations and implications from a broader societal and global perspective that are underappreciated. First, the aggregate economic value delivered by a new medicine will ultimately be determined by the different types and number of patients using it over its life cycle, and this may include totally new indications (often at different doses). Second, not only does the cost-effectiveness of a product vary among individual patients, but also the economic value will typically vary systematically across indications. Because the price per milligram of a pill often cannot be varied across indications, the actual cost-effectiveness achieved will generally vary. This creates a dilemma for manufacturers working within a reimbursement environment that simultaneously fixes the price at launch for the duration of the product life cycle and evaluates the product by applying a cost-effectiveness threshold: should potential future indications have some impact on the proposed initial product price?

In oncology, for example, both for reasons of safety and for risk–benefit, it is common to first conduct clinical trials in the most severely ill patients (e.g., patients with metastatic disease who have failed first- or second-line therapies). If efficacy is demonstrated at reasonable tolerability in this situation, then the therapy can be tested at earlier stages of disease progression, for example, moving from later stage, metastatic use to treatment at early diagnosis, called “adjuvant” use (e.g., in combination with surgery). This process of drug development and testing usually takes 8–12 years for the initial indication, and several more years for each subsequent indication [1]. Once a therapy is approved in an early-stage, adjuvant setting, the characteristics of the patients who progress to metastatic disease and those individuals who are newly diagnosed with metastases may be quite different from patients originally diagnosed with metastatic disease at the beginning of the product life cycle or during the registration clinical trials. Early therapy will affect the later incidence of metastatic disease as well as the treatment of adjuvant patients. Moreover, before a product is approved for use in the adjuvant setting, there can also be changes in the standard of care in the metastatic setting.

This analysis takes a broader long-term perspective, asking what is the overall cost-effectiveness across multiple indications...
throughout the product life cycle and what is the aggregate economic value delivered. The objective was to develop a dynamic life-cycle modeling (DLM) approach, and to apply it to a case example. We evaluated a targeted cancer therapy, trastuzumab (Herceptin®, Genentech, Inc., South San Francisco, CA), a monoclonal antibody approved for treatment of human epidermal growth factor receptor 2 (HER2)-positive breast cancer in combination with chemotherapy. This analysis considers both the treatment for metastatic HER2-positive breast cancer (approved by the Food and Drug Administration in 1998) and the more recently approved (2006) indication for adjuvant treatment of early-stage HER2-positive breast cancer. We chose this example because trastuzumab has a more recently approved second indication, national-level epidemiological data were available, and data from suitable indication-specific cost-effectiveness analyses were available.

Study Data and Methods

Background and Model Overview

A DLM approach combines information on expected incremental cost-effectiveness ratios (ICERs) with an epidemiologically based projection of utilization by indication over the product life cycle. The effective product life cycle for biologics is not fixed given the current lack of a regulatory pathway in the United States for so-called “biosimilars” or “follow-on biologics” and the potential entrance of competing branded biologics in the same class. For purposes of this exercise, the “end” of the product life cycle is assumed to be 10 years after the launch in adjuvant treatment (i.e., through 2016). This DLM projection thus has three major components: a forecast of the volume of trastuzumab use over the product life cycle—from initial launch in 1998 to 2016, an estimate of the average cost-effectiveness (i.e., the average ICER) in metastatic treatment over this period, and an estimate of the average ICER in adjuvant treatment following approval in this indication eight years after the initial launch.

Projecting Disease Incidence

Approximately 20% to 25% of patients with breast cancer will have HER2-positive disease, which is associated with a poor prognosis [2,3]. Trastuzumab is a monoclonal antibody that targets HER2, and is approved for the treatment of HER2-positive metastatic breast cancer either as first-line therapy in combination with paclitaxel or as second- or third-line monotherapy. For this analysis, we relied on previously published and publicly available data to project the economic impact of trastuzumab.

To our knowledge, there are no published projections of the volume of long-term trastuzumab use in the United States. The approach we used projects an increasing volume of the use of trastuzumab from 1998 to 2016 based on estimates of the anticipated annual incidence of metastatic breast cancer and early-stage breast cancer in the United States. These estimates were derived from Surveillance Epidemiology and End Results (SEER) data from the National Cancer Institute, and assumed that 25% of patients tested are HER2-positive [4]. In the years prior to approval of the adjuvant indication, the volume of utilization was based solely on metastatic breast cancer patients. After approval of the adjuvant indication in November 2006, the volume of use was projected to shift to include women receiving adjuvant therapy for EBC as well as women receiving first-line metastatic treatment, whether initially diagnosed, or for recurrent disease. Furthermore, given that data from the joint analysis of the National Surgical Adjuvant Breast and Bowel Project B-31 and the North Central Cancer Treatment Group N9831 trials demonstrated that the addition of trastuzumab to a standard adjuvant regimen reduced the risk of recurrence by 52% ($P < 0.001$) and improved survival by 33%, there is an anticipated reduction in the number of future metastatic breast cancer patients [5].

Projecting Volume of Trastuzumab Use

The projection of the volume of use over the product life cycle was based on demographic projections, epidemiological estimates, and assumptions about use rates among candidate (i.e., HER2-positive) patients. The female population was divided into five age groups: less than 21, 21–39, 40–54, 55–64, and more than 64. Data on US census age-specific subpopulations through 2016 were aggregated to within these age groups [6]. An analysis of the SEER registry for the period 1999–2001 provided the incidence rates of breast cancer used in the analysis: these are shown in Table 1 for both EBC and newly diagnosed MBC patients. An assumption was made that 25% of the previously diagnosed early patients in SEER would later suffer a recurrence. An analysis of recent SEER data was used to estimate the share of EBC patients who are node-negative low-risk, node-negative high-risk, and node positive. Based on these shares, the use rate in the base case for HER2-positive patients was assumed to be 60% for all patients in the metastatic setting and 70% for node-positive and high-risk node-negative patients in the adjuvant setting. This also reflects the impact of a number of factors, including that not all women will be tested, some will have

<table>
<thead>
<tr>
<th>Table 1</th>
<th>Input parameters for volume of use projection</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input parameters</td>
<td>Base-case</td>
</tr>
<tr>
<td>Incidence rates (per 100,000 women) for early breast cancer* by age:</td>
<td>SEER 1999–2001</td>
</tr>
<tr>
<td>&lt;21 years</td>
<td>0.02</td>
</tr>
<tr>
<td>21–39 years</td>
<td>23.2</td>
</tr>
<tr>
<td>40–54 years</td>
<td>161.5</td>
</tr>
<tr>
<td>55–64 years</td>
<td>371.1</td>
</tr>
<tr>
<td>65+ years</td>
<td>372.0</td>
</tr>
<tr>
<td>Incidence rates (per 100,000 women) for newly diagnosed metastatic disease*:</td>
<td>SEER 1999–2001</td>
</tr>
<tr>
<td>&lt;21 years</td>
<td>0</td>
</tr>
<tr>
<td>21–39 years</td>
<td>1.1</td>
</tr>
<tr>
<td>40–54 years</td>
<td>6.6</td>
</tr>
<tr>
<td>55–64 years</td>
<td>15.1</td>
</tr>
<tr>
<td>65+ years</td>
<td>19.4</td>
</tr>
<tr>
<td>Incidence of metastatic disease among women diagnosed previously</td>
<td>Assumption</td>
</tr>
<tr>
<td>Incidence of HER2-positive breast cancer among women diagnosed with breast cancer</td>
<td>25%</td>
</tr>
<tr>
<td>Early breast cancer</td>
<td>SEER</td>
</tr>
<tr>
<td>distribution at diagnosis:</td>
<td></td>
</tr>
<tr>
<td>Node-negative, low-risk</td>
<td>17%</td>
</tr>
<tr>
<td>Node-negative, high-risk</td>
<td>50%</td>
</tr>
<tr>
<td>Node positive</td>
<td>33%</td>
</tr>
<tr>
<td>Trastuzumab utilization rate (%) by status:</td>
<td>Assumptions</td>
</tr>
<tr>
<td>Node-negative, low-risk</td>
<td>0%</td>
</tr>
<tr>
<td>Node-negative, high-risk</td>
<td>70%</td>
</tr>
<tr>
<td>Node positive</td>
<td>70%</td>
</tr>
<tr>
<td>Metastatic</td>
<td>60%</td>
</tr>
</tbody>
</table>

*Both HER2-positive and HER2-negative.

SEER = Surveillance Epidemiology and End Results; HER2-positive, human epidermal growth factor receptor positive.
comorbidities that limit trastuzumab use, some may not have adequate insurance coverage, and some will choose not to receive treatment.

**Indication-Specific Cost-Utility Ratios**

Several studies report on the cost-effectiveness of trastuzumab in either the metastatic [7,8] or adjuvant settings [9-11] using Markov models to assess the incremental lifetime costs and QALYs of the addition of trastuzumab in metastatic and early-stage adjuvant breast cancer. The MBC models were based on the data from the trials reported in Slamon et al. (2001) [12]. The EBC models were based on the data from the trials reported in Romond et al. (2006) [5]. We developed the base case MBC and EBC cost-effectiveness ratios for trastuzumab based on the models described above.

The number of QALYs gained and the incremental lifetime cost per QALY gained were estimated from the studies cited above that used Markov models, relying in varying degrees of clinical trial data and published aggregate results. These assumptions are summarized in Table 2 for both metastatic and adjuvant treatment. The Markov models included costs for: 1) HER2 testing (using immunohistochemistry and/or fluorescence in situ hybridization); 2) trastuzumab therapy based on average wholesale price for medication costs and Medicare reimbursement rates for procedures and resources; 3) patients with EBC who were assumed to receive one year of trastuzumab in the adjuvant setting and were treated with trastuzumab until disease progression in the metastatic setting; 4) adverse event monitoring; and 5) treatment of adverse events. In the adjuvant setting, costs for treating metastatic disease were also included for those patients projected to progress over time. The indication-specific outcome measures for the model were QALYs and were based on reported survival, probability of recurrence for EBC, time to progression for MBC, and adverse events including the incidence of cardiac dysfunction.

The base case estimates for the costs and QALYs for trastuzumab in the metastatic and adjuvant indications are presented in Table 2. There were multiple sources available for the cost-effectiveness of trastuzumab in the metastatic setting. The assumed MBC ICER represents a combination of two studies, using utility weights from Elkin et al. [7] and survival estimates from Hornberger et al. [8], but also with updated drug cost. As Hornberger et al. had access to the original trial data and used propensity scoring to adjust for crossovers following progression, their survival estimates were used. For the cost-effectiveness of trastuzumab in the adjuvant setting, all data elements were based on the work of Garrison et al. [9], whose estimated ICER falls between two other recently published studies [10,11] and for which we had access to full model, which is necessary for these calculations.

### Dynamic Life-Cycle Cost-Utility

We examined life-cycle cost-effectiveness using several measures. The primary measure was an overall life-cycle cost-effectiveness ratio that was discounted to 1998 based on the projected costs and QALYs for all patients receiving trastuzumab between 1998 and 2016. This overall, cumulative ICER was computed by multiplying the mean QALYs gained and mean costs separately by the estimated number of patients with HER2-positive breast cancer in the respective adjuvant and metastatic settings in each year. QALYS gained and costs were discounted at 3% and were cumulated separately and then divided to calculate the overall life-cycle ICER.

**Overall Life-Cycle ICER**

\[
\text{Overall Life-Cycle ICER} = \frac{\text{Discounted } \Sigma (\text{MBC Costs} + \text{EBC Costs})}{\text{Discounted } \Sigma (\text{MBC QALYs} + \text{EBC QALYs})}
\]

Where:

- \(\text{MBC QALYs} = \text{mean discounted MBC QALY per patient} \times \Sigma (\text{MBC patients over the 19-year period})\)
- \(\text{EBC QALYs} = \text{mean discounted EBC QALY per patient} \times \Sigma (\text{EBC patients over the 19-year period})\)
- \(\text{MBC Costs} = \text{mean discounted MBC Cost per patient} \times \Sigma (\text{MBC patients over the 19-year period})\)
- \(\text{EBC Costs} = \text{mean discounted EBC Cost per patient} \times \Sigma (\text{EBC patients over the 19-year period})\)

We also calculated this figure on a "history to date" basis to examine how it changes over time. For comparison, we calculated two alternative "naive" estimates to understand how a proper accounting over the life cycle would compare. These two measures were a lifetime weighted average and annual weighted average, both without discounting to 1998 and based only on the mean ICERs and volume of use. Proper calculation of the overall life-cycle ICER requires specific information on both the numerator and denominators for each of the two indications.

**Aggregate Economic Value**

We defined net economic value as the potential "social surplus" as defined in economics as the sum of consumer surplus and producer surplus. Essentially, it is the amount by which aggregate societal willingness to pay for benefits exceeds the costs of providing them. This requires an assumption about the willingness to pay for a QALY gained. As there is no consensus about a specific value, we use a range of $50,000-$150,000, reflecting the variation of what analysts have used in practice [13,14]. The gross life-cycle economic value is defined as the discounted QALYs gained multiplied by the mean threshold value \(T\) that society places on a QALY:

**Gross Economic Value**

\[
\text{Gross Economic Value} = (T \times \text{ndMBC} + T \times \text{ndEBC})
\]

Where \(T\) is the threshold value society places on a QALY and is varied between $50,000/QALY and $150,000/QALY.

\(\text{ndMBC} = \text{discounted } \Sigma \text{MBC QALYs over the 19-year period}\)

\(\text{ndEBC} = \text{discounted } \Sigma \text{EBC QALYs over the 19-year period}\)
The net economic value was thus defined as the difference between this gross willingness to pay and the incremental costs of providing trastuzumab over the entire period, that is, the projected sales of trastuzumab minus the net of other treatment costs (discounted at 3%).

**Impact of Adjuvant Use on Metastatic Use**

Adjuvant use of trastuzumab is expected to reduce the downstream use of trastuzumab in the metastatic indication. Projected ICERs for EBC must include an assumption about whether future MBC patients will receive trastuzumab. For example, Garrison et al. [9] assumed this for their base case. This raises a question of potential double-counting in the epidemiological projections. In the life-cycle incidence forecasts, we have distinguished between newly diagnosed MBC and previously diagnosed MBC patients. This could lead to an overstatement of both the costs of trastuzumab and the benefits in terms of QALYs gained. We attempted to estimate the potential size of this bias by performing the life-cycle calculation without including the previously diagnosed patients, assuming they were fully reflected in the CE ratio for adjuvant patients.

**Sensitivity Analyses**

Our approach to sensitivity analyses was to vary deterministically four critical drivers: the uptake of trastuzumab in metastatic and adjuvant use, and the mean ICERs in metastatic and adjuvant use. The former ranged from 50% to 90%, while the mean ICER in metastatic use was varied from $70,000 to $115,000, and was varied from $15,000 to $40,000 in adjuvant use. These ranges should encompass the bulk of the uncertainty generated by all of the underlying variables in each of these components.

**Results**

As shown in Figure 1, by 2016, the number of patients treated annually with adjuvant trastuzumab was projected to be approximately three times the number of patients treated with trastuzumab in the metastatic setting. Over the entire period, an estimated 319,000 US women with HER2-positive breast cancer are projected to receive adjuvant trastuzumab and 161,000 are projected to receive treatment for metastatic breast cancer. Between 1998 and 2016, the cumulative net cost of trastuzumab was projected to be $15.4 billion and the cumulative QALYs gained were projected to be 432,547. The cumulative incremental cost-utility ratio was $35,590 per QALY gained. Changes in this cumulative ratio over time are reflected in Figure 2. As this is below most commonly cited thresholds, trastuzumab use appears to have a net surplus. When these QALYs were valued at varying rates of societal willingness to pay from $50,000 to $150,000, the projected gross economic value of trastuzumab treatment due to QALY gains ranged from $21.6 billion to $64.9 billion. Subtracting the cumulative net cost of trastuzumab, the economic value to society is projected to be between $6.2 billion to $49.5 billion. Or viewed alternatively, the proportional reward to the manufacturer represents as much as 71% of the social surplus generated to as little as about 24%, depending on the societal willingness to pay for QALYs.

The overall ICER results compared to the indication-specific estimates are depicted in Figure 3. Given the change in volumes of use for the two indications, this estimate is lower than either naive estimate—a mean annual ICER for 2016 of $40,652 per QALY or a mean cumulative (nondiscounted) ICER of $46,262 per QALY.

In the sensitivity analyses, as shown in Table 3, the overall life-cycle ICER varied from $32,914 to $42,129 as the utilization rate varied, and the overall life-cycle ICER varied from $21,210 to $52,842 as the indication-specific ICERs were varied.

Excluding metastatic patients who were previously diagnosed to adjust for potential double-counting reduces the overall life-cycle ICER by 18% to $29,357, and reduces total QALYs gained by 10% to 384,675. This also decreases the gross life-cycle economic value from $21.6 billion to $19.2 billion at the low
end, and from $64.9 billion to $57.7 billion at the high end. However, this probably overstates the reduction somewhat as Garrison et al. [9] conservatively assume that all patients who initially received adjuvant trastuzumab—as well as those who did not—receive trastuzumab after metastatic progression.

**Discussion**

We developed a DLM approach and applied the methodology to the breast cancer drug trastuzumab. Specifically, we found that the cumulative, life-cycle ICER trastuzumab was $35,590, with a range of $21,000–$53,000 in sensitivity analyses. The gross life-cycle value was estimated to be between $21.6 billion and $64.9 billion, with a corresponding net economic value of $6.2 billion–$49.5 billion. The overall life-cycle ICER for trastuzumab is less than one half of the projected ICER in the initial indication, and the aggregate economic value is much greater as a result of the second adjuvant indication. These results suggest that indication-specific models may have significant limitations for informing policy decisions for drugs with multiple indications licensed over time.

Applying a DLM approach to consider an innovative therapeutic agent highlights the importance of perspective. A short-term perspective focusing on indication-specific cost-effectiveness offers a different view than a longer-term perspective that recognizes the interdependence of drug development decisions and investment incentives. US private payers are generally expected to take a short-term perspective, considering value for money spent on an indication-by-indication basis. Reimbursement systems, particularly in Europe, do not adjust prices or reimbursement rates in response to new information on cost-effectiveness with new indications. In the United States, manufacturers can and do often increase the prices of branded products greater than the rate of general inflation. The net economic surplus would be lower in real terms if the real price of trastuzumab increases over this period. In any case, establishing evidence- and value-based reimbursement systems in this situation, particularly with fixed prices across indications at any point in time, is difficult.

Reimbursement systems that do not account for changing value across indications or over time may produce suboptimal, long-term societal outcomes. Under the patent system, drug prices for branded products can also be seen as a reward for innovation, with potentially far-reaching implications for incentives to undertake future innovative research and development. For example, the calculations of Philipson and Jena [15] for HIV drugs suggest that manufacturers are receiving only about 5% of the social surplus they create. They argue that this low proportion provides a much smaller incentive for innovation. In this trastuzumab case, the estimated range of surplus reward to the manufacturer varied between 24% and 71%, depending on the threshold value for a QALY. A role for the dynamic, long-term perspective is exemplified in oncology where early clinical trials are focused on the sickest patients, followed only many years later in patients diagnosed earlier where the potential benefit is greater. This has significant implications for reimbursement systems, the value of information generated by additional research, and incentives for investment. The recent debate between Claxton and Towse [16,17] about the “value-based” pricing proposal of the UK Office of Fair Trading report [18] has raised the possibility of ex post payments based on performance, including potentially different payments for different indications. The potential positive effects on incentives for investment have been noted by Thornton [19]. Lundin and Ramsberg [20] have recently argued theoretically that a

---

**Table 3** Sensitivity of life-cycle ICER to use rates (as a percentage of incident population) and to variations in indication-specific ICERs (base case in italics)

<table>
<thead>
<tr>
<th>Sensitivity to use rates as percentage of incident population</th>
<th>MBC</th>
</tr>
</thead>
<tbody>
<tr>
<td>EBC</td>
<td></td>
</tr>
<tr>
<td>90% $32,914 $33,849 $34,751 $36,465 $38,662 $40,188 $42,129</td>
<td></td>
</tr>
<tr>
<td>70% $34,471 $35,590 $36,662 $38,674</td>
<td></td>
</tr>
<tr>
<td>60% $35,567 $36,808 $37,989 $40,188</td>
<td></td>
</tr>
<tr>
<td>50% $37,008 $38,398 $39,710 $42,129</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sensitivity to variations in indication-specific ICERs</th>
<th>MBC ICER</th>
</tr>
</thead>
<tbody>
<tr>
<td>EBC ICER $70,000 $85,676 $115,000</td>
<td></td>
</tr>
<tr>
<td>$15,000 $21,210 $21,658 $22,191</td>
<td></td>
</tr>
<tr>
<td>$26,417 $34,398 $35,590 $37,053</td>
<td></td>
</tr>
<tr>
<td>$40,000 $47,602 $49,917 $52,842</td>
<td></td>
</tr>
</tbody>
</table>

Source: Authors’ calculations.

ICER, incremental cost-effectiveness ratio; MBC, metastatic breast cancer; EBC, early breast cancer.
“dynamic cost-effectiveness rule” could improve incentives for research and development.

There are several limitations to our analysis. First, these calculations are based on projections and may not represent actual practice in terms of either cost-effectiveness achieved or future volumes of use. The model is only from a US perspective, although a more global view would be appropriate for considering returns to R&D. The model also does not estimate any impact on volume of trastuzumab use or a revaluation of consumer surplus if there are follow-on competitive compounds (including the entry of branded competitors). If such a competition were to lower use, then the aggregate surplus generated by trastuzumab would be lower, although if price competition lowered prices, then the share of surplus going to the manufacturer would be reduced. And the model does not include any consideration of the use of trastuzumab in other indications, including “off-label” indications such as third- or fourth-line metastatic treatment, where formal testing has not been conducted. An oncology drug could be less cost-effective in such indications, which would raise its overall life-cycle cost-effectiveness ratio, although potentially still increasing aggregate net economic value.

Also, the time horizon to 2016 was chosen to represent a plausible period of limited competition from follow-ons which would also require years of testing to establish efficacy in both of these indications. A slightly shorter effective patent protection would reduce the gross economic value generated, but would have limited impact on the life-cycle ICER or the share of surplus reward, which varies more with the threshold value for a QALY. In addition, our estimates took a payer perspective, which did not include any indirect cost savings or time costs. Presumably, improved cancer survival improves the labor force participation and contributions of these patients, increasing the overall social surplus. This analysis was conducted in real terms (2006 US dollars) discounted to 1998. As is customary, “inflation” in either drug or other medical prices is not included in the projections. Also, we did not consider the impact of any fall in the real price of trastuzumab if there were any competition from a biosimilar product during this period, or of any real increase in price because, as discussed above, manufacturers in the United States often increase prices of branded products over time. Furthermore, we did not model the impact of the potential competitor lapatinib, an oral chemotherapeutic agent recently approved for HER2-positive patients whose metastatic disease has progressed after receiving regimens including trastuzumab.

Our calculations to consider the impact of double-counting suggest that our projection of the life-cycle ICER could be too conservative (i.e., biased upward) since we assume that all recurrent MBC patients receive trastuzumab. Note that although we project that 84% of metastatic use in 2016 would be for previously diagnosed patients, the vast majority of overall use by then would be for EBC patients receiving adjuvant therapy. Nonetheless, this could be an important area for future research, as it would be useful to have a reliable estimate of the impact of improved adjuvant outcomes on metastatic incidence and costs.

Conclusions

Indication-specific cost-utility or cost-effectiveness models do not account for the important interdependence of drug development and expanded indications over time: the development and approval of subsequent indications is contingent upon success in those developed first. Average ICERS for multi-indication compounds can increase or decrease over the product life cycle. This is especially true in oncology where initial research on safety and efficacy occurs in the most critically ill patients: only after efficacy is demonstrated in this situation can new therapies be tested at earlier stages of disease. The field of pharmacoeconomics and reimbursement policies should give greater attention to dynamic, long-term aspects of drug pricing and reimbursement policies and how they affect incentives for innovation and drug development.

The authors wish to thank Deborah Labeck, Deepa Lalla, and Carolina Reyes for their help with this research, and Marlene Gyldmark and Jamie Cross for useful comments on earlier drafts.

Source of financial support: This research was supported by unrestricted funding to the University of Washington from Genentech, Inc. This research was presented in part at The International Society for Pharmacoeconomics and Outcomes Research, European Meeting, October, 2006.

References

14 Ubel PA. What is the price of life and why doesn’t it increase at the rate of inflation? Arch Intern Med 2003;163:1637–41.
17 Towsa A. If it ain’t broke, don’t price fix it: the OCT and the PPRS. Health Econ 2007;16:653–65.
Costs and Health Utilities Associated with Extremely Preterm Birth: Evidence from the EPICure Study

Stavros Petrou, PhD,1,2 Giselle Abangma, MSc,2 Samantha Johnson, PhD,3 Dieter Wolke, PhD,4 Neil Marlow, FMedSci5

1National Perinatal Epidemiology Unit, University of Oxford, Oxford, UK; 2Health Economics Research Centre, University of Oxford, Oxford, UK; 3Academic Division of Child Health, Queens Medical Centre, University of Nottingham, UK; 4Department of Psychology and Health Science Research Institute, University of Warwick, UK; 5Academic Neonatology, Institute of Women’s Health, University College London, London, UK

ABSTRACT

Objective: To estimate costs and health utilities associated with extremely preterm birth at approximately 11 years of age using evidence from a whole population study (the EPICure study).

Methods: The study population comprised surviving children born at 20 through 25 completed weeks of gestation in all 276 maternity units in the United Kingdom and Republic of Ireland from March through December 1995 and a control group of classmates born at full term, matched for age, sex, and ethnic group. Estimates of utilization of health, social, and education services were combined with unit costs derived from primary and secondary sources. Generalized linear regression was used to estimate the impact of extremely preterm birth on public sector costs during the 11th year of life. Suboptimal levels of function for each of the eight attributes of the Health Utilities Index Mark III (cognition, vision, hearing, speech, ambulation, dexterity, emotion, and pain) and multiplicative multi-attribute utility scores were compared between the extremely preterm children and their classmates. Tobit regressions were performed to explore the effects of gestational age at birth on the Health Utilities Index Mark III multiattribute utility score.

Results: Mean (standard deviation [SD]) public sector costs over the 12-month period were £6484 (£3548) for the combined extremely preterm group and £4007 (£2537) for their classmates, generating a mean cost difference of £2477 (bootstrap 95% confidence interval [CI] £1605, £3360) that was statistically significant (P < 0.001). The generalized linear models revealed that compared to birth at term, birth at ≤23 completed weeks, 24th–24th weeks and 25th–25th weeks gestation increased public sector costs by an average of £2417 (95% CI £60, £4774; P = 0.044), £1528 (95% CI £129, £2927; P = 0.032) and £1501 (95% CI £428, £2574; P = 0.006), respectively. In all eight attributes of the Health Utilities Index Mark III, there were significantly higher proportions of suboptimal levels of function among the extremely preterm children (P ≤ 0.05). The mean (SD) multiattribute utility score for the extremely preterm children as a cohort was 0.789 (0.264), compared to 0.956 (0.102) for the classmates born at term, a mean difference in utility score of 0.167 (95% CI 0.124, 0.209) that was statistically significant (P < 0.001). The Tobit regressions revealed that, compared to birth at term, birth at ≤23 completed weeks, 24th–24th weeks and 25th–25th weeks gestation reduced the Health Utilities Index Mark III multiattribute utility score by an average of 0.312 (95% CI 0.169, 0.455; P < 0.001), 0.337 (95% CI 0.235, 0.439; P < 0.001) and 0.243 (95% CI 0.159, 0.327; P < 0.001), respectively.

Conclusions: The results of this study should be used to inform the development of future economic evaluations of interventions aimed at preventing extremely preterm birth or alleviating its effects.

Keywords: costs, extremely preterm birth, health-related quality of life, health utilities.

Introduction

The incidence of preterm birth, defined as birth before 37 completed weeks of gestation, has been reported at between 5% and 11% throughout the industrialized world with some of the highest rates reported in the United States [1]. The incidence of preterm birth has increased slightly since the 1980s, which has been attributed to increasing rates of multiple births, greater use of assisted reproduction, and increased obstetric intervention, such as induced labor and Caesarean section [2,3]. Developments in clinical practice, such as the use of ultrasonography to estimate gestational age, may also have had a slight effect on the incidence rate as may the trend toward registering a live birth for infants born at the limits of viability whose deaths might previously been classified as stillbirths or miscarriages [2]. Preterm birth has been associated with socioeconomic disadvantage, non-Caucasian ethnic background, substance misuse, and extremes in maternal age, with both teenage and older mothers at an increased risk [4,5].

Although the majority of preterm births occur between 33 and 36 completed weeks of gestation, it is possible for infants born as early as 22 weeks gestation to survive and the clinical outcomes following birth at extremely low gestational ages are those most widely reported in the literature. A substantial body of literature has reported that infants born at extremely low gestational ages are at an increased risk of a range of adverse neonatal outcomes including chronic lung disease [6], severe brain injury [7], retinopathy of prematurity [8], necrotizing enterocolitis [9], and neonatal sepsis [10]. In later life, these infants are also at an increased risk of motor and sensory impairment [11,12], learning difficulties [13–17], and behavioral problems [18–21]. Although the clinical sequelae of extremely preterm birth are well documented, relatively little is known about its consequences in cost or utility metrics that can be incorporated into cost-effectiveness modeling studies. Two recent systematic reviews of the economic literature in this area highlight the focus of studies upon costs incurred during the neonatal period with few attempts to estimate the costs of preterm birth during later childhood [22,23]. In addition, studies that have estimated the health-related quality of life of premature infants using preference-based measures have tended to categorize children in terms of birth weight, rather than gestational age at birth, despite
the limited prognostic capabilities of the birth weight measure [24–28]. One recent study described the health-related quality of life of British teenagers in mainstream schooling who were born before 29 weeks gestational age using the Health Utilities Index [29]. However, these children were identified from cohorts born in the early 1980s prior to improvements in perinatal practices and intensive care technologies, thus limiting the generalizability of the study results to the current clinical context.

The purpose of this research is to augment the limited economic evidence in this area by estimating costs and health utilities associated with extremely preterm birth during mid-childhood. In so doing, we provide a significant new resource to analysts modeling the cost-effectiveness of preventive or treatment interventions for extremely preterm birth.

**Methods**

**Study Population**

Children that participated in the EPICure study formed the basis of this investigation. The EPICure study is a whole population study of all infants born at 20 through 25 completed weeks of gestation in all 276 maternity units in the United Kingdom and Republic of Ireland from March to December 1995. A full description of the study population, recruitment methods, and neonatal assessment procedures is available elsewhere [30]. Of 307 surviving children, 241 (78.2%) were assessed at a median age of 6 years and 4 months (range: 5 years and 2 months to 7 years and 3 months) [31], while 219 (71.3%) were assessed at a median age of 10 years and 11 months (range: 10 years and 1 month to 12 years and 1 month) [32]. At the final follow-up, surviving children underwent a series of assessments of cognitive and functional disability and were then classified into four groups of overall disability (none, mild, moderate, and severe) on the basis of the most severe classification in any domain [32]. A control group of 153 mainstream school classmates who were born at full term and matched for age, sex, and ethnic group was also evaluated using the same assessment procedures. Ethical approval for the study was obtained from the Southampton and South West Hampshire Research Ethics Committee and approved by the Central Office for Research Ethics Committees.

**Estimation of Costs**

As part of the follow-up assessments conducted at 11 years, the main parent (usually the mother) was asked to complete a detailed postal questionnaire about their child’s resource utilization over the previous year. The questionnaire was piloted to ascertain its acceptability, comprehensiveness, and reliability, and reminder letters were sent to parents to increase the response and completion rates. The data collected from the main parent included their child’s use of hospital inpatient and day care services, community health services, prescribed medications, social services, and education services. All resource use data were entered directly from the research instruments into a purpose-built data collection program with in-built safeguards against inconsistent entries and then verified by dual coding. Estimates of service provision were derived from these data and usually expressed in terms of contact hours. For all hospital admissions, estimates of service provision were expressed in terms of patient days with part of a day at each level of care counted as a 24-hour period. For education services, estimates of services provision reflected the level of educational assistance within each type of educational establishment (mainstream school, mainstream school with special unit attached, special school for the physically disabled, and special school for children with learning difficulties).

UK unit costs were applied to each resource item to value total resource use for each study child over an annual basis. All unit costs employed followed recent guidelines on costing public services as part of economic evaluation [33–35]. The calculation of these costs was underpinned by the concept of opportunity cost, which can be defined as the value of the next best alternative for using these resources [34]. The costs of hospital inpatient and day care services were largely derived from English Department of Health reference costs based upon National Health Service trust financial returns [36]. The unit costs of community health and social services were largely derived from national sources [37], and took account of time spent by professionals on indirect activities, such as traveling and paper work. Some unit costs of health and social services were calculated from first principles using established accounting methods [33]. Drug costs were obtained from the British National Formulary [38]. Educational costs were based upon micro-costing exercises for different types of educational establishment and were obtained from the Department of Education and Skills in England (Department for Education and Skills in London, England, pers. comm.). All costs were expressed in pounds sterling and reflected values for the financial year 2006 to 2007. No inflation or deflation of costs to 2006 to 2007 prices using indices such as the National Health Service Hospital and Community Health Services Pay and Prices Index was required.

**Estimation of Health Utilities**

The postal questionnaire completed by the main parent around the child’s 11th birthday included a preference-based measure of health-related quality of life, namely the Health Utilities Index. The main parent was considered the appropriate subject for describing the child’s health-related quality of life as related research had indicated that the comprehension level for the Health Utilities Index is somewhat high for a pediatric sample where a number of children have developmental disabilities [39,40].

The Health Utilities Index is a family of preference-based multi-attribute utility measures [41]. The principal caregiver completed the unedited 15-item questionnaire for proxy-assessed usual health status assessment, which was obtained from the Health Utilities Index developers and covers both Mark II and Mark III health status classification systems. The “usual” health focus of the questions has previously been applied in population health surveys, where short-term illnesses such as the flu are not the major concern [42]. The Mark III classification system is now recommended by the developers because of its broad applicability in both clinical and general population health studies, improvements in a number of definitions, and an increased orthogonality of its attributes for structural independence [42]. It has been used in previous studies of extremely low birth weight children internationally [28]. The Health Utilities Index Mark III health status classification system covers eight attributes: cognition, vision, hearing, speech, ambulation, dexterity, emotion, and pain. Function within each attribute is graded on a 5- or 6-point scale corresponding to level of severity, ranging from normal function (level 1) to severe impairment (levels 5 or 6). Responses to the Health Utilities Index Mark III health status classification system were converted into multiplicative multi-attribute utility scores using a published utility function [43,44]. These multi-attribute utility scores are based on the permutation of responses across the eight attributes and are expressed on an interval scale ranging from −0.36 (representing the health state with the lowest level of function for all attributes) to 1.00 (representing the health state with the highest level of function for all attributes).
The multi-attribute utility scoring algorithm for the Health Utilities Index Mark III can be summarized as \( u^* = 1.371(b_1 \times b_2 \times b_3 \times b_4 \times b_5 \times b_6 \times b_7 \times b_8) - 0.371 \), where \( u^* \) is the utility score for the overall health state being measured and the \( b_i \)'s are substituted from a table of coefficients provided by the Health Utilities Index developers for the appropriate attribute and level [44]. To develop the multi-attribute utility scoring algorithm, a random sample of 504 general population adults living in the city of Hamilton, Canada had previously been asked to value selected health states using both a visual analog scaling technique and a standard gamble instrument. Further details on the utility algorithms for the Health Utilities Index Mark III are reported elsewhere [43,44].

**Statistical Analysis**

Differences in baseline sociodemographic and clinical characteristics between the extremely preterm children and their classmates were tested using the Pearson chi-squared test. Comparisons of each category of public sector costs and of total public sector costs were made between the extremely preterm children and their classmates, as well as between four pre-specified groups of children of varying gestational age at birth (≤23 completed weeks, 24–24.9 weeks, 25–25.9 weeks, term). Comparisons of costs are reported as mean values with standard deviations (SD) and mean differences in costs between the comparison groups with 95% confidence intervals (CIs) where applicable. As the data for costs were skewed, in addition to Student’s t-tests of cost differences, nonparametric bootstrap estimation was used to derive 95% CIs for mean cost differences between the comparison groups [45]. The bootstrap method does not rely on parametric assumptions concerning the underlying distribution of data, hence its usefulness for generating CIs for skewed data [46]. Using a large number of simulations, and based on sampling with replacement from the original data, the bootstrap method estimates the distribution of a sampling statistic [46]. Using a large number of simulations, and based on sampling with replacement from the original data, the bootstrap method estimates the distribution of a sampling statistic [46].

Each of the CIs surrounding mean cost differences was calculated using 1000 bias-corrected bootstrap replications. In addition, the comparison groups were tested using two-sample t-tests of unequal variance. Finally, we performed Tobit regressions to explore the effects of gestational age at birth on the Health Utilities Index Mark III utility score. Tobit regression was required to account for the censoring of the dependent variable, the utility score, which has an upper value of 1.0 [47]. As with costs, two regression models for health utilities were performed, the first applying the dichotomous variable for gestational age at birth, and the second applying the further disaggregated variable for gestational age at birth. The same covariates applied in the generalized linear regression on costs were applied in the Tobit regressions on health utilities.

All analyses were performed with a microcomputer using the Statistical Package for the Social Sciences (SPSS) (version 15.0; SPSS Inc, Chicago, IL) software and STATA (version 9.0; College Station, TX: StataCorp LP) software. P-values of 0.05 or less were considered statistically significant.

**Results**

Postal questionnaires reporting costs and health utilities were returned for 190 extremely preterm children (representing 61.9% of eligible surviving children and 86.8% of children assessed for cognitive and functional disability at a median age of 10 years and 11 months) and 141 classmates (representing 92.2% of classmates recruited into the concurrent clinical study [32]). Examination of longitudinal data for the extremely preterm children revealed that those not assessed at a median age of 10 years and 11 months were more likely to be born at 25–25.9 weeks, be of nonwhite ethnic origin, have had an operation for necrotizing enterocolitis, to have unemployed parents, and to have had lower cognitive scores or cognitive impairment at 2.5 and 6 years (P ≤ 0.05). There were no significant differences in the sociodemographic and clinical characteristics at 11 years between the extremely preterm children and their classmates, for whom postal questionnaires were returned, with the exception of language spoken at home, overall disability status and gestational age a birth (Table 1).

The resource use values for each category of resource use for the extremely preterm children and their classmates, and the respective unit costs of each resource item, are presented in Table 2. Mean public sector costs over the 12-month period were £7514 (SD: £5662) among children born at ≤23 completed weeks, £6821 (±5984) among children born at 24–24.9 weeks, £6132 (±5304) among children born at 25–25.9 weeks, and £4007 (±2537) among the classmates born at term. Mean public sector costs over the 12-month period were £6484 (£5548) for the combined extremely preterm group and £4007 (±2537) for their classmates, generating a mean cost difference of £2477 (bootstrap 95% CI £1605, £3360) that was statistically significant (P < 0.001) (Table 3). When the data were analyzed by cost category, extremely preterm birth was associated with an increase of £297 in hospital inpatient care costs (bootstrap 95% CI £115, £523; P = 0.007), £108 in hospital outpatient and day care costs (bootstrap 95% CI £46, £166; P = 0.001), £405 in total hospital costs (bootstrap 95% CI £204, £668; P = 0.001), £344 in community health and social care costs (bootstrap 95% CI £223, £461; P < 0.001), £760 in total health and social care.
costs (bootstrap 95% CI £493, £1052; \( P < 0.001 \)), and £1716 in education costs (bootstrap 95% CI £877, £2517; \( P < 0.001 \)). A more detailed breakdown of the costs of each study group is available upon request.

Relationships between the clinical and sociodemographic characteristics of the study population and total public sector costs over the 12-month period are shown in Table 4. The generalized linear models revealed that even after controlling for clinical and sociodemographic confounders, extremely preterm birth was associated with significantly increased public sector costs. Model 1 revealed that, compared to birth at term, extremely preterm birth increased public sector costs by an

---

Table 1: Sociodemographic and clinical characteristics of study participants; n (%)

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Extremely preterm children (n = 190)</th>
<th>Classmates (n = 141)</th>
<th>P-value*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>86 (45.3)</td>
<td>59 (41.8)</td>
<td>0.535</td>
</tr>
<tr>
<td>Female</td>
<td>104 (54.7)</td>
<td>82 (58.2)</td>
<td></td>
</tr>
<tr>
<td>Maternal marital status</td>
<td></td>
<td></td>
<td>0.105</td>
</tr>
<tr>
<td>Married</td>
<td>119 (62.6)</td>
<td>104 (73.8)</td>
<td></td>
</tr>
<tr>
<td>Single</td>
<td>10 (5.3)</td>
<td>11 (7.8)</td>
<td></td>
</tr>
<tr>
<td>Cohabiting</td>
<td>24 (12.6)</td>
<td>9 (6.4)</td>
<td></td>
</tr>
<tr>
<td>Widowed</td>
<td>3 (1.6)</td>
<td>1 (0.7)</td>
<td></td>
</tr>
<tr>
<td>Separated/divorced</td>
<td>32 (16.8)</td>
<td>16 (11.3)</td>
<td></td>
</tr>
<tr>
<td>Missing</td>
<td>2 (1.1)</td>
<td>0 (0.0)</td>
<td></td>
</tr>
<tr>
<td>Respondent parent’s age (years)</td>
<td></td>
<td></td>
<td>0.525</td>
</tr>
<tr>
<td>&lt;30</td>
<td>2 (1.1)</td>
<td>2 (1.4)</td>
<td></td>
</tr>
<tr>
<td>30–39</td>
<td>73 (38.4)</td>
<td>57 (40.4)</td>
<td></td>
</tr>
<tr>
<td>40–49</td>
<td>95 (50.0)</td>
<td>73 (51.8)</td>
<td></td>
</tr>
<tr>
<td>≥50</td>
<td>10 (5.3)</td>
<td>3 (2.1)</td>
<td></td>
</tr>
<tr>
<td>Missing</td>
<td>10 (5.3)</td>
<td>6 (4.3)</td>
<td></td>
</tr>
<tr>
<td>Type of accommodation</td>
<td></td>
<td></td>
<td>0.133</td>
</tr>
<tr>
<td>Owner occupied</td>
<td>133 (70.0)</td>
<td>109 (77.3)</td>
<td></td>
</tr>
<tr>
<td>Rented</td>
<td>41 (21.6)</td>
<td>19 (13.5)</td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td>12 (6.3)</td>
<td>12 (8.5)</td>
<td></td>
</tr>
<tr>
<td>Missing</td>
<td>4 (2.1)</td>
<td>1 (0.7)</td>
<td></td>
</tr>
<tr>
<td>Access to car</td>
<td></td>
<td></td>
<td>0.602</td>
</tr>
<tr>
<td>Yes</td>
<td>178 (93.7)</td>
<td>133 (94.3)</td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>9 (4.7)</td>
<td>5 (3.5)</td>
<td></td>
</tr>
<tr>
<td>Missing</td>
<td>3 (1.6)</td>
<td>3 (2.1)</td>
<td></td>
</tr>
<tr>
<td>Highest parental qualification</td>
<td></td>
<td></td>
<td>0.688</td>
</tr>
<tr>
<td>Vocational/NVQ/CSE</td>
<td>29 (15.3)</td>
<td>17 (12.1)</td>
<td></td>
</tr>
<tr>
<td>O-level/GCSE/Scottish standards</td>
<td>51 (26.8)</td>
<td>40 (28.4)</td>
<td></td>
</tr>
<tr>
<td>BTEC diploma/A-level/Scottish higherers</td>
<td>30 (15.8)</td>
<td>16 (11.3)</td>
<td></td>
</tr>
<tr>
<td>Diploma or HND</td>
<td>19 (10.0)</td>
<td>14 (9.9)</td>
<td></td>
</tr>
<tr>
<td>University degree</td>
<td>20 (10.5)</td>
<td>22 (15.6)</td>
<td></td>
</tr>
<tr>
<td>Postgraduate qualification</td>
<td>12 (6.3)</td>
<td>9 (6.4)</td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td>10 (5.3)</td>
<td>12 (8.5)</td>
<td></td>
</tr>
<tr>
<td>None</td>
<td>12 (6.3)</td>
<td>8 (5.7)</td>
<td></td>
</tr>
<tr>
<td>Missing</td>
<td>7 (3.7)</td>
<td>3 (2.1)</td>
<td></td>
</tr>
<tr>
<td>Highest household occupational status</td>
<td></td>
<td></td>
<td>0.064</td>
</tr>
<tr>
<td>Professional/manageral</td>
<td>79 (41.6)</td>
<td>77 (54.6)</td>
<td></td>
</tr>
<tr>
<td>Intermediate occupation</td>
<td>44 (23.2)</td>
<td>23 (16.3)</td>
<td></td>
</tr>
<tr>
<td>Routine and manual occupation</td>
<td>46 (24.2)</td>
<td>35 (24.8)</td>
<td></td>
</tr>
<tr>
<td>Long-term unemployed</td>
<td>11 (5.8)</td>
<td>3 (2.1)</td>
<td></td>
</tr>
<tr>
<td>Missing</td>
<td>10 (5.3)</td>
<td>3 (2.1)</td>
<td></td>
</tr>
<tr>
<td>Language spoken at home†</td>
<td></td>
<td></td>
<td>0.023</td>
</tr>
<tr>
<td>English only</td>
<td>162 (85.3)</td>
<td>132 (93.6)</td>
<td></td>
</tr>
<tr>
<td>English and other language(s)</td>
<td>25 (13.2)</td>
<td>8 (5.7)</td>
<td></td>
</tr>
<tr>
<td>Missing</td>
<td>3 (1.6)</td>
<td>1 (0.7)</td>
<td></td>
</tr>
<tr>
<td>Number of smokers in home</td>
<td></td>
<td></td>
<td>0.437</td>
</tr>
<tr>
<td>0</td>
<td>115 (60.5)</td>
<td>91 (64.5)</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>39 (20.5)</td>
<td>27 (19.1)</td>
<td></td>
</tr>
<tr>
<td>≥2</td>
<td>31 (16.3)</td>
<td>16 (11.3)</td>
<td></td>
</tr>
<tr>
<td>Missing</td>
<td>5 (2.6)</td>
<td>7 (5.0)</td>
<td></td>
</tr>
<tr>
<td>Disability status at 11 years</td>
<td></td>
<td></td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>None</td>
<td>33 (17.4)</td>
<td>102 (72.3)</td>
<td></td>
</tr>
<tr>
<td>Mild</td>
<td>80 (42.1)</td>
<td>37 (26.2)</td>
<td></td>
</tr>
<tr>
<td>Moderate</td>
<td>55 (28.9)</td>
<td>2 (1.4)</td>
<td></td>
</tr>
<tr>
<td>Severe</td>
<td>22 (11.6)</td>
<td>0 (0.0)</td>
<td></td>
</tr>
<tr>
<td>Gestational age at birth</td>
<td></td>
<td></td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>22+0–22+6 weeks</td>
<td>1 (0.5)</td>
<td>—</td>
<td></td>
</tr>
<tr>
<td>23+0–23+6 weeks</td>
<td>18 (9.5)</td>
<td>—</td>
<td></td>
</tr>
<tr>
<td>24+0–24+6 weeks</td>
<td>59 (31.1)</td>
<td>—</td>
<td></td>
</tr>
<tr>
<td>25+0–25+6 weeks</td>
<td>112 (58.9)</td>
<td>—</td>
<td></td>
</tr>
<tr>
<td>Term</td>
<td>—</td>
<td>141 (100.0)</td>
<td></td>
</tr>
</tbody>
</table>

*χ² test.
†Distinguished from ethnicity for which the comparison groups were matched. To our knowledge, all study children spoke English with varying levels of proficiency.

A-level, advanced level; BTEC, Business and Technology Education Council; CSE, certificate of secondary education; GCSE, general certificate of secondary education; HND, higher national diploma; NVQ, national vocational qualification; O-level, ordinary level.
Table 2  Resource use and unit costs of resource items (UK pound sterling, 2006 to 2007 prices). Resource use values are given as means (SD) unless otherwise stated

<table>
<thead>
<tr>
<th>Resource use variable, unit</th>
<th>Extremely preterm children (n = 190)</th>
<th>Classmates (n = 141)</th>
<th>Unit cost or range†</th>
</tr>
</thead>
<tbody>
<tr>
<td>Community and social care services</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>General practitioner, contacts</td>
<td>1.91 (1.28)</td>
<td>1.24 (1.40)</td>
<td>£34.00 per contact1</td>
</tr>
<tr>
<td>Practice nurse, contacts</td>
<td>0.37 (1.11)</td>
<td>0.27 (0.70)</td>
<td>£30.00 per contact1</td>
</tr>
<tr>
<td>Community nurse, contacts</td>
<td>0.03 (0.19)</td>
<td>0.02 (0.14)</td>
<td>£23.00 per contact1</td>
</tr>
<tr>
<td>Community pediatrician, contacts</td>
<td>0.2 (0.54)</td>
<td>0.06 (0.47)</td>
<td>£242.90 per contact1</td>
</tr>
<tr>
<td>Dentist, contacts</td>
<td>1.74 (1.22)</td>
<td>1.65 (0.93)</td>
<td>£70.02 per contact1</td>
</tr>
<tr>
<td>Orthodontist, contacts</td>
<td>0.25 (0.73)</td>
<td>0.29 (1.08)</td>
<td>£135.00 per contact1</td>
</tr>
<tr>
<td>Optician, contacts</td>
<td>0.86 (1.52)</td>
<td>0.51 (0.76)</td>
<td>£26.15 per contact1</td>
</tr>
<tr>
<td>Chiropodist, contacts</td>
<td>0.06 (0.54)</td>
<td>0.01 (0.08)</td>
<td>£16.00 per contact1</td>
</tr>
<tr>
<td>Physiotherapist, contacts</td>
<td>1.35 (6.07)</td>
<td>0.08 (0.56)</td>
<td>£40.00 per contact1</td>
</tr>
<tr>
<td>Speech therapist, contacts</td>
<td>1.20 (4.57)</td>
<td>0.01 (0.12)</td>
<td>£40.00 per contact1</td>
</tr>
<tr>
<td>Audiologist, contacts</td>
<td>0.24 (0.74)</td>
<td>0.04 (0.35)</td>
<td>£170.67 per contact1</td>
</tr>
<tr>
<td>Social worker, contacts</td>
<td>0.27 (1.25)</td>
<td>0.09 (1.01)</td>
<td>£39.00 per contact1</td>
</tr>
<tr>
<td>Home visitor/volunteer, contacts</td>
<td>0.04 (0.51)</td>
<td>0.00 (0.00)</td>
<td>£19.00 per contact1</td>
</tr>
<tr>
<td>Counselor, contacts</td>
<td>0.26 (2.40)</td>
<td>0.03 (0.34)</td>
<td>£34.00 per contact1</td>
</tr>
<tr>
<td>Psychologist, contacts</td>
<td>0.18 (0.88)</td>
<td>0.04 (0.35)</td>
<td>£67.00 per contact1</td>
</tr>
<tr>
<td>Psychiatrist, contacts</td>
<td>0.03 (0.20)</td>
<td>0.01 (0.17)</td>
<td>£256.00 per contact1</td>
</tr>
<tr>
<td>Osteopath, contacts</td>
<td>0.14 (1.16)</td>
<td>0.04 (0.51)</td>
<td>£42.89 per contact1</td>
</tr>
<tr>
<td>Home teacher (portage), contacts</td>
<td>0.11 (1.45)</td>
<td>0.00 (0.00)</td>
<td>£36.61 per contact1</td>
</tr>
<tr>
<td>Home teacher (other), contacts</td>
<td>2.56 (11.74)</td>
<td>0.38 (4.38)</td>
<td>£36.61 per contact1</td>
</tr>
<tr>
<td>Orthoptist, contacts</td>
<td>0.08 (0.46)</td>
<td>0.01 (0.08)</td>
<td>£57.57 per contact1</td>
</tr>
<tr>
<td>Orthotist, contacts</td>
<td>0.16 (0.99)</td>
<td>0.00 (0.00)</td>
<td>£30.24 per contact1</td>
</tr>
<tr>
<td>Other community healthcare professionals, contacts</td>
<td>0.32 (1.52)</td>
<td>0.03 (0.27)</td>
<td>£16.00–135.00 per contact1†,‡</td>
</tr>
<tr>
<td>Hospital outpatient and day care services</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Accident and emergency care, attendances</td>
<td>0.18 (0.46)</td>
<td>0.29 (1.00)</td>
<td>£39.25 per attendance1</td>
</tr>
<tr>
<td>Hospital day unit, attendances</td>
<td>0.10 (0.77)</td>
<td>0.00 (0.00)</td>
<td>£34.00–395.00 per attendance1</td>
</tr>
<tr>
<td>Other outpatient care, attendances</td>
<td>0.84 (1.87)</td>
<td>0.22 (0.85)</td>
<td>£34.00–395.00 per attendance1†,‡</td>
</tr>
<tr>
<td>Hospital inpatient services</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Breathing difficulties, days</td>
<td>0.04 (0.51)</td>
<td>0.00 (0.00)</td>
<td>£329.90–6195.11 per day1</td>
</tr>
<tr>
<td>Surgery, days</td>
<td>0.17 (1.10)</td>
<td>0.02 (0.19)</td>
<td>£329.90–6195.11 per day1</td>
</tr>
<tr>
<td>ICU, days</td>
<td>0.00 (0.00)</td>
<td>0.00 (0.00)</td>
<td>£329.90–6195.11 per day1</td>
</tr>
<tr>
<td>Long-term illness, days</td>
<td>0.28 (3.13)</td>
<td>0.04 (0.51)</td>
<td>£329.90–6195.11 per day1</td>
</tr>
<tr>
<td>Other, days</td>
<td>0.23 (2.03)</td>
<td>0.01 (0.12)</td>
<td>£329.90–6195.11 per day1</td>
</tr>
<tr>
<td>Education services</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mainstream school, n (%)</td>
<td>158 (83.2)</td>
<td>136 (96.5)</td>
<td>£3,152.00 per annum1</td>
</tr>
<tr>
<td>Mainstream school with special unit attached, n (%)</td>
<td>15 (7.9)</td>
<td>5 (3.5)</td>
<td>£16,434.00 per annum1</td>
</tr>
<tr>
<td>Special school, n (%)</td>
<td>17 (8.9)</td>
<td>0 (0.0)</td>
<td>£16,434.00 per annum1</td>
</tr>
</tbody>
</table>

*Bootstrap estimation using 1000 replications, bias corrected.
†Values calculated using Student’s t-test.
‡Confidence interval.
§Primary research.
ICU, intensive care unit; SD, standard deviation.

average of £1608 (95% CI £686, £2530; P = 0.001). Model 2 revealed that, compared to birth at term, birth at ≥23 completed weeks, 24–24.6 weeks and 25–25.6 weeks gestation increased public sector costs by an average of £2417 (95% CI £660, £4774; P = 0.044), £1528 (95% CI £129, £2927; P = 0.032) and £1501 (95% CI £428, £2574; P = 0.006), respectively. No other clinical or sociodemographic covariate had a significant impact upon public sector costs with the exception of long-term unemployment of both parents, which was associated with an average increase of £5088 (95% CI £193, £9983; P = 0.042) in model 1 and £5264 (95% CI £314, £10,213; P = 0.037) in model 2.

Comparisons of the frequency and proportion of suboptimal levels of function between the extremely preterm children and their classmates born at term are shown in Table 5 for each of the eight attributes of the Health Utilities Index Mark III. In all eight attributes (vision, hearing, speech, emotion, pain, ambulation,
Table 4  Clinical and sociodemographic factors predicting public sector costs (UK pound sterling, 2006 to 2007 prices) during the 11th year of life, generalized linear model with gamma distribution and linear (identity) link function

<table>
<thead>
<tr>
<th>Factor</th>
<th>Model 1</th>
<th>Model 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Adjusted regression coefficient*</td>
<td>Robust SE</td>
</tr>
<tr>
<td>Gestational age at birth</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Classmate norm§</td>
<td></td>
<td></td>
</tr>
<tr>
<td>All extremely preterm (model 1)</td>
<td>1.607.9</td>
<td>470.6</td>
</tr>
<tr>
<td>23 weeks (model 2)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>24 weeks (model 2)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>25 weeks (model 2)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gender</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male†</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Single</td>
<td>2,303.8</td>
<td>1,286.9</td>
</tr>
<tr>
<td>Cohabiting</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Widowed</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Separated/divorced</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Respondent parent’s age§</td>
<td></td>
<td></td>
</tr>
<tr>
<td>≤30†</td>
<td></td>
<td></td>
</tr>
<tr>
<td>30–39</td>
<td></td>
<td></td>
</tr>
<tr>
<td>≥50</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Type of accommodation§</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Owner occupied†</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tenant</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Access to car§</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes†</td>
<td></td>
<td></td>
</tr>
<tr>
<td>No†</td>
<td>−1,283.8</td>
<td>1,192.8</td>
</tr>
<tr>
<td>Highest parental qualification§</td>
<td></td>
<td></td>
</tr>
<tr>
<td>O-level/GCSE/Sc standards</td>
<td>30.0</td>
<td>721.4</td>
</tr>
<tr>
<td>BTEC diploma/A-level/Sc higher§</td>
<td>351.0</td>
<td>849.7</td>
</tr>
<tr>
<td>Diploma or HND §</td>
<td>−373.5</td>
<td>885.1</td>
</tr>
<tr>
<td>University degree§</td>
<td>−325.5</td>
<td>893.8</td>
</tr>
<tr>
<td>Postgraduate qualification§</td>
<td>2,186.7</td>
<td>1,325.9</td>
</tr>
<tr>
<td>Other§</td>
<td>1,098.5</td>
<td>1,082.1</td>
</tr>
<tr>
<td>Highest occupational status§</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Professional/managerial§</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intermediate occupation§</td>
<td>265.1</td>
<td>647.2</td>
</tr>
<tr>
<td>Routine and manual occupation§</td>
<td>−5.0</td>
<td>571.1</td>
</tr>
<tr>
<td>Long-term unemployed§</td>
<td>5,088.0</td>
<td>2,497.5</td>
</tr>
<tr>
<td>Language spoken at home§</td>
<td></td>
<td></td>
</tr>
<tr>
<td>English only§</td>
<td>−231.5</td>
<td>861.3</td>
</tr>
<tr>
<td>English and other language(s)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of smokers in home§</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0†</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1‡</td>
<td>−232.8</td>
<td>587.0</td>
</tr>
<tr>
<td>2‡</td>
<td>1,493.6</td>
<td>822.7</td>
</tr>
<tr>
<td>Constant</td>
<td>6,381.1</td>
<td></td>
</tr>
</tbody>
</table>

*Interpreted as the additional costs over and above the reference group after adjustment for covariates.

†Reference group.

‡Includes child born at ≥21–22+ weeks.

§Cases with missing information omitted from analyses.

A-level, advanced level; BTEC, Business and Technology Education Council; CI, confidence interval; CSE, certificate of secondary education; GCSE, general certificate of secondary education; HND, higher national diploma; NVQ, national vocational qualification; O-level, ordinary level; Sc, Scottish; SE, standard error.
dexterity, and cognition), there were significantly higher proportions of suboptimal levels of function among the extremely preterm children ($P \leq 0.05$). When each extremely preterm subgroup was compared to the classmates born at term, there were significantly higher proportions of suboptimal levels of function across all attributes with the exception of hearing ($P = 0.214$), pain ($P = 0.430$), ambiguity ($P = 0.119$) and dexterity ($P = 0.070$) for children born at 23 completed weeks gestation, emotion ($P = 0.058$) for children born at 24th–24th weeks gestation and hearing ($P = 0.089$) for children born at 25th–25th weeks gestation.

Table 6 presents descriptive statistics for the multi-attribute utility scores for the comparison groups. These multi-attribute utility scores summarize population preferences for the overall health state of the child across the eight attributes of the Health Utilities Index Mark III. The mean (SD) multiattribute utility score for the extremely preterm children as a cohort was 0.789 (0.264), compared to 0.956 (0.102) for the classmates born at term, a mean difference in utility score of 0.167 (95% CI 0.124, 0.209) that was statistically significant ($P < 0.001$). The mean multiattribute utility scores were also significantly lower in each of the three individual extremely preterm subgroups than in the classmates born at term ($P \leq 0.05$).

Finally, the Tobit regressions revealed that even after controlling for clinical and sociodemographic confounders, extremely preterm birth was associated with significantly reduced multiattribute utility scores (Table 7). Model 1 revealed that, compared to birth at term, extremely preterm birth reduced the Health Utilities Index Mark III multiattribute utility score by an average of 0.280 (95% CI 0.204, 0.357; $P < 0.001$). Model 2 revealed that, compared to birth at term, birth at 23 completed weeks, 24th–24th weeks and 25th–25th weeks gestation reduced the Health Utilities Index Mark III multiattribute utility score by an average of 0.312 (95% CI 0.169, 0.455; $P < 0.001$), 0.337 (95% CI 0.235, 0.439; $P < 0.001$) and 0.243 (95% CI 0.159, 0.327; $P < 0.001$), respectively. The only other clinical or sociodemographic covariates associated with significantly reduced multiattribute utility scores were living in rented rather than owner occupied accommodation, a highest parental qualification of “other” compared to vocational or equivalent, a highest parental occupational status of routine and manual compared to professional or managerial, and a highest parental occupational status of long-term unemployed compared to professional or managerial ($P \leq 0.05$).

**Discussion**

Although the clinical sequelae of extremely preterm birth are well documented, relatively little is known about its consequences in cost or utility metrics that can inform cost-effectiveness modeling studies. Previous studies of the costs of extremely preterm birth have been criticized for their relatively poor methodological quality when assessed against current guidelines for conduct of health economic studies [22,23]. They have also been criticized for their narrow perspective, and their focus upon costs incurred during the neonatal period with few attempts to estimate the costs of the condition during later childhood [22,23]. Previous studies that have estimated the health-related quality of life associated with extremely preterm birth using preference-based measures have tended to categorize children in terms of birth weight, rather than gestational age at birth, and have identified children born prior to recent improvements in perinatal practices and intensive care technologies, thereby limiting their generalizability to the current clinical context [24–29].

This study estimated the costs and health utilities associated with extremely preterm birth on the basis of the largest (to our knowledge) cohort study of extremely preterm children in the world [30–32]. The children were drawn from defined geographical areas, namely the whole of the United Kingdom and Republic of Ireland, rather than clinic-based populations and,

---

**Table 5** Number (%) of children with suboptimal levels of function* within each Health Utilities Index Mark III attribute

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Gestational age at birth</th>
<th>All extremely preterm</th>
<th>Classmates</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>=23 weeks†</td>
<td>24 weeks</td>
<td>25 weeks</td>
</tr>
<tr>
<td>Vision</td>
<td>7 (36.8)</td>
<td>19 (32.2)</td>
<td>35 (31.2)</td>
</tr>
<tr>
<td>Hearing</td>
<td>1 (5.3)</td>
<td>5 (8.5)</td>
<td>5 (4.5)</td>
</tr>
<tr>
<td>Speech</td>
<td>5 (26.3)</td>
<td>16 (27.1)</td>
<td>16 (14.3)</td>
</tr>
<tr>
<td>Emotion</td>
<td>5 (26.3)</td>
<td>9 (15.3)</td>
<td>21 (18.8)</td>
</tr>
<tr>
<td>Pain</td>
<td>3 (15.8)</td>
<td>14 (23.7)</td>
<td>23 (20.5)</td>
</tr>
<tr>
<td>Ambulation</td>
<td>1 (5.3)</td>
<td>8 (13.6)</td>
<td>7 (6.3)</td>
</tr>
<tr>
<td>Dexterity</td>
<td>2 (10.5)</td>
<td>7 (11.9)</td>
<td>10 (8.9)</td>
</tr>
<tr>
<td>Cognition</td>
<td>9 (47.4)</td>
<td>33 (55.9)</td>
<td>57 (50.9)</td>
</tr>
</tbody>
</table>

*Suboptimal levels of function defined as less than normal (below level 1) function for each attribute.
†Includes child born at 22nd–23rd weeks.
‡Calculated using two-sample t-test for unequal variance.

---

**Table 6** Health Utilities Index Mark III multiattribute utility scores

<table>
<thead>
<tr>
<th>Group</th>
<th>N</th>
<th>Mean</th>
<th>(SD)</th>
<th>Mean decrement from classmates</th>
<th>P-value‡</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classmates</td>
<td>141</td>
<td>0.956</td>
<td>0.0102</td>
<td></td>
<td></td>
</tr>
<tr>
<td>=23 weeks†</td>
<td>19</td>
<td>0.772</td>
<td>0.291</td>
<td>-0.184</td>
<td>0.016</td>
</tr>
<tr>
<td>24 weeks</td>
<td>58</td>
<td>0.717</td>
<td>0.333</td>
<td>-0.239</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>25 weeks</td>
<td>112</td>
<td>0.830</td>
<td>0.208</td>
<td>-0.126</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>All extremely preterm</td>
<td>190</td>
<td>0.789</td>
<td>0.264</td>
<td>-0.167</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>

*Calculated using Fisher’s exact test comparing all children born extremely preterm with classmate controls.
†Includes child born at 22nd–23rd weeks.
‡Calculated using two-sample t-test for unequal variance.
Table 7 Clinical and sociodemographic factors predicting Health Utilities Index Mark III multiattribute utility scores during the 11th year of life, Tobit regression

<table>
<thead>
<tr>
<th>Factor</th>
<th>Model 1</th>
<th></th>
<th>Model 2</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Gestational age at birth</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Classmate norms†</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All extremely preterm (model 1)</td>
<td>-0.280</td>
<td>0.039</td>
<td>(-0.357,-0.204)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>≥23 weeks (model 2)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24 weeks (model 2)</td>
<td>-0.312</td>
<td>0.073</td>
<td>(-0.455,-0.169)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>25 weeks (model 2)</td>
<td>-0.337</td>
<td>0.052</td>
<td>(-0.439,-0.235)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Gender</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male‡</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female†</td>
<td>0.098</td>
<td>0.036</td>
<td>(0.028,0.168)</td>
<td>0.006</td>
</tr>
<tr>
<td>Maternal marital status§</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Married†</td>
<td>0.035</td>
<td>0.105</td>
<td>(-0.172,0.243)</td>
<td>0.737</td>
</tr>
<tr>
<td>Single‡</td>
<td>-0.070</td>
<td>0.061</td>
<td>(-0.189,0.050)</td>
<td>0.253</td>
</tr>
<tr>
<td>Cohabiting‡</td>
<td>0.056</td>
<td>0.150</td>
<td>(-0.239,0.351)</td>
<td>0.708</td>
</tr>
<tr>
<td>Widowed‡</td>
<td>0.028</td>
<td>0.056</td>
<td>(-0.083,0.138)</td>
<td>0.624</td>
</tr>
<tr>
<td>Separated/divorced‡</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Respondent parent’s age‡</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;30‡</td>
<td>0.076</td>
<td>0.182</td>
<td>(-0.283,0.434)</td>
<td>0.678</td>
</tr>
<tr>
<td>30–39‡</td>
<td>0.013</td>
<td>0.183</td>
<td>(-0.347,0.372)</td>
<td>0.945</td>
</tr>
<tr>
<td>≥40‡</td>
<td>-0.115</td>
<td>0.200</td>
<td>(-0.508,0.278)</td>
<td>0.364</td>
</tr>
<tr>
<td>Type of accommodation§</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Owner occupied†</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rented‡</td>
<td>-0.131</td>
<td>0.058</td>
<td>(-0.244,-0.017)</td>
<td>0.024</td>
</tr>
<tr>
<td>Other‡</td>
<td>-0.081</td>
<td>0.071</td>
<td>(-0.220,0.039)</td>
<td>0.254</td>
</tr>
<tr>
<td>Access to car§</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes‡</td>
<td>0.307</td>
<td>0.124</td>
<td>(0.062,0.552)</td>
<td>0.014</td>
</tr>
<tr>
<td>No‡</td>
<td>0.276</td>
<td>0.124</td>
<td>(0.032,0.521)</td>
<td>0.027</td>
</tr>
<tr>
<td>Highest parental qualification§</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>O-level/GCSE/Sc standards</td>
<td>-0.095</td>
<td>0.060</td>
<td>(-0.214,0.024)</td>
<td>0.117</td>
</tr>
<tr>
<td>BTEC diploma/A level/Sc higherers</td>
<td>-0.133</td>
<td>0.069</td>
<td>(-0.268,0.002)</td>
<td>0.053</td>
</tr>
<tr>
<td>Diploma or HND</td>
<td>-0.181</td>
<td>0.077</td>
<td>(-0.332,-0.030)</td>
<td>0.019</td>
</tr>
<tr>
<td>University degree</td>
<td>-0.091</td>
<td>0.078</td>
<td>(-0.244,0.062)</td>
<td>0.245</td>
</tr>
<tr>
<td>Postgraduate qualification</td>
<td>-0.111</td>
<td>0.091</td>
<td>(-0.291,0.068)</td>
<td>0.224</td>
</tr>
<tr>
<td>Other</td>
<td>-0.263</td>
<td>0.086</td>
<td>(-0.433,-0.093)</td>
<td>0.003</td>
</tr>
<tr>
<td>None</td>
<td>-0.195</td>
<td>0.098</td>
<td>(-0.387,-0.003)</td>
<td>0.047</td>
</tr>
<tr>
<td>Highest occupational status†</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Professional/managerial†</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intermediate occupation</td>
<td>-0.096</td>
<td>0.050</td>
<td>(-0.195,0.002)</td>
<td>0.055</td>
</tr>
<tr>
<td>Routine and manual occupation</td>
<td>-0.134</td>
<td>0.050</td>
<td>(-0.233,-0.036)</td>
<td>0.008</td>
</tr>
<tr>
<td>Long-term unemployed</td>
<td>-0.304</td>
<td>0.097</td>
<td>(-0.495,-0.113)</td>
<td>0.002</td>
</tr>
<tr>
<td>Language spoken at home§</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>English only†</td>
<td>-0.013</td>
<td>0.065</td>
<td>(-0.142,0.115)</td>
<td>0.838</td>
</tr>
<tr>
<td>English and other language(s)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0†</td>
<td>0.001</td>
<td>0.050</td>
<td>(-0.097,0.099)</td>
<td>0.986</td>
</tr>
<tr>
<td>1†</td>
<td>-0.017</td>
<td>0.054</td>
<td>(-0.123,0.089)</td>
<td>0.755</td>
</tr>
<tr>
<td>Constant‡</td>
<td>0.864</td>
<td></td>
<td>0.846</td>
<td>0.403</td>
</tr>
</tbody>
</table>

*Interpreted as the additional utility over and above the reference group after adjustment for covariates.
†Reference group.
‡Includes child born at 22-26 weeks.
§Cases with missing information omitted from analyses.
CI, confidence interval; Sc, Scottish; SE, standard error.
consequently, selection biases are unlikely to represent a major problem. The analysis also used a contemporaneous classroom control group born at full term and matched for age, sex, and ethnic group, rather than control data from siblings, which are prone to biases due to continuously changing developmental profiles throughout childhood, or comparisons with British population norms for which limited data are available [31,48]. The study used validated and reliable approaches to measuring and valuing the costs and health-related quality of life preference-based outcomes associated with extremely preterm birth. The study cost accounting was comprehensive and conducted in line with the methodological requirements for modern health economic evaluation. Notably, in our opinion, the study detected statistically significant differences in each of the broad categories of cost and preference-based health-related quality of life outcomes between the extremely preterm and term born children. The annual cost difference of almost £2500 per annum between the extremely preterm children and their term counterparts exceeds that identified for several other childhood conditions [49]. Similarly, the mean decrement in the multi-attribute utility score of 0.167, or once adjusted for confounders of 0.280, far exceeds the 0.030 minimally important difference in utility score postulated in the literature as clinically important for evaluative purposes [50,51].

There are a number of caveats to the study findings, which should be borne in mind by readers. First, given that the control group comprised solely classmates attending mainstream schools, it might be argued that this is a healthier group than would be expected in the normal population and we have over-estimated cost and utility differences between the study groups. However, in the United Kingdom and Republic of Ireland, children with special education needs are largely integrated in mainstream schools. The prevalence of childhood disability requiring education in special schools is 1.1% in England [52]. Consequently, the inclusion of classmates for each extremely preterm child in separate schooling would most likely bias the comparison group. Moreover, separate standardized tests of academic attainment conducted in our study population revealed that classmates achieved mean scores expected for the normal population (reading score: mean 99, SD 12; math score: mean 99, SD 15) [53]. We are therefore confident that our control group adequately reflects the degree of health impairment in 11-year-old children. It is worth further noting that the extremely preterm children lost to follow-up at 11 years in our study were more likely to have had lower cognitive scores or cognitive impairment at 2.5 and 6 years, suggesting that, if anything, we might have underestimated the true extent of health impairment among these children and, by extension, the cost and utility differences between the study groups. A second caveat to the study findings is that the analysis of cost differences was conducted from a public sector perspective and encompassed costs to health, social, and education services. It is likely that birth at the borderline of viability has an economic impact upon other sectors of the economy and upon families and carers [22], suggesting that adopting a broader perspective would increase the cost differences between the study groups. A third caveat is that our cost estimates are based on parental reports of their child’s resource utilization over the previous year of life. Previous research has indicated that parents accurately recall their child’s hospital service utilization over extended periods when validated against medical records, but tend to under-report their child’s community service utilization [54]. If this were the case for our study, our absolute costs for community service utilization may be underestimated. A fourth caveat is that each child’s health-related quality of life was assessed by the main parent (usually the mother) rather than the child itself. At the outset of the study, it was considered on the basis of preliminary research that the comprehension level required for the Health Utilities Index Mark III is somewhat high for our pediatric sample where a number of children have developmental disabilities. Empirical evidence of the concordance between child and parent ratings of attributes of the child’s health-related quality of life suggests that parents are able to accurately rate observable behaviors, such as physical functioning and physical symptoms, but are less successful at identifying social or emotional impairments [55,56]. However, there is no consistent evidence to suggest that parents consistently either under-report or over-report social or emotional impairments [57], which suggests that there are unlikely to be systematic biases in the measurement of health-related quality of life in our study. Furthermore, our findings are broadly in keeping with the responses of adolescents aged 12–16 years in an international comparison study of extremely low birth weight adolescents in Canada, Germany, and The Netherlands [28]. A final caveat is that although the Health Utilities Index is the most widely used of the multi-attribute utility measures within the childhood context, the underlying preference weights for the Mark III health status classification system have been derived from a survey of Canadian adults. Recent research suggests that our approach of indirectly estimating health utilities by attaching population-derived utility scores to Health Utilities Index Mark III health states may be a poor substitute for directly measured utility scores at the individual level [58]. However, the cognitive requirements entailed in directly estimating utility scores for health states using techniques such as the visual analog, standard gamble, and time trade-off approaches precluded a direct measurement approach among our pediatric sample [55]. Moreover, many decision-making bodies, such as the National Institute for Health and Clinical Excellence in England and Wales, highlight the importance of valuing health outcomes using population-based preferences of the type we have used for the broader comparative purposes of economic evaluation [59].

How might the results of our study be used to inform economic evaluations of preventive and treatment interventions for extremely preterm birth? It is our view that cost-effectiveness assessments of new interventions in this area should ideally be based on evidence from randomized controlled trials with prospective assessments of costs using a validated client service receipt inventory and outcomes using a multi-attribute utility measure underpinned by population-based preferences. This approach is attractive in terms of its internal validity, minimization of bias and low incremental cost given the large fixed costs incurred by the prospective collection of clinical data [34]. The selection of the appropriate cost and multi-attribute utility measures for such studies should be informed by national technology assessment guidelines and an understanding of their psychometric properties within the childhood context, including their practicality, reliability, and validity. There are many circumstances, however, when prospective cost and utility measurement of this type is either insufficient or unfeasible. For example, within the context of trial-based economic evaluations, the analyst is often faced with extrapolating long-term costs and outcomes beyond the time horizon of the trial. Within the context of decision analytic modeling-based economic evaluations, the analyst is often faced with estimating costs and health utilities for a large number of health conditions or states with limited resources or time. Under such circumstances, our mean results and their associated distributions can act as data inputs for models of cost-effectiveness of preventive or treatment interventions for extremely preterm birth. Indeed, it might be argued that even where estimates of costs and health utilities associated with
extremely preterm birth are already available or can be estimated prospectively. As part of an economic evaluation, our results could be viewed as an additional resource that should be pooled with the totality of the evidence base [46]. It should be noted, however, that analysts may face a particular methodological challenge when the time horizon for the cost-effectiveness model spans the entire period of childhood or further into adulthood. Under these circumstances, the impact of age on costs and health utilities should be estimated from data gathered in large-scale longitudinal studies as they become available. When such data are not available, techniques such as meta-regression of data across a number of studies should be considered as a means of disentangling age impacts [27].

In conclusion, the results of this study should be used to inform the development of future economic evaluations of interventions aimed at preventing extremely preterm birth or alleviating its effects. Further research is required that identifies, measures, and values the longer-term economic impacts of the condition in a valid and reliable manner.

We are indebted to the EPICure Study Group, which includes pediatricians in 276 maternity units in the United Kingdom and Republic of Ireland who identified the original cohort, contributed perinatal data, and whose help was invaluable. We would also like to thank the children who participated in EPICure Study and the parents who completed the relevant research instruments.

Source of financial support: The 11-year EPICure follow-up study was supported by the UK Medical Research Council. Dr. Petrou is supported by a UK Medical Research Council Senior Non-Clinical Research Fellowship. Dr. Johnson is funded by a fellowship from the James Tudor Foundation. The National Perinatal Epidemiology Unit, University of Oxford, is funded by the Department of Health in England. The Health Economics Research Centre, University of Oxford, is funded by the National Co-ordinating Centre for Research Capacity Development in England. The views contained in this paper are those of the authors and, not necessarily, of the funding bodies.

References


Comparative Analysis of Length of Stay, Total Costs, and Treatment Success between Intravenous Moxifloxacin 400 mg and Levofloxacin 750 mg among Hospitalized Patients with Community-Acquired Pneumonia

Howard Friedman, PhD,1 Xue Song, PhD,2 Simone Crespi, MPH,3 Prakash Navaratnam, MPH, PhD4

1Analytic Solutions, LLC, New York, NY, USA; 2Thomson Reuters Healthcare, Cambridge, MA, USA; 3Schering-Plough Pharmaceuticals, Kenilworth, NJ, USA; 4Informagenics, LLC, Worthington, OH, USA

ABSTRACT

Objective: This study aimed to evaluate the length of stay (LOS), costs, and treatment consistency among patients hospitalized with community-acquired pneumonia (CAP) initially treated with intravenous (IV) moxifloxacin 400 mg or IV levofloxacin 750 mg.

Methods: Adults with CAP receiving IV moxifloxacin or IV levofloxacin for ≥3 days were identified in the Premier Perspective comparative database. Primary outcomes were LOS and costs. Secondary outcomes included treatment consistency, which was defined as 1) no additional IV moxifloxacin or levofloxacin after ≥1 day off study drug; 2) no switch to another IV antibiotic; and 3) no addition of another IV antibiotic.

Results: A total of 7720 patients met inclusion criteria (6040 receiving moxifloxacin; 1680 receiving levofloxacin). Propensity matching created two cohorts (1300 patients each) well matched for demographic, clinical, hospital, and payor characteristics. Before the patients were matched, mean LOS (5.87 vs. 5.46 days; P = 0.0004) and total costs per patient ($7302 vs. $6362; P < 0.0001) were significantly greater with moxifloxacin. After the patients were matched, mean LOS (5.63 vs. 5.51 days; P = 0.462) and total costs ($6524 vs. $6473; P = 0.476) were comparable in both cohorts. Treatment consistency was higher for moxifloxacin before (81.0% vs. 78.9%; P = 0.048) and after matching (82.8% vs. 78.0%; P = 0.002).

Conclusions: In-hospital treatment of CAP with IV moxifloxacin 400 mg or IV levofloxacin 750 mg was associated with similar hospital LOS and costs in propensity-matched cohorts.

Keywords: community-acquired pneumonia, cost, hospital, length of stay, levofloxacin, moxifloxacin, treatment outcomes.

Introduction

Community-acquired pneumonia (CAP) occurs in an estimated 5 to 6 million persons annually in the United States and results in approximately 60,000 deaths [1,2]. Each year, CAP is responsible for an estimated 10 million physician visits and more than 1 million hospitalizations [3,4]. A cost-of-illness study found that the total direct cost for treating CAP was $8.4 billion (in 1995 dollars), of which $4.8 billion was for patients ≥65 years of age [1]. Eighty-nine percent of the total cost, or $7.5 billion, was for inpatient care. According to the 2005 Nationwide Inpatient Sample, the average hospital length of stay (LOS) for CAP was 5.52 days, and in-hospital mortality was 4% [5].

The Infectious Diseases Society of America (IDSA) and the American Thoracic Society (ATS) recently issued consensus guidelines for the management of adults with CAP to identify patients who should be hospitalized, as well as antibiotics for empiric use before a causative pathogen has been isolated [6]. Several studies have shown that the implementation of the IDSA/ATS guidelines leads to improved patient care with concomitant reductions in hospital LOS, costs, and readmissions [7–9]. Other studies have demonstrated that such reductions in LOS produce substantial cost savings without adversely affecting mortality, readmission rates, or the time needed to return to normal activities [10,11].

The decision to hospitalized a patient is based on the severity of illness and the clinicians’ determination of a range of factors, including the likelihood that the patient will reliably take oral medications [6]. The IDSA/ATS guidelines recommend hospitalization of, where available and appropriate, intensive, in-home health-care services, for patients with confusion, urea, respiratory rate, blood pressure, and age ≥65 years scores ≥2. Empiric antibiotic therapy in hospitalized patients should consist of a respiratory fluoroquinolone (e.g., moxifloxacin or levofloxacin) or alternatively, a beta-lactam (e.g., cefotaxime, ceftriaxone, ampicillin, or for selected patients, etanopenem) plus macrolide regimen [6]. When patients are admitted directly to an intensive care unit (ICU), empiric therapy should consist of a beta-lactam plus either a respiratory fluoroquinolone or azithromycin.

The safety and efficacy of respiratory fluoroquinolones in hospitalized patients with CAP have been demonstrated in numerous studies [12–15]. Comparisons to beta-lactam–macrolide regimens or nonstandardized regimens suggest that fluoroquinolones lead to earlier hospital discharge, which in some studies has led to cost savings [16–18]. In the Community-Acquired Pneumonia Recovery in the Elderly study, a prospective, randomized, double-blind trial, treatment with moxifloxacin 400 mg daily was associated with significantly faster clinical recovery than treatment with levofloxacin 500 mg daily in hospitalized elderly patients with CAP, although the clinical cure rates did not differ significantly when assessed 5 to 21 days after completion of treatment [15]. Nevertheless, a recent retrospective database analysis of hospitalized patients with CAP suggested that initial treatment with intravenous (IV) levofloxacin 750 mg reduced hospital LOS by 0.5 day when compared with initial treatment with IV moxifloxacin 400 mg [19]. Comparisons between levofloxacin and moxifloxacin in that study may have been limited by methodological issues. To
address these issues, we also conducted a retrospective database analysis to evaluate LOS and costs, as well as treatment consistency, among patients with CAP treated with IV moxifloxacin 400 mg or IV levofloxacin 750 mg daily. Our objective was to compare treatment costs and outcomes (LOS and treatment consistency) with moxifloxacin and levofloxacin from the payor’s perspective, in propensity-matched cohorts of hospitalized patients with CAP.

Methods

Data Source

Data from the Premier Perspective comparative database (PCD; Charlotte, NC) from April 2003 to March 2006 were analyzed. The PCD contains inpatient data from more than 500 acute-care facilities in the United States that represent all geographic areas, urban and rural facilities, teaching and nonteaching hospitals, and a broad range of hospital sizes [20]. The database includes standard hospital admission and discharge information as well as date-stamped logs of all billed items for procedures, medications, and laboratory, diagnostic, and therapeutic services at the individual patient level. Hospitals submit data to the PCD on a monthly or quarterly basis. The data undergo extensive quality assurance and data validation checks, and the cost information is reconciled with the hospitals’ financial statements before the data are made available for research.

Eligibility Criteria

Patients ≥18 years of age with a principal diagnosis of CAP [International Classification of Diseases, Ninth Revision (ICD-9) codes: 481, 482.xx, 483.xx, 485, 486, and 487.x] who were treated for ≥3 days with either IV moxifloxacin 400 mg or IV levofloxacin 750 mg beginning on the date of hospital admission or on the following day were identified [21,22]. Patients who were admitted from or discharged to another acute-care hospital, nursing home, or other long-term care facility and those discharged from the hospital in the previous month were excluded to ensure that the pneumonia episode was community acquired and not nosocomial and that only complete episodes of inpatient care were examined. Patients were also excluded if they were discharged with surgical diagnosis-related groups (DRGs) (i.e., only patients with DRGs 79, 80, 89, 90, 475, and 565 were eligible), received a first IV dose with an antibiotic other than moxifloxacin or levofloxacin, switched antibiotic therapy during the first 3 days of hospitalization, had a hospital LOS of <3 days, or had a discharge status of death.

Patient Population

The patient population was characterized by a series of demographic and clinical variables, payor and provider variables, and medication-related variables. The demographic variables included patient age, sex, race, and year of admission. Clinical variables included the type of pneumonia based on the three-digit level principal discharge ICD-9 code, and comorbidities and CAP complications derived from the secondary diagnosis ICD-9 codes in the admission record. The CAP complications provide a measure of initial CAP severity and include sepsis, respiratory failure, pleural effusion and empyema, abscess, renal failure, and congestive heart failure. In addition, the severity of illness and risk of mortality were assessed by using the All Patient Refined DRG (APR-DRG). Other severity-related clinical variables that were captured included intubation at any time during an admission, respiratory therapy within the first 24 hours after admission, and total length of stay in the ICU. Payor data were grouped into the following categories: Medicare, Medicaid, private insurance, uninsured, other, and unknown. Provider data were characterized by region (Northeast, Midwest, South, and West), location (urban or rural), teaching hospital status, number of beds, admission from the emergency department, and specialty of the admitting and attending physicians. Medication-related variables included the number of doses of study drug that were administered during the hospital stay and the average daily dose of the study drug.

Outcome Variables

The primary outcome variables were total costs (in US dollars) per hospital admission with CAP and LOS per admission with CAP. The LOS represented the total number of days in the hospital, from the day of admission to the day of discharge. Because this study covered a 3-year period, total costs for the index hospitalization were calculated according to the discharge month and then standardized into March 2006 dollars by using the corresponding Consumer Price Index Medical Care for that month. The costs for the components of care—a secondary outcome variable—were identified by using UB-92 revenue codes and standardized by using the Consumer Price Index Medical Care. These costs included room and board (UB-92 revenue codes 110–219), pharmacy (codes 25x and 63x), IV therapy supplies (code 26x), respiratory therapy (code 41x), and all other costs (identified by other revenue codes). Other secondary outcome variables included treatment consistency. Treatment consistency was achieved if the patients met all of the following three criteria: 1) they did not require an additional dose of IV moxifloxacin or levofloxacin during the same hospital stay after being off the study drug for at least 1 day (retreatment); 2) they did not switch to another IV antibiotic (switch); and 3) they did not require the addition of another IV antibiotic (add-on).

Statistical Analyses

Statistical analysis was performed by using Statistical Analysis Software (SAS; SAS Institute Inc., Cary, NC). The conditional logistic regression analysis was conducted by using SAS 9.1, and the rest of the analyses were conducted by using SAS 8.2. Baseline demographic and other information were presented because either counts (%) for categorical data or mean (SD) for continuous data. Descriptive profiles were calculated for all variables before and after propensity score matching. Categorical variables were evaluated by using chi-square tests, and continuous variables were analyzed by using nonparametric rank-sum tests.

Outcomes were compared in three different manners: pre-matched unadjusted comparison, postmatched unadjusted comparison, and postmatched comparison adjusted for factors thought to influence outcome. Because patients were not randomly allocated to study treatment, propensity score matching was used to develop comparable cohorts of patients treated with IV moxifloxacin and IV levofloxacin having similar distributions of patient characteristics [23]. The probability that a patient received moxifloxacin as the index drug was modeled by using demographics, hospital characteristics, and baseline clinical characteristics. Variables that affect treatment choice and outcomes were included in the matching process, including age, sex, race, type of pneumonia, severity measures, comorbidities, CAP complications, type of payor, hospital teaching status, hospital size, region, location, admitting and attending physician specialty, emergency department admission, and year of admission. The logistic regression model was constructed in a stepwise manner to predict the probability of moxifloxacin use by each patient.
After the logistic regression model was estimated, individuals in the moxifloxacin cohort were matched one to one with the pool of levofloxacin users who had similar propensity scores using a greedy match [24]. Sampling without replacement was used when creating the propensity-matched samples. The quality of the matched was examined by using descriptive statistics tests, including chi-square and rank-sum tests. A multivariate analysis was performed on the matched samples to examine the marginal effects of specific factors on outcomes of interest. Generalized linear models were used when costs and cost components were the outcome (Gamma distribution), count data models (Poisson regression) were used for LOS, and a conditional logistic model stratified on the match was used for treatment consistency. Independent variables included in these models were demographic and clinical characteristics, payor, hospital characteristics, physician specialty, emergency department admission, year of admission, and treatment with moxifloxacin or levofloxacin. For the multivariate analysis related to LOS and the GLM related to total costs, P-values and Wald 95% confidence intervals (CI) are presented from the models without any adjustments.

Results

Patient Cohorts

A total of 34,287 patients with CAP who were discharged after receiving either moxifloxacin or levofloxacin during the 3-year study period were identified in the PCD. Of these, 6040 (25.4%) of 23,746 patients who received moxifloxacin and 1680 (15.9%) of 10,541 patients who received levofloxacin met eligibility criteria and were included in this analysis.

The unmatched moxifloxacin and levofloxacin cohorts differed significantly in terms of a variety of demographic, clinical, and hospital characteristics. Patients who received moxifloxacin tended to be older [mean (SD) age, 70.5 (15.2) vs. 68.4 (15.7) years; P < 0.0001], were more likely to have comorbid cardiovascular disease, and were less likely to have chronic obstructive pulmonary disorder or asthma (all P < 0.0001) (Table 1). The severity of illness estimated by APR-DRG severity and risk of mortality were generally higher in the moxifloxacin group (both P < 0.0001), whereas CAP complications were generally evenly balanced between groups, apart from congestive heart failure, which was more common in the moxifloxacin group (45.7% vs. 40.4%; P = 0.0001), and sepsis, which was more common in the levofloxacin group (2.9% vs. 5.7%; P < 0.0001). In terms of hospital characteristics, patients in the moxifloxacin group were more likely to be treated at a teaching hospital (49.1% vs. 39.7%; P < 0.0001), and to be admitted from the emergency department (84.7% vs. 76.3%; P < 0.0001) (Table 1). The distribution by hospital region, year of admission, and admitting physician specialty also differed significantly between groups (all P < 0.0001).

Propensity matching produced a total sample of 2600 patients, equally divided between the moxifloxacin and levofloxacin cohorts. After matching, there were no statistically significant differences between the two cohorts in terms of demographic, clinical, hospital, or payor characteristics (Tables 1 and 2). For the two combined cohorts, the mean age was 69 years; the majority were female (58%) and/or had comorbid asthma (76%) or chronic obstructive pulmonary disorder (63%). Most patients had at least one CAP complication (85%), most commonly respiratory failure (66%), and slightly more than half (55%) had moderate APR-DRG severity. Less than 1% of

### Table 1  Demographic and clinical characteristics of the moxifloxacin and levofloxacin cohorts before and after matching

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Moxifloxacin Before matching</th>
<th>Levofloxacin Before matching</th>
<th>P-value</th>
<th>Moxifloxacin After matching</th>
<th>Levofloxacin After matching</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Characteristic</td>
<td>(N = 6040)</td>
<td>(N = 1680)</td>
<td></td>
<td>(N = 1300)</td>
<td>(N = 1300)</td>
<td></td>
</tr>
<tr>
<td>Sex, no. (%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>2442 (40.4)</td>
<td>686 (40.8)</td>
<td>0.766</td>
<td>553 (42.5)</td>
<td>536 (41.2)</td>
<td>0.499</td>
</tr>
<tr>
<td>Female</td>
<td>3598 (59.6)</td>
<td>994 (59.2)</td>
<td></td>
<td>747 (57.5)</td>
<td>764 (58.8)</td>
<td></td>
</tr>
<tr>
<td>Age, mean (SD), year</td>
<td>70.5 (15.2)</td>
<td>68.4 (15.7)</td>
<td>&lt;0.0001</td>
<td>69.2 (15.2)</td>
<td>69.1 (15.6)</td>
<td>0.986</td>
</tr>
<tr>
<td>Comorbid conditions, no. (%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cancer</td>
<td>417 (6.9)</td>
<td>127 (7.6)</td>
<td>0.353</td>
<td>100 (7.7)</td>
<td>92 (7.1)</td>
<td>0.549</td>
</tr>
<tr>
<td>Diabetes</td>
<td>1941 (32.1)</td>
<td>510 (30.4)</td>
<td>0.166</td>
<td>399 (30.7)</td>
<td>403 (31.0)</td>
<td>0.865</td>
</tr>
<tr>
<td>COPD</td>
<td>3239 (53.6)</td>
<td>1061 (63.2)</td>
<td>&lt;0.0001</td>
<td>812 (62.5)</td>
<td>814 (62.6)</td>
<td>0.935</td>
</tr>
<tr>
<td>Asthma</td>
<td>3769 (62.4)</td>
<td>1320 (78.6)</td>
<td>&lt;0.0001</td>
<td>991 (76.2)</td>
<td>989 (76.1)</td>
<td>0.927</td>
</tr>
<tr>
<td>Cardiovascular disease</td>
<td>1898 (31.4)</td>
<td>431 (25.7)</td>
<td>&lt;0.0001</td>
<td>353 (27.2)</td>
<td>351 (27.0)</td>
<td>0.930</td>
</tr>
<tr>
<td>Secondary diagnoses, mean (SD), no.</td>
<td>7.63 (3.76)</td>
<td>7.38 (3.57)</td>
<td>0.072</td>
<td>7.72 (3.96)</td>
<td>7.44 (3.44)</td>
<td>0.369</td>
</tr>
<tr>
<td>Any CAP complication, no. (%)</td>
<td>5187 (85.9)</td>
<td>1406 (83.7)</td>
<td>0.025</td>
<td>1105 (85.0)</td>
<td>1099 (84.5)</td>
<td>0.743</td>
</tr>
<tr>
<td>Sepsis</td>
<td>177 (2.9)</td>
<td>96 (5.7)</td>
<td>&lt;0.0001</td>
<td>59 (4.5)</td>
<td>62 (4.8)</td>
<td>0.780</td>
</tr>
<tr>
<td>Respiratory failure</td>
<td>3903 (64.6)</td>
<td>1088 (64.8)</td>
<td>0.914</td>
<td>874 (67.2)</td>
<td>852 (65.5)</td>
<td>0.361</td>
</tr>
<tr>
<td>Pleural effusion/empyema</td>
<td>319 (5.3)</td>
<td>95 (5.7)</td>
<td>0.548</td>
<td>65 (5.0)</td>
<td>79 (6.1)</td>
<td>0.230</td>
</tr>
<tr>
<td>Abscess</td>
<td>16 (0.3)</td>
<td>8 (0.5)</td>
<td>0.169</td>
<td>4 (0.3)</td>
<td>5 (0.4)</td>
<td>0.738</td>
</tr>
<tr>
<td>Renal failure</td>
<td>1548 (25.6)</td>
<td>427 (25.4)</td>
<td>0.860</td>
<td>361 (27.8)</td>
<td>364 (28.0)</td>
<td>0.896</td>
</tr>
<tr>
<td>Congestive heart failure</td>
<td>2758 (45.7)</td>
<td>678 (40.4)</td>
<td>0.0001</td>
<td>535 (41.2)</td>
<td>532 (40.9)</td>
<td>0.905</td>
</tr>
<tr>
<td>APR-DRG severity, no. (%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Minor</td>
<td>485 (8.0)</td>
<td>171 (10.2)</td>
<td>&lt;0.0001</td>
<td>106 (8.2)</td>
<td>124 (9.5)</td>
<td>0.270</td>
</tr>
<tr>
<td>Moderate</td>
<td>3154 (52.2)</td>
<td>934 (55.6)</td>
<td>0.0001</td>
<td>701 (53.9)</td>
<td>726 (55.8)</td>
<td>0.977</td>
</tr>
<tr>
<td>Major</td>
<td>2170 (35.9)</td>
<td>532 (31.7)</td>
<td></td>
<td>454 (34.9)</td>
<td>417 (32.1)</td>
<td></td>
</tr>
<tr>
<td>Extreme</td>
<td>231 (3.8)</td>
<td>43 (2.6)</td>
<td></td>
<td>39 (3.0)</td>
<td>33 (2.5)</td>
<td></td>
</tr>
<tr>
<td>APR-DRG mortality risk, no. (%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Minor</td>
<td>1784 (29.5)</td>
<td>620 (36.9)</td>
<td>&lt;0.0001</td>
<td>452 (34.8)</td>
<td>455 (35.0)</td>
<td>0.977</td>
</tr>
<tr>
<td>Moderate</td>
<td>3312 (54.8)</td>
<td>854 (50.8)</td>
<td></td>
<td>678 (52.2)</td>
<td>678 (52.2)</td>
<td></td>
</tr>
<tr>
<td>Major</td>
<td>822 (13.6)</td>
<td>179 (10.7)</td>
<td></td>
<td>153 (11.8)</td>
<td>148 (11.4)</td>
<td></td>
</tr>
<tr>
<td>Extreme</td>
<td>122 (2.0)</td>
<td>27 (1.6)</td>
<td></td>
<td>17 (1.3)</td>
<td>19 (1.5)</td>
<td></td>
</tr>
<tr>
<td>Patients intubated, no. (%)</td>
<td>59 (1.0)</td>
<td>8 (0.5)</td>
<td>0.050</td>
<td>6 (0.5)</td>
<td>8 (0.6)</td>
<td>0.592</td>
</tr>
<tr>
<td>Patients receiving respiratory therapy, no. (%)</td>
<td>5130 (84.9)</td>
<td>1528 (91.0)</td>
<td>&lt;0.0001</td>
<td>1166 (89.7)</td>
<td>1168 (89.8)</td>
<td>0.897</td>
</tr>
</tbody>
</table>

APR-DRG, All Patient Refined Diagnostic-Related Groups; CAP, community-acquired pneumonia; COPD, chronic obstructive pulmonary disorder. P-values are the result of bivariate comparisons.
patients were intubated. Most patients were treated at an urban hospital (85%), had Medicare coverage (69%), were admitted in the year 2005 (52%), and/or were admitted after presentation to the emergency department (79%).

Outcomes

Before propensity matching, the mean (SD) LOS was significantly longer (0.41 day) in the moxifloxacin than in the levofloxacin cohort [5.87 (4.10) vs. 5.46 (3.45) days, respectively; \( P = 0.0004 \) (Fig. 1). Nevertheless, after propensity score matching, there was no significant difference in LOS between the moxifloxacin and levofloxacin groups [5.63 (3.50) vs. 5.51 (3.52) days; \( P = 0.462 \)]. Total hospital costs showed a similar profile, with moxifloxacin versus levofloxacin having higher average total costs in the unmatched population [$7302 ($10,754) vs. $6362 ($4654), respectively; \( P < 0.0001 \)], but no significant differences in the propensity-matched cohorts [$6624 ($5576) vs. $6473 ($4782); \( P = 0.476 \)] (Fig. 2).

Similarly, when individual cost components were evaluated, higher room and board charges (\( P < 0.0001 \)) and pharmacy costs (\( P = 0.0003 \)) were found in the moxifloxacin group than in the levofloxacin group before matching. After propensity matching, room and board charges (\( P = 0.239 \)) and pharmacy costs (\( P = 0.905 \)) did not differ significantly between the moxifloxacin and levofloxacin cohorts (Table 3). The numeric difference between the two matched cohorts was $103 for room and board costs and $33 for pharmacy costs. The cost of IV

Table 2 Payor and hospital characteristics of the moxifloxacin and levofloxacin cohorts before and after matching

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Before matching</th>
<th>After matching</th>
<th>( P )-value</th>
<th>Before matching</th>
<th>After matching</th>
<th>( P )-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Payor type, no. (%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Medicare</td>
<td>4152 (68.7)</td>
<td>1122 (66.8)</td>
<td>&lt;0.0001</td>
<td>894 (68.8)</td>
<td>896 (68.9)</td>
<td>0.979</td>
</tr>
<tr>
<td>Medicaid</td>
<td>299 (5.0)</td>
<td>114 (6.8)</td>
<td></td>
<td>76 (5.8)</td>
<td>81 (6.2)</td>
<td></td>
</tr>
<tr>
<td>Private insurance</td>
<td>1296 (21.5)</td>
<td>295 (17.6)</td>
<td></td>
<td>241 (18.5)</td>
<td>233 (17.9)</td>
<td></td>
</tr>
<tr>
<td>Uninsured</td>
<td>184 (3.0)</td>
<td>81 (4.8)</td>
<td></td>
<td>48 (3.7)</td>
<td>51 (3.9)</td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td>109 (1.8)</td>
<td>68 (4.0)</td>
<td></td>
<td>41 (3.2)</td>
<td>39 (3.0)</td>
<td></td>
</tr>
<tr>
<td>Hospital region, no. (%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Northeast</td>
<td>1572 (26.0)</td>
<td>196 (11.7)</td>
<td>&lt;0.0001</td>
<td>188 (14.5)</td>
<td>196 (15.1)</td>
<td>0.585</td>
</tr>
<tr>
<td>Midwest</td>
<td>1301 (21.5)</td>
<td>415 (24.7)</td>
<td></td>
<td>332 (25.5)</td>
<td>343 (26.4)</td>
<td></td>
</tr>
<tr>
<td>South</td>
<td>2817 (46.6)</td>
<td>868 (51.7)</td>
<td></td>
<td>655 (50.4)</td>
<td>622 (47.8)</td>
<td></td>
</tr>
<tr>
<td>West</td>
<td>350 (5.8)</td>
<td>201 (12.0)</td>
<td></td>
<td>125 (9.6)</td>
<td>139 (10.7)</td>
<td></td>
</tr>
<tr>
<td>Population density, no. (%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rural</td>
<td>399 (6.6)</td>
<td>371 (22.1)</td>
<td>&lt;0.0001</td>
<td>200 (15.4)</td>
<td>187 (14.4)</td>
<td>0.474</td>
</tr>
<tr>
<td>Urban</td>
<td>5641 (93.4)</td>
<td>1309 (77.9)</td>
<td></td>
<td>1100 (84.6)</td>
<td>1113 (85.6)</td>
<td></td>
</tr>
<tr>
<td>Teaching hospital, no. (%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rural</td>
<td>399 (6.6)</td>
<td>371 (22.1)</td>
<td>&lt;0.0001</td>
<td>200 (15.4)</td>
<td>187 (14.4)</td>
<td>0.474</td>
</tr>
<tr>
<td>Urban</td>
<td>5641 (93.4)</td>
<td>1309 (77.9)</td>
<td></td>
<td>1100 (84.6)</td>
<td>1113 (85.6)</td>
<td></td>
</tr>
<tr>
<td>Hospital size, mean (SD) beds, no.</td>
<td>465.1 (226.4)</td>
<td>378.6 (213.9)</td>
<td>&lt;0.0001</td>
<td>389.0 (260.3)</td>
<td>390.5 (208.9)</td>
<td>0.637</td>
</tr>
<tr>
<td>Year of admission, no. (%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2003</td>
<td>1119 (18.5)</td>
<td>16 (1.0)</td>
<td>&lt;0.0001</td>
<td>19 (1.5)</td>
<td>16 (1.2)</td>
<td>0.559</td>
</tr>
<tr>
<td>2004</td>
<td>1988 (32.9)</td>
<td>201 (12.0)</td>
<td></td>
<td>217 (16.7)</td>
<td>193 (14.8)</td>
<td></td>
</tr>
<tr>
<td>2005</td>
<td>2267 (37.5)</td>
<td>861 (51.3)</td>
<td></td>
<td>681 (52.4)</td>
<td>704 (54.2)</td>
<td></td>
</tr>
<tr>
<td>2006</td>
<td>666 (11.0)</td>
<td>602 (35.8)</td>
<td></td>
<td>383 (29.5)</td>
<td>387 (29.8)</td>
<td></td>
</tr>
<tr>
<td>Admission from emergency department, no. (%)</td>
<td>5117 (84.7)</td>
<td>1282 (76.3)</td>
<td>&lt;0.0001</td>
<td>1036 (79.7)</td>
<td>1024 (78.8)</td>
<td>0.562</td>
</tr>
<tr>
<td>Admitting physician specialty, no. (%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hospitalist</td>
<td>29 (0.5)</td>
<td>122 (7.3)</td>
<td>&lt;0.0001</td>
<td>29 (2.2)</td>
<td>14 (1.1)</td>
<td>0.207</td>
</tr>
<tr>
<td>Infectious disease</td>
<td>10 (0.2)</td>
<td>1 (0.1)</td>
<td></td>
<td>2 (0.2)</td>
<td>1 (0.1)</td>
<td></td>
</tr>
<tr>
<td>Primary care</td>
<td>4360 (72.2)</td>
<td>1238 (73.7)</td>
<td></td>
<td>1012 (77.8)</td>
<td>1015 (78.1)</td>
<td></td>
</tr>
<tr>
<td>Pulmonologist</td>
<td>299 (5.0)</td>
<td>114 (6.8)</td>
<td></td>
<td>89 (6.8)</td>
<td>93 (7.2)</td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td>1342 (22.2)</td>
<td>205 (12.2)</td>
<td></td>
<td>168 (12.9)</td>
<td>177 (13.6)</td>
<td></td>
</tr>
</tbody>
</table>

\( P \)-values are the result of bivariate comparisons.
therapy supplies was higher in the levofloxacin group than in the moxifloxacin group both before \((P < 0.0001)\) and after \((P = 0.0006)\) propensity matching; the differences between treatment cohorts were $10 and $25, respectively. Respiratory therapy costs and other costs did not differ between groups in the prematched cohorts, whereas after matching, respiratory therapy costs were $112 higher in the moxifloxacin group \((P < 0.0001)\), and other costs were $71 higher in the levofloxacin group \((P = 0.0043)\).

Moxifloxacin was associated with a significantly higher treatment consistency than levofloxacin before propensity matching \((81.0\% \text{ vs. } 78.9\%, \text{ respectively; } P = 0.0481)\) as well as after matching \((82.8\% \text{ vs. } 78.0\%; P = 0.0018)\) (Fig. 3). These findings reflected differences in retreatment rates (defined as requiring an additional dose of IV moxifloxacin or levofloxacin during the same hospital stay after being off the study drug for at least 1 day), which were higher in the levofloxacin group before propensity matching \((13.6\% \text{ vs. } 11.9\%, \text{ respectively; } P = 0.060)\) and significantly higher in the levofloxacin group after matching \((14.1\% \text{ vs. } 9.2\%; P < 0.0001)\). Frequencies of regimen changes from one study drug to another or to add-on therapy did not differ between the moxifloxacin and levofloxacin cohorts before or after propensity matching.

Factors Influencing Outcomes

Tables 4 and 5 present factors that significantly affected LOS (Table 4) and total costs (Table 5). In the multivariate analysis, the choice of index drug (moxifloxacin vs. levofloxacin) was not significant in the estimation of the LOS (Table 4; 95% CI for index drug regression coefficient = -0.027 to 0.039) or the total costs (Table 5; 9% CI for index drug regression coefficient = -0.048 to 0.025). On the other hand, a number of other demographic, clinical, and hospital factors were significant predictors of one or both of these outcomes, resulting in models with some predictive power; adjusted \(R^2 = 0.33\) and adjusted \(R^2 = 0.24\) for the multivariate model for total costs and LOS, respectively.

Clinical factors that were significantly associated with LOS included the need for intubation, the presence of an abscess, an ICD-9 code for other bacterial pneumonia, the presence of pleural effusion or empyema, the presence of congestive heart failure, and the number of secondary diagnoses at admission \((all \ P < 0.0001)\) as well as APR-DRG severity \((P = 0.0001)\). Female sex \((P < 0.0001)\) and advanced age \((P = 0.001)\) were the only demographic factors that emerged as significant predictors of LOS, whereas Medicare insurance \((P = 0.0029)\) and other forms of insurance (i.e., not Medicare, Medicaid, or private insurance) \((P = 0.0009)\) were significantly associated with shorter mean LOS. Hospital factors significantly associated with longer LOS included nonteaching facility, Northeast region, and admission in 2004 or 2005.

In general, the foregoing factors were also significantly associated with total hospital costs (Table 5). The proportion of the LOS spent in the ICU strongly influenced total costs \((P < 0.0001)\), as did the need for intubation \((P < 0.0001)\). Several hospital characteristics, including nonteaching status, Northeast region, rural location, admission in 2004, and admission from 2005.

<table>
<thead>
<tr>
<th>Table 3</th>
<th>Medical costs and length of stay of moxifloxacin and levofloxacin cohorts before and after matching</th>
</tr>
</thead>
<tbody>
<tr>
<td>Outcome</td>
<td>Before matching</td>
</tr>
<tr>
<td></td>
<td>Moxifloxacin ((N = 6040))</td>
</tr>
<tr>
<td>Mean (SD) cost per patient, US$</td>
<td></td>
</tr>
<tr>
<td>Total cost</td>
<td>7,302 (10,754)</td>
</tr>
<tr>
<td>Room and board</td>
<td>4,108 (8,709)</td>
</tr>
<tr>
<td>Pharmacy</td>
<td>624 (951)</td>
</tr>
<tr>
<td>IV therapy supplies</td>
<td>194 (251)</td>
</tr>
<tr>
<td>Respiratory therapy</td>
<td>467 (792)</td>
</tr>
<tr>
<td>Other</td>
<td>1,099 (3,922)</td>
</tr>
<tr>
<td>Mean (SD) length of stay, d</td>
<td>5.87 (4.10)</td>
</tr>
</tbody>
</table>

IV, intravenous. P-values are the result of bivariate comparisons.
the emergency department, were significantly predictive of total costs (Table 5), but payor factors were not (data not shown).

Patients treated with moxifloxacin were more likely to achieve treatment consistency than those receiving levofloxacin [odds ratio (OR) = 1.40; P = 0.0048, global null hypothesis likelihood ratio (OR) = 0.914; P = 0.001] or a higher APR-DRG risk of mortality (OR = 0.638; P = 0.015). No other demographic, clinical, hospital, or payor characteristic was a significant predictor of treatment consistency.

**Discussion**

Results of the present study indicate that neither clinical nor formulary decisions concerning levofloxacin or moxifloxacin for CAP can be made strictly on the basis of different costs of care, including LOS. This retrospective database analysis demonstrated that daily IV treatment with moxifloxacin 400 mg or levofloxacin 750 mg was associated with similar hospital LOS and total costs in a matched cohort of patients with CAP. Before propensity matching, the moxifloxacin and levofloxacin cohorts differed considerably in demographic and clinical characteristics known to influence LOS and costs, such as advanced age, illness severity, and mortality risk, as well as in various hospital- and payor-based characteristics that can also impact these outcomes. Because patients were not randomly allocated to moxifloxacin or levofloxacin treatment, estimation of treatment effects on LOS and costs may be biased by such imbalances between treatment groups. Accordingly, comparisons of outcomes between cohorts receiving one of the two fluoroquinolones before successful propensity matching are not a reliable means of concluding that LOS or total costs differ between moxifloxacin and levofloxacin.

Propensity score matching was developed to reduce bias between two imbalanced study groups. Heckman and colleagues

**Table 4 Results of multivariate analysis of variables associated with length of stay (LOS)**

<table>
<thead>
<tr>
<th>Variable</th>
<th>Estimated coefficients</th>
<th>Wald 95% CI</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Moxifloxacin</td>
<td>0.0057</td>
<td>-0.027 to 0.039</td>
<td>0.734</td>
</tr>
<tr>
<td>Demographic factors</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>0.0027</td>
<td>0.001–0.004</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Female</td>
<td>0.0707</td>
<td>0.036–0.105</td>
<td></td>
</tr>
<tr>
<td>Clinical factors</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ICD-9 code 482 (other bacterial pneumonia)</td>
<td>0.2588</td>
<td>0.198–0.320</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Comorbid cancer</td>
<td>0.1026</td>
<td>0.040–0.166</td>
<td>0.0014</td>
</tr>
<tr>
<td>Comorbid cardiovascular disease</td>
<td>-0.0552</td>
<td>-0.16 to -0.094</td>
<td>0.0057</td>
</tr>
<tr>
<td>Number of secondary diagnoses</td>
<td>0.0324</td>
<td>0.027–0.038</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Respiratory failure</td>
<td>0.0679</td>
<td>0.017–0.118</td>
<td>0.0085</td>
</tr>
<tr>
<td>Pleural effusion and empyema</td>
<td>0.2117</td>
<td>0.146–0.278</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Abscess</td>
<td>0.5152</td>
<td>0.326–0.705</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Congestive heart failure</td>
<td>0.0845</td>
<td>0.045–0.124</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>APR-DRG severity</td>
<td>0.0721</td>
<td>0.037–0.107</td>
<td>0.0001</td>
</tr>
<tr>
<td>Intubation</td>
<td>0.6014</td>
<td>0.432–0.770</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Hospital factors</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nonteaching hospital</td>
<td>0.0583</td>
<td>0.014–0.102</td>
<td>0.0095</td>
</tr>
<tr>
<td>Northeast region</td>
<td>0.1537</td>
<td>0.077</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Admission in 2004</td>
<td>0.0852</td>
<td>0.032–0.138</td>
<td>0.0016</td>
</tr>
<tr>
<td>Admission in 2005</td>
<td>0.0534</td>
<td>0.015–0.092</td>
<td>0.0068</td>
</tr>
<tr>
<td>Payor factors</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Medicare</td>
<td>-0.1456</td>
<td>-0.050 to -0.242</td>
<td>0.0029</td>
</tr>
<tr>
<td>Other</td>
<td>-0.2277</td>
<td>-0.093 to -0.362</td>
<td>0.0009</td>
</tr>
</tbody>
</table>

Overall model: Adjusted $R^2 = 0.24$, chi-square test for model P < 0.0001.

APR-DRG, All Patient Refined Diagnostic-Related Groups; CI, confidence interval; ICD-9, International Classification of Diseases, Ninth Revision.
suggested that up to 85% of the bias resulting from unequal distributions in patient characteristics can be neutralized by matching patients by using propensity scores [23]. Multiple methods have been developed for conducting propensity matching, including stratified matching, nearest-neighbor matching, radius matching, kernel matching, and Mahalanobis matching [23–26]. When there is considerable overlap in the estimated propensity score between groups, as is the case in the present study, each matching method should provide similar estimated treatment effects [26]. After propensity score matching, the moxifloxacin and levofloxacin cohorts were well balanced, with no significant differences between groups in admission demographic, clinical, hospital, or payor characteristics. This supports the conclusion that hospital LOS and charges for inpatient CAP management do not differ significantly between the IV moxifloxacin and levofloxacin regimens studied. Furthermore, multivariate analyses of the total cohort showed that the choice of treatment (moxifloxacin or levofloxacin) was not a significant factor in predicting the hospital LOS or total charges.

The mean LOS of 5.51 to 5.63 days in the propensity-matched cohorts in this study is similar to mean values of 5.52 days reported in the 2005 Nationwide Inpatient Sample (NIS) and 5.27 days reported in the 2005 National Hospital Discharge Survey (NHDS) [5]. The NIS data were drawn from hospitals in 37 states that represented 78% of US community hospitals, whereas the NHDS data covered all hospitals across the 50 states. The NIS and NHDS data included all patients, whereas the present study evaluated only those ≥18 years of age.

Our findings differed somewhat from those of a similar database analysis conducted by Schein and colleagues, who retrospectively evaluated the PCD database to compare moxifloxacin and levofloxacin treatment outcomes in patients with CAP from January 2004 to December 2005 [19]. Apart from the different study time period, Schein and colleagues included patients who were hospitalized for ≥3 days but ≤90 days and who received IV moxifloxacin or IV levofloxacin through the first 3 days of hospitalization. In comparison, the present study did not limit total LOS but consequently did not exclude patients with LOS of ≥90 days, and patients were eligible for the present analysis if they started IV moxifloxacin or IV levofloxacin treatment on the day of admission or the following day and continued the regimen for ≥3 days. In addition, Schein and colleagues did not exclude patients with surgical DRGs, which may be expected to confound LOS and costs. Patients in our study tended to be older (69 years vs. 64 years) and were more likely to be female (58% vs. 52%), reside in the Northeast (15% vs. 12%), not the South (49% vs. 58%), and were admitted from the emergency department (79% vs. 71%) than those in the Schein et al. analysis.

Mean LOS and total costs in the moxifloxacin and levofloxacin cohorts also differed across the two studies, both before and after propensity matching. For example, in the propensity-matched cohorts, Schein and colleagues found mean LOS to be 6.37 days with moxifloxacin and 5.83 days with levofloxacin (P = 0.02) [19], compared to 5.63 days and 5.51 days, respectively, in this study. Given the considerable overlap in time periods between the two studies, differences in management practices, such as greater use of short-course therapy, cannot explain the differences between studies.

Several methodological factors may have contributed to differences between the two studies. First, Schein and colleagues performed propensity matching on approximately 60 variables, and, after matching, only the number of patients who were admitted to urban hospitals differed between treatment groups [19]. Although this appears to be a good match, the mean propensity score still differed significantly between the moxifloxacin and levofloxacin matched cohorts (P < 0.001) possibly because an unusually high caliper score of 0.7 was used. The standard method is to use one quarter of the standard deviation of the estimated propensity score, so it is unclear why 0.7 was chosen in their analysis. Second, Schein and colleagues excluded all patients with LOS of ≥90 days, whereas no LOS limit was placed on patient inclusion in this analysis. Although the impact of this exclusion criterion is unknown, it does represent a factor that would influence the calculation of LOS and cost. Third, the multivariate regression analyses performed after propensity

### Table 5 Results of generalized linear model analysis of variables associated with total costs

<table>
<thead>
<tr>
<th>Variable</th>
<th>Estimated coefficients</th>
<th>Wald 95% CI</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Moxifloxacin</td>
<td>-0.0155</td>
<td>-0.048 to 0.025</td>
<td>0.532</td>
</tr>
<tr>
<td>Demographic factors</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>0.0611</td>
<td>0.023–0.099</td>
<td>0.0015</td>
</tr>
<tr>
<td>Clinical factors</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ICD-9 code 482 (other bacterial pneumonia)</td>
<td>0.233</td>
<td>0.156–0.310</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Comorbid cancer</td>
<td>0.1831</td>
<td>0.110–0.257</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Comorbid COPD</td>
<td>0.0163</td>
<td>0.059–0.154</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Number of secondary diagnoses</td>
<td>0.0349</td>
<td>0.029–0.041</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Pleural effusion and empyema</td>
<td>0.1942</td>
<td>0.110–0.278</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Abscess</td>
<td>0.5179</td>
<td>0.205–0.831</td>
<td>0.0012</td>
</tr>
<tr>
<td>Renal failure</td>
<td>0.0648</td>
<td>0.017–0.113</td>
<td>0.0083</td>
</tr>
<tr>
<td>Congestive heart failure</td>
<td>0.1571</td>
<td>0.113–0.201</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>APR-DRG severity</td>
<td>0.0759</td>
<td>0.037–0.115</td>
<td>0.0001</td>
</tr>
<tr>
<td>Intubation</td>
<td>0.5931</td>
<td>0.325–0.861</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Respiratory therapy</td>
<td>0.1189</td>
<td>0.049–0.189</td>
<td>0.0008</td>
</tr>
<tr>
<td>% of LOS spent in ICU</td>
<td>0.5987</td>
<td>0.396–0.802</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Hospital factors</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rural location</td>
<td>0.0657</td>
<td>0.017–0.114</td>
<td>0.0079</td>
</tr>
<tr>
<td>Northeast region</td>
<td>0.1246</td>
<td>0.040–0.210</td>
<td>0.0040</td>
</tr>
<tr>
<td>Rural location</td>
<td>-0.063</td>
<td>-0.094 to -0.164</td>
<td>0.0003</td>
</tr>
<tr>
<td>Admission in 2004</td>
<td>0.1521</td>
<td>0.094–0.211</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Emergency department admission</td>
<td>-0.0909</td>
<td>-0.045 to -0.137</td>
<td>0.0001</td>
</tr>
</tbody>
</table>

Overall model: Adjusted R² = 0.33, chi-square test for model P < 0.0001.

APR-DRG, All Patient Refined Diagnostic-Related Group; CI, confidence interval; COPD, chronic obstructive pulmonary disorder; ICD-9, International Classification of Diseases, Ninth Revision; ICU, intensive care unit; LOS, length of stay.
matching in the study by Schein and colleagues seemed to include only a limited number of variables (i.e., urban hospital location, treatment, and the interaction between urban location and treatment). Urban location was included as an independent variable because it remained significantly different between treatment cohorts after propensity matching. Notably, variables that are expected to affect LOS and costs and that were statistically significant in our models, such as comorbidities, severity, complications, hospital teaching status, and hospital admission via the emergency department, were omitted from the regression models used by Schein and colleagues.

Treatment consistency—defined in the present study as the absence of retreatment with the first study drug or switching to or adding another IV antibiotic—was evaluated as a secondary outcome in the present study. In the propensity-matched cohort, treatment consistency was achieved by 82.8% of patients who received moxifloxacin and 78.0% of those who received levofloxacin. Logistic regression analysis also demonstrated that treatment with moxifloxacin significantly increased the likelihood of treatment consistency when compared with levofloxacin.

Potential Study Limitations

The retrospective database design of this study has potential limitations. First, because patients were not randomly allocated to treatment, propensity score matching was needed to generate well-balanced cohorts for comparisons between moxifloxacin and levofloxacin, based on measured demographic and clinical characteristics of the patients as well as measured payor and hospital characteristics. Nevertheless, propensity score matching can be conducted only, based on observable characteristics (demographic, clinical, hospital, and payor characteristics) in the database [27]. Characteristics not captured in the database, such as physician preference, formulary restriction, causative pathogen, and antimicrobial resistance rates, could still be different between the two cohorts. In addition, as in any other retrospective administrative claims database analysis, patient-level data were somewhat limited. Although we feel that our propensity score matching protocol, including multivariate analysis post matching, effectively generated two very similar cohorts, our findings do not serve as a substitute for randomization of patients into the two treatment groups in terms of excluding certain forms of bias (e.g., selection bias, treatment-selection bias/confounding by indication). As another potential limitation, treatment consistency, as operationally defined for the first time in the present study as the absence of retreatment with the first study drug or switching to or adding another IV antibiotic, needs to be further evaluated and/or validated in distinct populations.

Conclusions

Inpatient management of CAP using IV moxifloxacin 400 mg or IV levofloxacin 750 mg daily was associated with similar hospital LOS and total costs in balanced patient populations in the present retrospective database analysis. Initial treatment with IV moxifloxacin significantly increased the likelihood of treatment consistency when compared with initial IV levofloxacin treatment using our study criteria. Both fluoroquinolones are recognized as appropriate options for empiric therapy of CAP in hospitalized patients. On the basis of the present findings, there were no significant differences between these fluoroquinolones in hospital LOS or overall costs for the management of CAP in the hospital setting.

Assistance in article preparation was provided by Stephen W. Gutkin, Rete Biomedical Communications Corp. (Wyckoff, NJ USA), with support from the study sponsor.

Source of Financial Support: Prakash Navaratnam is an employee of Informagemcs, LLC (Worthington, OH), and Howard Friedman of Analytical Solutions, LLC (New York, NY). Xue Song is an employee of Thomson Reuters Healthcare (Cambridge, MA). All three companies provide consultation with the pharmaceutical industry, and both Drs. Navaratnam and Friedman were remunerated by the study sponsor. Dr. Friedman is also a paid consultant to Bayer. Simone Crespi is an employee of the study sponsor. This study and its report were supported by Schering-Plough Pharmaceuticals, Kenilworth, NJ. The sponsor had a role in study design, data acquisition, and interpretation, drafting, and editing the article, and the decision to publish the findings.

References

14 Welte T, Petermann W, Schürrmann D, et al. Treatment with sequential intravenous or oral moxifloxacin was associated with faster clinical improvement than was standard therapy for hospitalized patients with community-acquired pneumonia who received initial parenteral therapy. Clin Infect Dis 2005;41:1697–705.


Evaluating Minimal Important Differences for the FACT-Melanoma Quality of Life Questionnaire

Robert L. Askew, BA,1 Yan Xing, MD, PhD, MS,1 J. Lynn Palmer, PhD,2 David Cella, PhD,3 Lemuel A. Moye, MD, PhD,4 Janice N. Cormier, MD, MPH1

1Department of Surgical Oncology, The University of Texas M.D. Anderson Cancer Center, Houston, TX, USA; 2Department of Palliative Care and Rehabilitation Medicine, The University of Texas M.D. Anderson Cancer Center, Houston, TX, USA; 3Center on Outcomes, Research, and Education, Evanston Northwestern Healthcare, Evanston, IL, USA; 4Department of Biostatistics, The University of Texas School of Public Health, Houston, TX, USA

ABSTRACT

Objectives: Minimal Important Differences (MIDs) establish benchmarks for interpreting mean differences in clinical trials involving quality of life outcomes and inform discussions of clinically meaningful change in patient status. The purpose of this study was to assess MIDs for the Functional Assessment of Cancer Therapy-Melanoma (FACT-M).

Methods: A prospective validation study of the FACT-M was performed with 273 patients with stages I through IV melanoma. FACT-M, Karnofsky Performance Scales, and Eastern Cooperative Oncology Group Performance Status scores were obtained at baseline and 3 months following enrollment. Anchor- and distribution-based methods for assessing MIDs were compared, and pattern-mixture modeling was employed to derive multivariate adjusted estimates.

Results: This study indicates that an approximate range for MIDs of the FACT-M subscales is between 3 to 9 points for the Trial Outcome Index, 4 to 6 points for the Melanoma Combined Subscale, 2 to 4 points for the Melanoma Subscale, and 1 to 2 points for the Melanoma Surgery Subscale. Each method produced similar but not identical ranges of MIDs.

Conclusions: The properties of the anchor instrument employed to derive MIDs directly affect resulting MID ranges and point values. When MIDs are offered as supportive evidence of a clinically meaningful change, the anchor instrument used to derive clinically meaningful thresholds of change should be clearly stated along with information supporting the choice of anchor instrument as the most appropriate for the domain of interest.

Keywords: Functional Assessment of Cancer Therapy, melanoma, minimal important differences, patient reported outcomes, quality of life.

Introduction

Quality of life (QOL) measures in cancer research have been shown to be independent predictors of both survival and response to therapy [1–5], and for melanoma in particular, QOL has been shown to be an independent predictor of survival for patients with advanced disease [2,6]. Patient reported outcomes such as QOL can serve multiple purposes, including acting as validation measures of treatment efficacy in the context of clinical trials and serving as reference points for clinical decision-making when modest differences in survival are anticipated among various treatment modalities [7–9].

With respect to measuring and reporting patient reported outcomes in the context of clinical trials, the US Food and Drug Administration published 2006 draft guidance on the methods to derive and interpret Minimal Important Differences (MIDs) for QOL instruments [7]. In this document, which was primarily intended to serve as a guidance for industry related to product/drug labeling, MIDs are defined as the minimum change observed in a patient reported outcome measure (e.g., QOL score) between treatment groups that can be correlated with or interpreted as a treatment benefit. With the increasingly acknowledged importance of patient reported outcomes in the context of clinical trials [7,10], the assessment of MIDs for QOL instruments is central to interpreting study results, in that they can both establish benchmarks for interpreting mean differences and inform the discussion of what constitutes a clinically meaningful change. A variety of techniques have been employed to determine MIDs [7], including distribution-based methods [11–13], anchor-based methods [14,15], empirical rules [16], or combinations of methods [17–23]. These techniques have been applied to many cancer-related QOL instruments, including the Functional Assessment of Cancer Therapy (FACT)-Lung, FACT-Prostate, FACT-Colorectal, and the FACT-Breast [17–22,24], but to date, there has been no comprehensive assessment of MIDs for the FACT-Melanoma (FACT-M) QOL questionnaire.

The Functional Assessment of Chronic Illness and Therapy is a patient reported outcome measurement system composed of a general health-related QOL component for patients with chronic disease coupled with disease-specific modules [25,26]. This hybrid approach for measuring patient reported outcomes has the advantage of allowing for both a more focused assessment of disease-specific profiles and symptoms while retaining comparability across populations due to the common core items [27]. For example, for patients undergoing treatment for cancer, the FACT-General (FACT-G) serves as the general component, and for those with specific malignancies such as melanoma, the general items of the FACT-G are supplemented with the melanoma-specific items of the FACT-M. The FACT-M was developed at the University of Texas M.D. Anderson Cancer Center and has been validated as a patient-reported QOL measure for melanoma patients with American Joint Committee on Cancer stages I through IV disease [28,29].

The objectives of this study were to derive MIDs for the FACT-M using anchor-based methods, to conduct a comparative distribution-based analysis, and to assess the correspondence
among MID ranges derived from each of the employed methods.

Methods

Clinical, demographic, and QOL data were collected prospectively at the Melanoma and Skin Center of the University of Texas M.D. Anderson Cancer Center in Houston, Texas with the approval of the institutional review board for the protection of human subjects. From 2004 to 2005, new patients and those within 3 years of melanoma diagnosis were recruited when they presented for scheduled appointments, with recruitment efforts targeting an equal proportion of patients with local, regional, and distant metastatic disease. With 80 patients per disease stage grouping (I and II, III, IV), a statistically significant difference in effect sizes of 0.45 or greater could be detected with 80% statistical power. Initial sampling targets were set at 300 participants to account for attrition, while assuring sufficient sample size at analysis. Study participants had to be at least 18 years of age and fluent in English, and enrollment required pathological confirmation of melanoma. Exclusion criteria included inability to consent due to other medical illnesses and disorientation to person, place, or time. Written informed consent was obtained for each of the study participants, and questionnaires were completed at baseline and during regularly scheduled follow-up visits. Clinical data were abstracted from patient files at baseline and for respective follow-up time points.

Instruments

The FACT-M is composed of items from the FACT-G, melanoma-specific items, and items related to melanoma surgery [28,29]. Four constructs comprise the FACT-G scale, with seven items assessing Physical Well-Being (PWB), seven items assessing Social/Family Well-Being, six items assessing Emotional Well-Being, and seven items assessing Functional Well-Being (FWB). With the addition of the 16-item Melanoma Subscale (MS) and the 8-item Melanoma Surgery Subscale (MSS)—collectively known as the Melanoma Combined Scale (MCS), the number of items of the FACT-M totals 51. Each of these subscales has been shown to have high levels of internal consistency (Cronbach’s $\alpha$: 0.71–0.95) and high test–retest reliability ($r$: 0.71–0.90) [29]. Higher scores on any subscale or total score indicate higher levels of QOL [26]. The Trial Outcome Index (TOI), which is often the primary outcome measure of QOL in melanoma trials [22], has been defined as the summed score of the seven PWB and seven FWB items from the FACT-G and the disease-specific subscale items.

The Karnofsky Performance Scale (KPS) is designed to assess a patient’s status in terms of functional impairment. With a range of 0 (dead) to 100 (normal—no evidence of disease), this scale assists clinicians and caretakers in gauging a patient’s ability to perform activities basic to daily living [30,31]. The KPS was administered at the time of study entry and 3 months later. The Eastern Cooperative Oncology Group Performance Status (ECOG-PS) scale is a measure of disease progression and its effect on a patient’s daily living [32,33]. The range of the scale is from 0 (representing fully active, pre-disease performance status) to 5 (death). Clinicians assessed patient performance status at baseline and at 3 months to correspond in timing with the other assessments.

Statistical Analysis

Baseline demographic and clinical data were summarized, and FACT-M TOI, MCS, MS, and MSS scores were calculated using STATA statistical software (v. SE9.2, StataCorp, College Station, TX, USA). Mean FACT-M scores were compared across groups based on relevant clinical indicators previously assessed for statistically significant score differences (e.g., disease stage and treatment status) [29]. Patient performance status was assessed at the time of study entry (baseline) and again after 3 months. For the KPS, an increase or decrease of 10 or more units on a 100-unit scale over time was considered clinically meaningful. The ECOG-PS scale is inverted, and an increase or a drop of one or more units on the six-unit scale was considered clinically meaningful. Patient groups (improving, remaining stable, and declining) were created based on these respective performance measures, and mean per-patient change in QOL scores were calculated for each of these groups. The differences between the QOL change scores of the various patient performance groups were identified and served as minimal units of clinically meaningful change in patient QOL [22].

To facilitate a patient-visit level analysis and to mitigate the effect of missingness that is common to studies involving QOL outcomes, a pattern mixture model analysis was employed following previously established methods [23,34,35]. Using SAS version 9.1 (SAS Institute, Cary, NC, USA) for Windows, separate analyses were conducted with the clinical anchors serving as independent variables, the FACT-M subscales serving as dependent variables, and several clinical variables with group-level differences in QOL such as disease stage, treatment status, and drop out status serving as covariates in the model.

Several distribution-based methods were also included for comparison, though the authors recognize that distribution-based methods provide no direct information about differences that are minimal and important [36,37]. From the methods of Wyrwich et al, the standard error of measurement (SEM) method was used to approximate minimum relevant thresholds of change, as the SEM has been shown to align with MID estimates for other important QOL instruments [13]. The SEM method employs the formula

$$SEM = \sigma_x (1 - relxx)^{1/2}$$  \hspace{1cm} (1)

where $\sigma_x$ is the standard deviation of the sample and $relxx$ is an appropriate measure of reliability for the instrument—in this case, Cronbach’s alpha. Separate SEMs were calculated for each time point (baseline and 3 months) and served as comparative reference points for the MID. Measures of effect size have also served as distribution-based reference points for MIDs [38–40]; however, effect sizes are distributionally standardized and thus not in the form of a scale score. Based on the correspondence of modified standard deviation measures (1/3 SD and 1/2 SD) to small and moderate effect sizes, respectively [38], the standard deviation measures of the FACT-M scale scores were divided by two and separately by three to reflect standards of change in the form of scale scores. This step was repeated for each measurement time point (i.e., baseline and 3 months), and again for the standard deviation of per-patient change.

As missing data is a common issue for studies assessing QOL [41], it was important to examine how missingness may have affected the MID estimates. Baseline demographic and clinical characteristics of study participants who completed 3 months of follow-up were compared to those for whom no follow-up data were available. Subsequent analyses included tests of association between missingness and each of the demographic and clinical variables, while the influence of potential covariates was examined.
Performance measures

Quality of Life measures

Surgical procedures

AJCC disease stage

Primary tumor site

Education

Marital status

Gender

Table 1 Demographic and clinical profile

<table>
<thead>
<tr>
<th>Median age in years (range)</th>
<th>52 (20–79)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Race/ethnicity</td>
<td></td>
</tr>
<tr>
<td>White, non-Hispanic</td>
<td>268 (98.1)</td>
</tr>
<tr>
<td>Hispanic</td>
<td>3 (1.1)</td>
</tr>
<tr>
<td>African-American</td>
<td>1 (0.4)</td>
</tr>
<tr>
<td>Other</td>
<td>1 (0.4)</td>
</tr>
<tr>
<td>Gender</td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>159 (58.2)</td>
</tr>
<tr>
<td>Female</td>
<td>114 (41.8)</td>
</tr>
<tr>
<td>Marital status</td>
<td></td>
</tr>
<tr>
<td>Married</td>
<td>218 (79.8)</td>
</tr>
<tr>
<td>Never married</td>
<td>27 (9.9)</td>
</tr>
<tr>
<td>Separated or divorced</td>
<td>21 (7.7)</td>
</tr>
<tr>
<td>Widowed</td>
<td>7 (2.6)</td>
</tr>
<tr>
<td>Education</td>
<td></td>
</tr>
<tr>
<td>&lt;High school</td>
<td>40 (14.6)</td>
</tr>
<tr>
<td>Some college</td>
<td>84 (30.8)</td>
</tr>
<tr>
<td>College graduate</td>
<td>92 (33.7)</td>
</tr>
<tr>
<td>Graduate school</td>
<td>53 (19.4)</td>
</tr>
<tr>
<td>Missing/Unknown</td>
<td>4 (1.5)</td>
</tr>
<tr>
<td>Primary tumor site</td>
<td></td>
</tr>
<tr>
<td>Head/Neck</td>
<td>41 (15.0)</td>
</tr>
<tr>
<td>Trunk/Back</td>
<td>69 (25.3)</td>
</tr>
<tr>
<td>Upper extremity</td>
<td>54 (19.8)</td>
</tr>
<tr>
<td>Lower extremity</td>
<td>79 (28.9)</td>
</tr>
<tr>
<td>Unknown primary</td>
<td>30 (11.0)</td>
</tr>
<tr>
<td>AJCC disease stage</td>
<td></td>
</tr>
<tr>
<td>Stages I and II</td>
<td>102 (37.4)</td>
</tr>
<tr>
<td>Stage III</td>
<td>100 (36.6)</td>
</tr>
<tr>
<td>Stage IV</td>
<td>71 (26.0)</td>
</tr>
<tr>
<td>Treatment status</td>
<td></td>
</tr>
<tr>
<td>Active treatment</td>
<td>75 (27.5)</td>
</tr>
<tr>
<td>Follow-up surveillance</td>
<td>198 (72.5)</td>
</tr>
<tr>
<td>Surgical procedures</td>
<td></td>
</tr>
<tr>
<td>Complete node dissection</td>
<td>133 (48.7)</td>
</tr>
<tr>
<td>Wide local excision and sentinel node biopsy</td>
<td>105 (38.5)</td>
</tr>
<tr>
<td>Wide local excision only</td>
<td>31 (11.4)</td>
</tr>
<tr>
<td>Fine needle aspiration</td>
<td>4 (1.5)</td>
</tr>
</tbody>
</table>

AJCC, American Joint Committee on Cancer.

Results

A total of 273 patients were enrolled in the prospective study, and 163 completed assessments at 3 months. Relatively even patient distributions were observed across levels of primary tumor site and education (Table 1). More than 25% of patients were undergoing active treatment at the time of study enrollment with the remainder in post-treatment follow-up surveillance. Stratifying change in QOL by baseline clinical characteristics uncovered significant group differences in QOL scale scores between patients with local-regional and advanced disease ($P < 0.001$) and between those in active treatment or in follow-up surveillance ($P < 0.001$). Differences in TOI scale scores ranged from 12 to 14 points, and for the MCS, they ranged from 6 to 7 points. For the MS and the MSS, differences ranged from 4 to 5 points and 1 to 3 points, respectively. These point differences were expected to provide theoretical upper limits to the MID estimates, as these patient groups were derived from clinically observed distinctions among patients and their corresponding treatment and QOL outcomes.

Table 2 outlines patient QOL as measured by mean FACT-M subscale scores. The mean TOI increased from baseline (130.8) to 3 months (136.9), and on the patient level, the mean per-patient change in TOI score was 5.6. Similar trends were seen for the subscales. As QOL scores improved over time, patient-reported performance status improved, as reflected in both the KPS and ECOG-PS scores. With increasing scores reflecting improvement in patient performance status, the baseline mean KPS score of 93.0 at baseline rose to 95.4 at 3 months with a mean per-patient change of 1.7. Similarly, performance status improved over time on the ECOG-PS with decreasing scale scores representing better health states. The mean ECOG-PS scale score at baseline was 0.3 and at 3 months was 0.2 with a mean per-patient change of –0.1 points.

Table 2 summarizes mean per-patient change in FACT-M subscale scores stratified by the anchor-based performance-change groups. Specifically, changes in QOL scores for patients whose performance status improved were compared with scores for patients whose performance scores decreased and with patients whose performance status remained stable. Differences in QOL change scores between the KPS performance groups ranged from 1.4 to 1.8 for the MSS, 1.9 to 3.4 for the MS, 3.7 to 4.8 for the MCS, and 5.1 to 7.1 for the TOI. Similarly, for the ECOG-PS performance groups, differences in QOL change scores ranged from 0.9 to 2.3 for the MSS, 2.6 to 4.6 for the MS, 4.9 to 5.5 for the MCS, and 8.3 points for the TOI.

Pattern mixture model analysis resulted in statistically significant MID estimates for each of the QOL subscales while controlling for the effect of covariates (Table 4). With the KPS serving as predictor, MID point estimates for the TOI, MCS, MS, and MSS were 9.0, 7.1, 4.6, and 2.6, respectively. When the ECOG-PS served as the predictor, MID point estimates were 12.2, 9.6, 6.1, and 3.5 for the respective subscales.

MID reference points from the standard deviation methods were similar with ranges of 1.3 to 3.0 for the MSS, 1.7 to 3.8 for the MS, 2.5 to 5.9 for the MCS, and 4.1 to 9.7 for the TOI (Table 5). Likewise, the SEM method yielded ranges of 1.5 to 2.3 for the MSS, 2.0 to 2.9 for the MS, 2.6 to 4.1 for the MCS, and 4.1 to 6.5 for the TOI. Figure 1 illustrates the ranges of MIDs for each of the FACT-M subscales stratified by method of analysis. Pattern mixture model estimation resulted in higher MID ranges than the other methods with the confidence interval acting as the...
upper and lower bounds of the suggested MID range. Average MID estimates from the anchor-based methods were higher than those from the distribution-based methods, and on nearly every scale, the MIDs from the ECOG-PS were higher than those from the KPS.

Because a nearly 40% attrition rate for this study was observed at 3 months, group comparisons between patients with and without follow-up data were made for all baseline demographic and clinical characteristics. Although no statistically significant group differences emerged from comparisons of treatment status, a significant difference was found for baseline disease stage \( (P < 0.001) \), with a larger proportion of patients with metastatic disease (regional and distant) among the missing (71.8%) than among those with follow-up data at 3 months (56.4%). Furthermore, those with metastatic disease at baseline were nearly twice as likely to have missing data at 3 months (OR 1.97; 95% CI [1.17–3.30], \( P = 0.01 \)), suggesting a statistical association. Subsequent examination of the relationship between disease stage and QOL scores revealed that baseline QOL scores differed for those with and without metastatic disease \( (P < 0.001) \). However, when controlling for disease stage, no association was observed between baseline QOL and missingness.

**Discussion**

A comprehensive assessment of MIDs makes use of multiple approaches to triangulate on a range of MID estimates [37], and

---

**Table 3** Mean per-patient change in FACT-M scale scores stratified by performance-change groups

<table>
<thead>
<tr>
<th>Change in Karofsky Performance Status at 3 months</th>
<th>Change in ECOG Performance Status at 3 months</th>
</tr>
</thead>
<tbody>
<tr>
<td>Improved (n = 40)</td>
<td>( \pm 1.4 )</td>
</tr>
<tr>
<td>Stable (n = 110)</td>
<td>( \pm 10.3 )</td>
</tr>
<tr>
<td>Declined (n = 13)</td>
<td>( \pm 14.1 )</td>
</tr>
</tbody>
</table>
| Diff (Improved \(
\rightarrow\n\) Stable) | \( \pm 7.6 \) | \( \pm 8.4 \) |
| Diff (Stable \(
\rightarrow\n\) Declined) | \( \pm 2.8 \) | \( \pm 6.2 \) |

**Table 4** Pattern mixture model estimates* stratified by performance-anchors

<table>
<thead>
<tr>
<th>Karnofsky Performance Scale</th>
<th>Estimate</th>
<th>95% CI</th>
<th>( P ) value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trial Outcome Index</td>
<td>8.95</td>
<td>(7.67–10.24)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Melanoma Combined Scale</td>
<td>7.12</td>
<td>(6.14–8.11)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Melanoma Subscale</td>
<td>4.61</td>
<td>(3.94–5.28)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Melanoma Surgery Subscale</td>
<td>2.58</td>
<td>(2.04–3.13)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Eastern Cooperative Oncology Group-Performance Scale (ECOG-PS)†</td>
<td>12.17</td>
<td>(9.99–14.36)</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>

**Table 5** Distribution-based ranges for minimal important differences

<table>
<thead>
<tr>
<th>FACT-Melanoma Trial Outcome Index (TOI)</th>
<th>SD baseline</th>
<th>SD 3 months</th>
<th>SD per-patient change</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard deviation (SD) method</td>
<td>19.46</td>
<td>16.41</td>
<td>12.31</td>
</tr>
<tr>
<td>Medium effect size (TOI/2)</td>
<td>9.73</td>
<td>8.21</td>
<td>6.16</td>
</tr>
<tr>
<td>Small effect size (TOI/3)</td>
<td>6.49</td>
<td>5.47</td>
<td>4.10</td>
</tr>
<tr>
<td>Melanoma Combined Subscale (MCS)</td>
<td>11.73</td>
<td>9.22</td>
<td>7.35</td>
</tr>
<tr>
<td>Medium effect size (MCS/2)</td>
<td>5.86</td>
<td>4.61</td>
<td>3.67</td>
</tr>
<tr>
<td>Small effect size (MCS/3)</td>
<td>3.91</td>
<td>3.07</td>
<td>2.45</td>
</tr>
<tr>
<td>Melanoma Subscale (MS)</td>
<td>7.60</td>
<td>6.11</td>
<td>5.16</td>
</tr>
<tr>
<td>Medium effect size (MS/2)</td>
<td>3.80</td>
<td>3.06</td>
<td>2.58</td>
</tr>
<tr>
<td>Small effect size (MS/3)</td>
<td>2.53</td>
<td>2.04</td>
<td>1.72</td>
</tr>
<tr>
<td>Melanoma Surgery Subscale (MSS)</td>
<td>5.91</td>
<td>4.79</td>
<td>3.84</td>
</tr>
<tr>
<td>Medium effect size (MSS/2)</td>
<td>2.95</td>
<td>2.39</td>
<td>1.92</td>
</tr>
<tr>
<td>Small effect size (MSS/3)</td>
<td>1.97</td>
<td>1.60</td>
<td>1.28</td>
</tr>
</tbody>
</table>

*Model accounts for effects of disease stage, treatment status, surgery, patient drop-out, and interactions.

†ECOG-PS scale is inverted, but absolute values are reported in the table for consistency.

---

FACT, Functional Assessment of Cancer Therapy.
the results of this study indicate that ranges for MIDs of the FACT-M scales are between 5 and 9 points for the TOI, 4 and 6 points for the MCS, 2 and 4 points for the MS, and 1 and 2 points for the MSS. These estimates are similar in range to those of other cancer-specific FACT subscales [18,21,22,42]. Although the anchor- and distribution-based methods employed to assess MIDs resulted in similar ranges, it is important to note that they were not the same, so the strengths and weaknesses of each method warrant further discussion.

One strength of the distribution-based methods is that they can be derived from a single sampling time point (e.g., cross-sectional study design), and the analysis requires few statistical elements to derive the ranges (e.g., the standard deviation of the sample and a measure of the instrument’s reliability). However, distribution-based methods lack clinical reference points, and because of this, a range of MIDs derived from distribution-based methods employed in isolation are likely to be viewed with skepticism in regards to their clinical meaning. In contrast, anchor-based methods for deriving MIDs make use of external reference points, and because of this, anchor-based methods generally require assessments of change in patient health status, and measures of change require multiple assessments of a study group over time along with the administration of multiple instruments.

Pattern mixture modeling is an anchor-based method that minimizes the effect of missingness by accounting for it in the model. Additionally, it allows visit-level analysis resulting in increased statistical power when multiple visits are present in the dataset. However, on several of the scales in this analysis, mixed model estimates were higher than group estimates derived from clinical distinctions that served as upper-bounds for the MIDs (i.e., treatment status, and disease stage). Our findings that patients with advanced disease were more likely to have missing data at 3 months and that they were the group expected to experience larger changes in QOL over time were not surprising. Because mixed modeling methods incorporate all visits and covariates, it is logical that MID estimates from mixed modeling were higher than those of the other anchor-based methods. Because the study objective was to ascertain the thresholds of minimal difference that are clinically meaningful, when mixed model estimates were above the upper thresholds set by clinical group distinctions, the estimates derived from mixed models were given less weight when triangulating the final suggested range for each subscale. There is still potential for bias from the lack of follow-up data for patients with rapidly progressing disease [41]; however, larger differences in change scores would likely result in larger MIDs, not smaller ones. Furthermore, the apparent association between baseline QOL and missingness was not
Minimal Important Differences of the FACT-M

significant when controlling for covariates, and as such, further adjustment of the MID ranges to account for missingness may not be warranted.

At issue is what the MID ranges derived from each method actually represent. One previously noted distinction between the distribution- and anchor-based methods for deriving MIDs is that they reflect separate underlying types of differences—the former reflecting a minimal statistically quantifiable difference in scores and the latter reflecting a minimal clinically meaningful difference in scores [43]. With this distinction in mind, comments by de Vet et al. underscore the importance of maintaining an integrated approach, in that by demonstrating that the clinically meaningful differences (e.g., anchor-based MIDs) are above the thresholds set by minimum detectable differences (e.g., distribution-based MIDs), the choice of the anchor instrument is supported [43], as was found in this study. However, minor discordance was observed among the MID ranges derived from the anchor instruments, which warranted further examination of the properties of each anchor instrument.

Although both the KPS and the ECOG-PS scale scores have been strongly correlated with patient survival [44], there remains an important distinction in their general mode of administration, as the ECOG-PS is typically a physician-assessed measure, while the KPS is often patient-assessed. Although others have suggested that the ECOG-PS is a viable substitute for the KPS given its demonstrated validity and reliability [45], the findings of this study illustrate that substitution yields different point ranges for MIDs. In the FACT-M validation study [29], FACT-M scores were found to be more sensitive to change in performance status when performance status was assessed by the KPS. The observed increase in sensitivity with the KPS may be in-part due to issues of instrument granularity as the KPS operates on a scale of 0 to 100 while the ECOG-PS operates on scale of 0 to 5. However, differences in granularity are unlikely to account for all of the observed differences in MID estimates, as interrater variability generally increases with the increasing number of scale increments available to raters [45]. Given that the acknowledged limitations of this study include a relatively small sample size, missingness of data, and positively skewed patient status measures, it is possible that the MIDs would realign with a larger population of patients.

Evidence that patients provide valuable and reliable evaluations of their own performance status and QOL that is better correlated than that of their clinicians [45–47] supports the view that MID estimates may be more appropriately derived from patient-assessed anchor instruments—particularly when using QOL instruments like the FACT-M that are also patient-assessed. This point is particularly relevant to the clinical trial setting, as the definition of a responder depends on more than just statistical precision [7]: it depends upon the validity of the reference instruments imparting clinical meaning to the observed mean differences. Our findings that the properties of the anchor instrument directly affect the resulting MIDs support the view that when MIDs are cited, the source anchor instrument should be included as part of the evaluation process.

Source of financial support: National Cancer Institute, Clinical Oncology Research Development Program, 5-K12-CA088084.

References


Mapping the Cancer-Specific EORTC QLQ-C30 to the Preference-Based EQ-5D, SF-6D, and 15D Instruments

Nick Kontodimopoulos, PhD¹, Vassilis H. Aletras, PhD¹,² Dimitris Paliouras, MD,¹,³ Dimitris Niakas, PhD¹

¹Hellenic Open University, Faculty of Social Sciences, Patras, Greece; ²University of Macedonia, Department of Business Administration, Thessaloniki, Greece; ³Thaegonio Cancer Hospital, Thessaloniki, Greece

ABSTRACT

Objectives: To estimate models, via ordinary least squares regression, for predicting Euro Qol 5D (EQ-5D), Short Form 6D (SF-6D), and 15D utilities from scale scores of the European Organization for Research and Treatment of Cancer Quality of Life Questionnaire (EORTC QLQ-C30).

Methods: Forty-eight gastric cancer patients, split up into equal subgroups by age, sex, and chemotherapy scheme, were interviewed, and the survey included the QLQ-C30, SF-36, EQ-5D, and 15D instruments, along with sociodemographic and clinical data. Model predictive ability and explanatory power were assessed by root mean square error (RMSE) and adjusted $R^2$ values, respectively. Pearson’s $r$ between predicted and reported utility indices was compared. Three random subsamples, half in size the initial sample, were created and used for “external” validation of the modeling equations.

Results: Explanatory power was high, with adjusted $R^2$ reaching 0.909, 0.833, and 0.611 for 15D, SF-6D, and EQ-5D, respectively. After normalization of RMSE to the range of possible values, the prediction errors were 12.0, 5.4, and 5.6% for EQ-5D, SF-6D, and 15D, respectively. The estimation equations produced a range of utility scores similar to those achievable by the standard scoring algorithms. Predicted and reported indices from the validation samples were comparable thus confirming the previous results.

Conclusion: Evidence on the ability of QLQ-C30 scale scores to validly predict 15D and SF-6D utilities, and to a lesser extent, EQ-5D, has been provided. The modeling equations must be tried in future studies with larger and more diverse samples to confirm their appropriateness for estimating quality-adjusted life-year in cancer-patient trials including only the QLQ-C30.

Keywords: utilities, mapping, EQ-5D, SF-6D, 15D, EORTC QLQ-C30.

Introduction

Cost-utility analyses (CUA) require the calculation of quality-adjusted life-years (QALYs), which account for quality (morbidity) and duration (mortality) of life in a single outcome measure. The quality adjustment is based on a set of weights (utilities), which reflect the preferences of the community for a certain health state relative to others. These weights are assumed to provide interval scaled information, where 1 and 0 refer to full health and death, respectively. Negative utilities may also occur if very severe health states are evaluated as being worse (i.e., less preferable) than death. Utilities are preference based and derived by the community either via direct valuation techniques, such as standard gamble (SG), time trade-off (TTO), or rating scale, or indirectly via generic health-related quality of life (HRQoL) measures, such as the 15D [1], the Health Utility Index (HUI) [2,3], the Euro Qol 5D (EQ-5D) [4] or the Short Form 6D (SF-6D) [5], for which scoring algorithms have been developed to yield community-based health utility estimates. According to recent review study, HRQoL instruments were used in 76% of the studies in which QALYs were calculated from patient measurements, with approximately 60% of them involving the EQ-5D [6].

When the interest is in measuring the HRQoL of patients, disease-specific instruments are often preferred to generic ones because they focus on particular health problems and tend to be more sensitive to clinically important differences [7]. Nevertheless, disease-specific measures do not have utility scoring systems, and CUA requires a single, preference-based score so that QALYs can be calculated. On the other hand, using both types of instruments adds to the burden imposed on patients for completing questionnaires. A solution to this problem is to attempt to predict utilities by “mapping,” i.e., regressing preference-based indices against dimension or item scores of disease-specific measures and thus obtain estimation models that can be used to subsequently calculate QALYs [8].

Mapping between measures is a fairly new research area, and, according to a recent review of studies mapping nonpreference to preference-based measures, the majority of such studies have been published or produced after 2000 [9]. The most popular target measure has been by far the EQ-5D, followed by the HUI2/HUI3 and the SF-6D. Interestingly, the 15D has been involved in only one mapping effort, in which the starting measure was a self-assessed health question (and not a condition-specific instrument) [10]. Disease-specific measures having been used in mapping studies covering asthma [8], inflammatory bowel disease [11], obesity [12], angina [13], dental conditions [14], intermittent claudication [15], and other diseases. As for cancer, the European Organization for Research and Treatment of Cancer Quality of Life Questionnaire (EORTC QLQ-C30) has been used to predict EQ-5D indices and dimensions with ordinary least squares (OLS) and probit regression, respectively. The model explaining EQ-5D values predicted well, whereas prediction of EQ-5D values based on predicted levels for each dimension was less successful [16]. The QLQ-C30 has also been used, along with the Functional Assessment of Cancer Therapy—Prostate Questionnaire, to construct and validate an EQ-5D prediction model for metastatic hormone-refractory prostate cancer patients, and showed good predictive validity [17].

The purpose of this study was to establish, using a common data set and OLS regression, empirical mapping relationships between the cancer-specific QLQ-C30 and three
preference-based generic instruments, the EQ-5D, SF-6D, and 15D, and to compare performance of the resulting models in terms of predictive ability and goodness of fit. The sample comprised of gastric cancer patients currently on chemotherapy. Although the QLQ-C30 has been designed to measure different HRQoL concepts than the preference-based instruments, demonstration of robust relationships and testing of the models with external “validation” samples can increase the confidence placed on utilities obtained in trials not including any generic instruments.

**Methods**

**Instruments**

The EuroQol EQ-5D consists of a descriptive health state classification system with five domains (mobility, self-care, usual activities, pain/discomfort, and anxiety/depression) and three severity levels in each (no problems, some problems, and extreme problems) [4]. Combining one level from each domain defines 243 different health states ranging from full to worst health. Direct valuations for 42 health states were elicited from 3395 persons of the UK general public using the TTO method. Regression techniques were applied to interpolate values for the other health states, and the utilities range from 0.59 to 1.00 [18]. The EQ-5D was found to be applicable and adaptable to the Greek environment [19], and its construct validity was demonstrated using a large sample of the Greek general population [20].

The SF-6D consists of a health status classification system with six domains (physical functioning, role limitations, social functioning, emotions, mental health and vitality) and four to six severity levels in each, giving 18,000 unique health states. The scoring model is based on the SG utilities of 836 members of the UK general population, and, like the EQ-5D, regression models predicted utility scores ranging from 0.30 (worst health) to 1.00 (full health) for each of the possible health states [5]. The SF-6D is derived from the SF-36 (used in this study), which has been translated into Greek, and its reliability and validity were established in a sample of 1007 adults living in the greater Athens area [21,22]. The only Greek study having used the SF-6D in a cost-utility analysis demonstrated it to be satisfactorily valid [23].

The 15D is a generic HRQoL instrument with 15 dimensions: mobility, vision, hearing, breathing, sleeping, eating, speech, elimination, usual activities, mental function, discomfort and symptoms, depression, distress, vitality, and sexual activity, with each divided into five possible response levels, structured from the best to the worst possible health condition [1]. The valuation system of the 15D is based on an application of the multiattribute utility theory. The single index (15D score) on a 0 to 1 scale (death and full health, respectively) is calculated from the health state descriptive system by using a set of population-based preference or utility weights. A weight for each level of each dimension is obtained by multiplying the level value by the importance weight of the dimension at that level. The level values on the 0 to 1 scale, reflecting the goodness of the levels relative to no problems on the dimension and to being dead, and the importance weights summing up to unity, have been elicited from representative population samples by using a combination of rating scale and magnitude estimation methods (ratio scale with quantifiers) [1]. The 15D Questionnaire is relatively brief, easy to use, and has been shown to be reliable, valid, acceptable, and responsive. It is the only utility instrument with a Greek scoring algorithm [24]. Nevertheless, because of the absence of national scoring weights for the other two instruments, the original Finnish scoring was used in this study.

The EORTC QLQ-C30 is a popular instrument for measuring general cancer quality of life [25]. Most of the 30 questions have four response levels (not at all, a little, quite a bit, and very much), with two questions relying on a 7-point scale. Raw questionnaire responses are transformed to produce scores on five function scales (physical, role, emotional, cognitive, and social functioning) and nine symptom scales along with a scale representing global quality of life. Higher functional scale scores (range: 0–100) indicate better HRQoL, but higher symptom scale/item scores indicate higher level of symptoms. The scales have undergone psychometric testing, based on classical test theory, which has yielded favorable results [26]. The Greek version has also been shown to be reliable in a sample of cancer patients under palliative care [27].

**Sample and Data Collection**

To ensure that major patient characteristics were represented in the sample, a nonprobability quota-based sampling procedure was employed. Equiproportional groups were formed according to sex, age (<60 and ≥60 years old) and treatment scheme (potent vs. mild which may result in more or less undesired side effects, respectively). All the information was collected between November 2007 and March 2008. Patients were approached for participation during their visit to the “Theagenio” Cancer Hospital in Thessaloniki for chemotherapy (before treatment). Selected patients had previously attended two to four chemotherapy ses-

| Table 1 Descriptive statistics for EQ-5D, SF-6D, and 15D utilities and QLQ-C30 scales |
|--------------------------------------|-------------------------------------|-------------------------------------|-------------------------------------|-------------------------------------|----------------------------------|
| Statistics | Utility scores | Functional scales | EORTC QLQ-C30 scales |
| EQ-5D | SF-6D | 15D | QL | PF | RF | EF | CF | SF |
| Mean (SD) | 0.550 (0.307) | 0.606 (0.094) | 0.685 (0.166) | 46.35 (17.10) | 57.92 (26.10) | 31.60 (30.47) | 50.93 (24.90) | 85.42 (21.09) | 42.71 (25.94) |
| 95% CI | 0.461–0.639 | 0.579–0.633 | 0.636–0.733 | 41.39–51.32 | 50.34–65.50 | 33.33–66.66 | 41.67–83.33 | 66.67–100.00 | 33.33–66.66 |
| 25–75 percentiles | 0.290–0.774 | 0.556–0.674 | 0.574–0.819 | 33.33–56.25 | 35.00–80.00 | 31.60–59.03 | 41.67–83.33 | 66.67–100.00 | 33.33–66.66 |
| Min.–max. | 0.00–1.00 | 0.00–1.00 | 0.00–1.00 | 0.00–100.00 | 0.00–100.00 | 0.00–100.00 | 0.00–100.00 | 0.00–100.00 | 0.00–100.00 |
| % Floor–% ceiling | 2.1–2.1 | 2.1–2.1 | 2.1–2.1 | 2.1–2.1 | 2.1–2.1 | 2.1–2.1 | 2.1–2.1 | 2.1–2.1 | 2.1–2.1 |
| Z’ value (P–sig.) | 1.283 (0.074) | 0.716 (0.684) | 0.586 (0.883) | 1.147 (0.144) | 0.960 (0.315) | 1.281 (0.075) | 0.835 (0.489) | 2.347 (<0.001) | 0.583 (0.023) |

*Higher functional scale scores indicate better health-related quality of life. |
*Higher symptom scale/item scores indicate higher level of symptoms. |
*Kolmogorov–Smirnov test. |

AF: appetite loss; CF: cognitive functioning; CI: confidence interval; CO: constipation; DI: diarrhoea; DT: dyspepsia; EF: emotional functioning; EQ-5D: Euro Qol 5D; EORTC: EORTC QLQ-C30; European Organization for Research and Treatment of Cancer; QoL: Quality of Life Questionnaire; FA: fatigue; FI: financial difficulties; NV: nausea/vomiting; PA: pain; PF: physical functioning; QL: global health status/QoL; RF: role functioning; SF: social functioning; SF-6D: Short Form 6D; SL: insomnia.
sions, and at least 20 days had passed since the last one. The final sample consisted of 48 (N = 48) gastric cancer patients, all of which had undergone surgery. No patients were suffering from metastases of the cancer to other organs, which could further affect their HRQoL negatively. To minimize missing data, patients were interviewed by the same interviewer, and the survey included the EQ-5D, SF-36, and 15D instruments, the cancer-specific QLQ-C30, and common demographic questions. Disease-related data were taken from the patients’ history. The interview lasted approximately 30 to 40 minutes, and none of the patients declined participation. The hospital’s Review Board ethically approved the study and all participants provided informed consent.

**Analysis**

The Kolmogorov-Smirnov test was used to demonstrate distribution normality of EQ-5D, SF-6D, 15D, and QLQ-C30 scale scores. Although this clearly was not a QLQ-C30 validation study, some validity evidence was required before actually testing the predictive ability of the instrument. Pearson’s correlations between QLQ-C30 and SF-36 scales were used to assess convergent construct validity. Based on the literature [28,29], it was hypothesized that scales measuring similar dimensions of HRQoL should be strongly correlated (at least >0.50). QLQ-C30 scale scores were also compared between subjects to assess known-groups construct validity. It was hypothesized that younger subjects and those not reporting coronary disease would have better HRQoL, as would those undergoing less potent chemotherapy. Internal consistency reliability of each QLQ-C30 scale was assessed using Cronbach’s alpha, and the 0.70 standard for group-level comparisons was adopted [30].

Ordinary least squares (OLS) regression was used to model the EQ-5D, SF-6D, and 15D using QLQ-C30 scale scores as predicting variables. The models defined in this study are additive, implying linear independence between predictor variables. Reported utility indices were modeled via a stepwise inclusion procedure, which identified significant QLQ-C30 predictors. Full OLS models (all QLQ-C30 scales entered as predictors) were created and used to test the modeling equations generated by the full sample. These samples were created by randomly selecting half the sample, and thus were at least partly different from the estimation sample. The procedure was repeated three times to enhance confidence in the validation exercise. All regressions were run using STATA version 8.0 (StataCorp LP, College Station, TX), and data were analyzed using SPSS version 13.0 (SPSS Inc, Chicago, IL).

**Results**

Sex and age group as important demographic characteristics and treatment scheme as a clinical factor were taken as the HRQoL-influencing variables that should be represented in the study sample, which was comprised of 48 patients. Based on equiproportional sampling, each sex group contained an equal number of participants from four age groups: 40 to 49, 50 to 59, 60 to 69, and >70, and each was equally divided into patients on milder chemotherapy and on a more potent scheme. From the overall sample, 25 patients (52.1%) had 9 years (or less) of schooling, and 30 (62.5%) were in the workforce. As for comorbid conditions, 14 patients (29.2%) reported diabetes, and the same percentage suffered from coronary disease. Finally, 12 patients (25.0%) were taking tranquillizers to help them deal with their condition.

EQ-5D, SF-6D, 15D, and QLQ-C30 descriptive statistics are presented in Table 1. Regarding the three utilities, the 15D gave the highest reported index (0.685), a finding that agrees with

<table>
<thead>
<tr>
<th>Symptom scales/items</th>
<th>FA</th>
<th>NV</th>
<th>PA</th>
<th>DY</th>
<th>SL</th>
<th>AP</th>
<th>CO</th>
<th>DI</th>
<th>FI</th>
</tr>
</thead>
<tbody>
<tr>
<td>74.07 (25.25)</td>
<td>73.61 (28.11)</td>
<td>26.39 (20.58)</td>
<td>52.78 (24.63)</td>
<td>41.67 (34.04)</td>
<td>38.33 (31.13)</td>
<td>8.33 (20.05)</td>
<td>56.25 (40.19)</td>
<td>39.58 (26.32)</td>
<td></td>
</tr>
<tr>
<td>77.78</td>
<td>83.33</td>
<td>33.33</td>
<td>66.67</td>
<td>66.67</td>
<td>66.67</td>
<td>0.00</td>
<td>66.67</td>
<td>33.33</td>
<td></td>
</tr>
<tr>
<td>58.33–100.00</td>
<td>0.00–100.00</td>
<td>0.00–33.33</td>
<td>33.33–66.67</td>
<td>0.00–66.67</td>
<td>33.33–91.67</td>
<td>0.00–0.00</td>
<td>8.33–100.00</td>
<td>33.33–66.67</td>
<td></td>
</tr>
<tr>
<td>0.00–100.00</td>
<td>0.00–100.00</td>
<td>0.00–66.67</td>
<td>0.00–100.00</td>
<td>0.00–100.00</td>
<td>0.00–100.00</td>
<td>0.00–100.00</td>
<td>0.00–100.00</td>
<td>0.00–100.00</td>
<td></td>
</tr>
<tr>
<td>2.1–29.2</td>
<td>2.1–41.7</td>
<td>31.3–8.3</td>
<td>4.2–10.4</td>
<td>29.2–12.5</td>
<td>8.3–25.0</td>
<td>81.3–2.1</td>
<td>25.0–35.4</td>
<td>16.7–6.3</td>
<td></td>
</tr>
<tr>
<td>1.126 (0.159)</td>
<td>1.682 (0.007)</td>
<td>2.070 (0.001)</td>
<td>1.831 (0.002)</td>
<td>1.284 (0.074)</td>
<td>1.425 (0.34)</td>
<td>3.282 (0.001)</td>
<td>1.497 (0.023)</td>
<td>2.993 (0.001)</td>
<td></td>
</tr>
<tr>
<td>–0.930</td>
<td>–0.725</td>
<td>0.083</td>
<td>0.196</td>
<td>0.219</td>
<td>–0.123</td>
<td>2.909</td>
<td>–0.722</td>
<td>0.460</td>
<td></td>
</tr>
</tbody>
</table>
results from another study in which the same three instruments were administered [31]. The SF-6D expectedly outscored the EQ-5D because it has been shown to give higher scores in samples with impaired health, in contrast to the EQ-5D, which usually gives higher scores in healthier populations [32,33]. The EQ-5D scores were the most negatively skewed (i.e., they clustered at the upper extreme); however, being a highly morbid sample, only one respondent was at the EQ-5D ceiling. Floor and ceiling effects were minimal for the other two measures as well. None of the instruments reached their theoretically achievable range. Kolmogorov–Smirnov Z-tests showed a normal data distribution for all three measures. As for the EQ-5D scales, cognitive functioning was the highest scoring functional scale (85.42), and role functioning the lowest (31.60), indicating best health status was significant in all three models, whereas role functioning was the only functional scale not appearing in any model. For all utility indices, at least two functional scales were significant predictors, with the physical functioning and emotional functioning scales appearing in two of the models. Four symptoms scales (pain, constipation, dyspnea, and insomnia) also appeared in at least one of the models as significant predictors. These stepwise regression models can be used to predict EQ-5D, SF-6D, and 15D utility scores, according to equations that are created from the information in Table 3.

The explanatory power of the regression models—represented by adjusted $R^2$—was relatively high, indicating that they explain large parts of the variation, especially in the case of

<table>
<thead>
<tr>
<th>Table 2</th>
<th>Pearson’s correlation coefficients between QLQ-C30 and SF-36 scales</th>
</tr>
</thead>
<tbody>
<tr>
<td>QLQ-C30 scale</td>
<td>PF</td>
</tr>
<tr>
<td>Global health status</td>
<td>0.62*</td>
</tr>
<tr>
<td>Physical functioning</td>
<td>0.87*</td>
</tr>
<tr>
<td>Role functioning</td>
<td>0.71*</td>
</tr>
<tr>
<td>Emotional functioning</td>
<td>0.24</td>
</tr>
<tr>
<td>Social functioning</td>
<td>0.61*</td>
</tr>
<tr>
<td>Fatigue</td>
<td>0.58*</td>
</tr>
<tr>
<td>Pain</td>
<td>0.56*</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 3</th>
<th>Multiple stepwise linear regression models</th>
</tr>
</thead>
<tbody>
<tr>
<td>Utility</td>
<td>QLQ-C30 predictors</td>
</tr>
<tr>
<td>EQ-5D</td>
<td>Physical functioning</td>
</tr>
<tr>
<td></td>
<td>Emotional functioning</td>
</tr>
<tr>
<td></td>
<td>Global health status</td>
</tr>
<tr>
<td></td>
<td>(Constant)</td>
</tr>
<tr>
<td>SF-6D</td>
<td>Social functioning</td>
</tr>
<tr>
<td></td>
<td>Global health status</td>
</tr>
<tr>
<td></td>
<td>Emotional functioning</td>
</tr>
<tr>
<td></td>
<td>Pain</td>
</tr>
<tr>
<td></td>
<td>Constipation</td>
</tr>
<tr>
<td></td>
<td>Dyspnea</td>
</tr>
<tr>
<td>15D</td>
<td>(Constant)</td>
</tr>
<tr>
<td></td>
<td>Physical functioning</td>
</tr>
<tr>
<td></td>
<td>Global health status</td>
</tr>
<tr>
<td></td>
<td>Insomnia</td>
</tr>
<tr>
<td></td>
<td>Cognitive functioning</td>
</tr>
<tr>
<td></td>
<td>(Constant)</td>
</tr>
</tbody>
</table>

*Corresponds to the normalized RMSE.
†Corresponds to the range of achievable utility values from the mapping equations compared to the standard scoring algorithms.

EQ-5D, Euro Qol 5D; EQ-5D, European Organization for Research and Treatment of Cancer Quality of Life Questionnaire; RMSE, root mean square error; SE, standard error of the mean; SF-6D, Short Form 6D.
15D and SF–6D (90.9% and 83.3%, respectively), and lower in EQ-5D (61.1%). High explanatory power does not necessarily imply good predictive ability as well; therefore, RMSE was the important indicator to be examined. In terms of predictive ability, the SF–6D model outperformed the others, with a RMSE of 0.038. The 15D model followed, showing slightly poorer predictive ability, i.e., RMSE = 0.050, and by far, the highest prediction errors corresponded to the EQ–5D model, with a RMSE of 0.192. To compare across instruments, normalized values (i.e., % RMSE) were calculated. The results for EQ–5D, SF–6D, and 15D prediction models were 12.0, 5.4, and 5.6% respectively, confirming that the EQ–5D model generates the weakest predictions.

For each instrument, the measuring range of the prediction equations was examined and compared with the range of utilities achievable by the standard scoring algorithms. At the lower end, the estimated SF–6D equation can generate lower values than the standard algorithm (0.212 vs. 0.301), whereas at the upper end the estimated EQ–5D slightly exceeded unity (1.186), as did 15D, but just marginally (1.019). Overall, the estimated 15D measuring limits were the ones closest to the standard range (i.e., 0.161–1.019 vs. 0.106–1.000). Estimated EQ–5D, SF–6D, and 15D utilities were compared with mean reported utilities, i.e., those derived directly from the instruments, and the correlation was high in all models, with Pearson’s r reaching 0.797, 0.924, and 0.959 (P < 0.01, throughout) for EQ–5D, SF–6D, and 15D, respectively.

Correlation between reported and predicted indices, from the three “external” data sets, was strong and statistically significant (P < 0.01) for all utility instruments (Table 4). It is also worth mentioning that reported utilities were well within the 95% confidence intervals of the predicted ones. A minimally important difference (MID)—defined as the smallest score difference which the patients perceive as beneficial [34]—was reached in the case of EQ–5D in the third validation sample because the estimated difference between reported and predicted utility indices was 0.03, which constitutes an EQ–5D MID typically reported in the literature [35]. The same MID (0.03) has been reported for the SF–6D [36] and the 15D [37], but was not exceeded here, and this increases confidence in the predicted utilities for these two instruments. In the case of EQ–5D, MID estimation efforts have resulted in higher values. In a review of eight longitudinal studies involving 11 patient groups, MIDs ranged from −0.011 to 0.140, and in nine of these groups exceeded 0.03 [38]. In any case, defining a valid MID for the EQ–5D calculated by the British tariff may be futile because of the large range of utility values generated.

### Discussion

Mapping can be valuable for analyzing data from older studies in which no preference-based instruments were used. When possible, however, it should be chosen as the alternative strategy because generic and disease-specific instruments serve different purposes, and one cannot substitute completely for the other. This study attempted to predict utility indices for three widely used instruments in economic evaluations, the EQ–5D, SF–6D, and 15D, using patient data from an also widely used disease-specific questionnaire. The obvious benefits would be the potential to perform cost-utility analyses with disease-specific questionnaires—something not usually possible—while exploiting the increased sensitivity of condition-specific instruments to aspects of a disease usually overseen by generic ones. The disease in question was gastric cancer, and the chosen questionnaire was the EORTC’s QLQ-C30, which is an integrated system for assessing the HRQoL of cancer patients.

The present study is the first, to our knowledge, to predict and compare utility indices for three preference-based instruments using the same data set. It is also the first to map any disease-specific questionnaire on to the 15D. OLS regression was used, and the models were compared mainly in terms of predictive ability, and, to a lesser extent, explanatory power. Equiproportional groups according to age, sex, and treatment scheme were formed, and the final sample comprised of 48 gastric cancer patients. Although the common source of data in similar studies are clinical trials or large survey panel data, there have been mapping attempts with small (<100) data sets [9].

Validation of the EORTC QLQ-C30 itself was not among the explicit objectives of this study; however, evidence was accumulated to support convergent and “known-groups” construct validity and internal consistency reliability. The SF–36, which has been shown to be valid and reliable in the Greek population [21,22], was used as the “standard,” and the hypothesized strong correlations between similar SF–36 and QLQ-C30 scales were confirmed. In light of the limited—until now—use of this instrument in Greek settings, such information is important in terms of the confidence to be placed on utility indices predicted from this cancer-specific instrument, but also adds to the existing knowledge from the only previous Greek study having tested and confirmed the psychometric properties of the QLQ-C30 [27].

The purpose of this study was to derive mapping algorithms to be applied to trial data to calculate utilities at various time points during the course of a treatment, and thus be used to derive QALYs. Although the prediction models were estimated in a specific context, i.e., in gastric cancer patients under chemotherapy, the equations can be applied to other cancer patients with similar QLQ-30 scores, or to cancer patients with similar scores who are receiving chemotherapy. Based on QLQ-C30 scale scores, EQ–5D, SF–6D, and 15D utility indices were predicted via OLS regression. The models excluded patient background characteristics such as age, sex, income, and others because this would increase model complexity, and as it has been shown in a recent review of mapping studies “Quite modest or negligible improvements were achieved from increasing model complexity” [9].

Explanatory power, particularly for the 15D and SF–6D models, was high (adjusted R² > 0.80), implying that QLQ-C30 scales explained large parts of the variation in these two utility indices. Because no 15D mapping attempts have been recorded in the literature, there were obviously no results with which to

| Table 4 Reported versus predicted utilities for three random validation samples |
|-----------------------------|-----------------------------|-------------------------------|-----------------------------|
| Validation samples         | Reported utilities          | Mean (95% CI) predicted utilities | Pearson’s r |
| Sample no. 1               | EQ–5D                       | 0.498 (0.412–0.596)           | 0.756*                     |
|                           | SF–6D                       | 0.582 (0.553–0.620)           | 0.909*                     |
|                           | 15D                         | 0.653 (0.589–0.716)           | 0.951*                     |
| Sample no. 2               | EQ–5D                       | 0.505 (0.408–0.619)           | 0.842*                     |
|                           | SF–6D                       | 0.587 (0.555–0.623)           | 0.878*                     |
|                           | 15D                         | 0.664 (0.594–0.731)           | 0.963*                     |
| Sample no. 3               | EQ–5D                       | 0.521 (0.435–0.668)           | 0.881*                     |
|                           | SF–6D                       | 0.605 (0.569–0.654)           | 0.934*                     |
|                           | 15D                         | 0.680 (0.610–0.761)           | 0.971*                     |

*P < 0.01. CI, confidence interval; EQ–5D, Euro Qol 5D; SF–6D, Short Form 6D.
compare $R^2$ or any other model parameters, for that matter. On the other hand, explanatory power for the SF-6D model was higher than relative values reported in mapping studies involving the SF-6D and disease-specific instruments (but not the QLQ-C30) [9,12]. As for the EQ-5D, the regression model explained 61% of its variation. This figure is in agreement with results from a study involving the QLC-C30 and the EQ-5D, in which two QLQ-C30 scales, global health status, and emotional functioning, were significant predictors, as in the present study [16]. In an attempt to improve explanatory power, we incorporated squared QLQ-C30 dimension scores into the EQ-5D estimation model; however, the improvement was modest ($R^2 = 0.687$). Inclusion of interaction or power terms in estimation models is common, but only rarely is there a noteworthy improvement. One exception is a mapping between BDQ to EQ-5D, where $R^2$ increased from 0.45 to 0.69 after incorporating squared terms of dimension scores [11].

Predictive ability of the models was highest for SF-6D (RMSE $= 0.038$), followed by 15D (RMSE $= 0.050$), and lower for EQ-5D (RMSE $= 0.192$), or in terms of normalized %RMSE values, 5.4, 5.6, and 12.0%, respectively. The strength of mapping usually depends on the overlap between two instruments in terms of their descriptive systems. By comparing the names of the domains, it appears that the EQ-5D is just as much covered by the QLQ-C30 as are the SF-6D and 15D instruments. On the other hand, the EQ-5D, because of its relative simplicity, does not address as many of the QLQ-C30 domains as the other two instruments, and this may explain in part the observed lower predictive ability of the EQ-5D model. Nevertheless, some dimensions (e.g., pain) are heavily weighted in the original scoring algorithms (e.g., SF-6D and EQ-5D), but are not necessarily predictive of the respective utility. For example, in this study, the QLC-C30 pain dimension is predictive of only SF-6D (and not EQ-5D) utilities, implying that overlap alone cannot explain the reduced predictive ability of the EQ-5D model.

The EQ-5D, SF-6D, and 15D estimation equations from this study produced an achievable range of utility scores, which was in close proximity to that given by the standard scoring algorithms, and, in the case of 15D, the predicted and standard scales were very close in magnitude and correlated strongly throughout (Pearson’s $r$ was similar to the values reported in Table 4). Although these results may be regarded as encouraging, at least for SF-6D and 15D, future studies involving larger and more diverse patient samples are needed to better assess estimated prediction models. The most obvious limitation in the present study was the size of the sample, which may be regarded as relatively small despite representing major patient demographic and clinical variables on an equal basis. Nevertheless, the results, which were tested using standard methods typically employed in similar studies, were quite encouraging, and when further validated, may contribute to the aim of performing cost-utility analyses by means of the EORTC’s QLQ-C30 alone.

References

36 Walters SJ, Brazier JE. What is the relationship between the minimally important difference and health state utility values? The case of the SF-6D. Health Qual Life Outcomes 2003;1:4.
38 Walters SJ, Brazier JE. Comparison of the minimally important difference for two health state utility measures: EQ-5D and SF-6D. Qual Life Res 2005;14:1523–32.

Pennifer Erickson, PhD,¹ Richard Willke, PhD,² Laurie Burke, RPh, MPH³

¹OLGA, State College, PA, USA; ²Global Outcomes Research, Pfizer Inc., Peapack, NJ, USA; ³Study Endpoints and Labeling, Office of New Drugs, Center for Drug Evaluation and Research, Food and Drug Administration, Silver Spring, MD, USA

Abstract

Objective: To facilitate development and evaluation of a PRO instrument conceptual framework, we propose two tools—a PRO concept taxonomy and a PRO instrument hierarchy. FDA’s draft guidance on patient reported outcome (PRO) measures states that a clear description of the conceptual framework of an instrument is useful for evaluating its adequacy to support a treatment benefit claim for use in product labeling. The draft guidance, however, does not propose tools for establishing or evaluating a PRO instrument’s conceptual framework.

Methods: We draw from our review of PRO concepts and instruments that appear in prescription drug labeling approved in the United States from 1997 to 2007. We propose taxonomy terms that define relationships between PRO concepts, including “family,” “compound concept,” and “singular concept.” Based on the range of complexity represented by the concepts, as defined by the taxonomy, we propose nine instrument orders for PRO measurement. The nine orders range from individual event counts to multitem, multiscale instruments.

Conclusion: This analysis of PRO concepts and instruments highlights the importance of the conceptual framework of an instrument to support treatment benefit claims. The tools described here offer a systematic approach to establishing and evaluating any instrument’s conceptual framework.

Introduction

The 2006 Food and Drug Administration draft guidance on patient-reported outcome (PRO) measures states that one of the first steps in the instrument selection or development process is the identification of the conceptual framework of each instrument [1]. The framework specifies the purpose for each item in terms of the instrument’s measurement goal and specifies how each item is to be used, either as a single-item concept or grouped together to form more complex concepts scored according to the instrument’s measurement structure and scoring system. The instrument can be deemed adequate to support a targeted statement of treatment benefit (i.e., claim) if the instrument measures the claimed concept in a well-defined and reliable way. By recommending the specification of the conceptual framework for each instrument, FDA recognizes the extensive variation that exists among PRO instruments. The tools described here offer a systematic approach to establishing and evaluating any instrument’s conceptual framework.

Instruments used in clinical research studies are known to differ in content depending on their intended application, for example, diagnosis, disease severity, and patient characteristics. These factors, in turn, determine the most relevant concepts for measuring treatment impact. Instruments may also differ according to developers’ perspectives on how to represent PRO concepts and their relationships; for example, researchers trained in medicine, psychology, and economics have developed instruments with different item formats, content, measurement structures, and scoring systems [2–5]. Reviews of compendia of health status and well-being measures present a more complete perspective of the diversity of concepts and measurement structures used in generating scoring systems for measures used in various fields, including pharmacoeconomics, health services research, geriatrics, mental health, and nursing [6–11].

Within the PRO field, researchers, including Fries, Guyatt, and Spilker [12–14], have proposed taxonomies for classifying health-related quality-of-life (HRQoL) concepts; these systems, however, have not as yet been operationalized. Existing classification operational systems, such as the Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition (DSM-IV), International Statistical Classification of Diseases and Related Health Problems (ICD), and International Classification of Functioning, Disability, and Health (ICF), [15–19], illustrate the clustering of concepts and diagnoses and their hierarchical arrangement into concepts of increasing complexity. These, however, have been designed for enumeration and epidemiologic analysis rather than for the type of evaluative decision-making required in the drug approval process. Our review of labeling approved by FDA indicated that PRO instruments of different complexities, from single items of event counts to multiscale, multiconcept instruments have been used to support claims of treatment benefit [20]. Furthermore, this review suggested that it would be possible to link an instrument’s content and measurement structure to the nature of a statement of treatment benefit. That is, there is an interrelationship between the intended claim and the measure that supports it.

The ability to identify and codify this relationship has several advantages to sponsors, regulators, as well as to outcomes researchers more broadly. First, the sponsor and FDA need to understand the complexity of the concept in the desired claim because it will determine the adequacy of the instrument used to...
support that claim. From FDA’s point of view, more complex claims are likely to require more comprehensive instruments that have been demonstrated to capture all the important aspects of the complex concept in the targeted patient population [21]. Second, matching the complexity of the claim to patients’ and physicians’ perspectives of disease burden and impact can be important to the external credibility and effect of the claim. Third, being able to link a PRO instrument explicitly to regulatory or clinical decision-making via the conceptual framework can be both a rewarding and challenging aspect of study design and implementation. Moreover, specification of the relationship between a statement of treatment benefit and the PRO instrument that supports this claim incorporates the need for using standard, well-established psychometric methods to demonstrate properties, such as content and construct validity, as integral components of the decision-making process.

To set forth a systematic method for depicting an instrument’s conceptual framework, this article proposes a “PRO Concept Taxonomy” and a “PRO Instrument Hierarchy.” These two tools endeavor to resolve inconsistency and confusion when conceptualizing and quantifying treatment benefit measured by PRO instruments. The PRO Concept Taxonomy incorporates key terms, including “singular” concept, “compound” concept, and “family” concept; usage of these terms is proposed as a way of adding clarity to the development of an instrument’s conceptual framework. This proposed classification system is generalizable across a wide range of families and concepts.

The PRO Instrument Hierarchy connects the conceptual content of a PRO instrument that has been selected to support the intended claim with the instrument’s measurement structure and scoring system, thereby completing the description of the instrument’s conceptual framework. By linking the claim made with the complexity of the instrument used to support it, we can plan a measurement strategy for future labeling goals.

Methods for Developing the PRO Concept Taxonomy and PRO Instrument Hierarchy

The first step in developing the taxonomy and hierarchy was to evaluate PRO concepts that were identified in our review of the Clinical Studies sections of the labeling for 215 new products approved in the United States from January 1997 through December 2002 [20]; labeling for 64 of these products was found to report at least one PRO. We attempted to identify the actual PRO instrument used to measure the PRO concept and each instrument was evaluated in terms of its conceptual framework to determine the instrument’s relationship to the PRO concept identified. In this article, we use the term “concept” to refer to an aspect of how patients feel or function that is expressed qualitatively; when measured by a PRO instrument, a concept is represented by items and domains.

The second step was to validate the taxonomy and hierarchy by evaluating the labeling for the 142 new products approved by FDA from January 2003 through December 2007; labeling for 36 products reported at least one PRO. The PRO concepts and instruments found in labeling for the 1997–2007 period can be found at: http://www.ispor.org/Publications/value/ ViHsupplementary/ViH12i8_Erickson.asp. The same methods were used for this review as for that of the 1997–2002 labeling. Third, we broadened the scope of our evaluation of PRO instruments to include formal scales beyond those that appeared in the new product labeling using information from the On-Line Guide to Quality-of-Life Assessment (OLGA) [6,22]. OLGA’s comprehensive database includes information on thousands of instruments that are of potential relevance for supporting a claim of treatment benefit. Based on selection criteria designed to identify instruments of diverse conceptual content and measurement structures, the conceptual frameworks of 25 instruments were formally evaluated. This step provided assurance that the taxonomy and hierarchy would be relevant not only to instruments used in previous labeling, but also to those that might appear after 2007.

These evaluations indicated that to fully understand the concept, or concepts, measured by a single instrument or battery of instruments, it is necessary to understand the relationships between the included concepts within the context of their use in the intended claim. For example, a claim of treatment benefit for a new migraine product is commonly stated in terms of five separate symptoms (defined below). Because there is no explicit specification of an interrelationship between them, five symptom-specific instruments are used to provide an implicit, rather than a measured, statement about treatment impact of the more general concept of migraine symptoms.

On the other hand, arthritis-related physical function is frequently expressed in terms of abilities to perform everyday activities, such as basic activities of daily living (ADLs) and instrumental ADLs (IADLs), for example, shopping, managing money, doing heavy housework, and mobility. When the relationships between the general and specific concepts is explicitly recognized, they can be measured using a single instrument, such as the Health Assessment Questionnaire Disability Index (HAQ-DI) [23], and the obtained scores can provide explicit information about treatment impact on both the more general concept as well as the specific abilities.

The PRO Concept Taxonomy

As a result of our evaluation of instruments, we define four nested levels of concepts that represent a practical limitation on the number of levels relevant for making meaningful statements about treatment benefit using PROs, a fifth level we define as concepts that are too basic for supporting meaningful claims (see Fig. 1). Concepts in lower levels of the nested arrangement are more specific than those in the higher levels. Understanding relationships between concepts enables researchers to apply an instrument that is appropriate for the purpose of measurement.

To facilitate a systematic method for depicting a conceptual framework, we define three terms: “family,” “compound concept,” and “singular concept.” A family is a taxonomic category that consists of subcategories, much like species and subspecies in biology. In the PRO context, families can be thought of as higher-level concepts that have subconcepts consisting of compound and singular concepts.

Families may be either generic or specific with respect to disease or condition. Generic families, such as mental, physical, and social function [24–26] are too general for meaningful, product-related discussions and measurement. Specific families, on the other hand, categorize concepts that are related to key diagnostic and therapeutic aspects and, thus, are useful for discussing treatment benefit; each specific family can be placed within a generic family. For example, the specific family of migraine symptoms, which is traditionally defined in terms of nausea, vomiting, pain, phonophobia, and photophobia, is located within the generic family of signs and symptoms.

Each family, whether generic or specific, comprises at least one singular concept that both patients and their health-care decision-makers could consider to be a meaningful goal of treatment benefit, for example, pain intensity. Singular concepts may have low-level singular concepts that are considered to be too basic for use in labeling, for example, ability to cut meat. A
compound concept is defined as consisting of at least two singular concepts; for example, the concept “basic activities of daily living” typically includes bathing, toileting, transferring, and dressing.

The PRO Concept Taxonomy is intended to provide structure to the task of establishing and reviewing a conceptual framework. This task requires identification of the concepts represented by instrument scores, identifying all items that contribute to that score, and diagramming the nesting of concepts within one or more families where appropriate, as illustrated in Figure 2. Singular concepts, and low-level singular concepts, are the most fundamental units in the taxonomy and can be considered as the “building blocks” of compound concepts. A compound concept may be made up of two types of singular concepts: 1) those that include low-level singular concepts, as shown in Family 1; and 2) those that can be measured with one item, as shown in Family 2. The type and number of these singular concepts depends on the disease and its treatment as well as the compound concept that represents the goal of measurement and corresponds to the labeling targets. A statement of benefit may be based on information about a single family or multiple families.
A PRO Concept Taxonomy and Instrument Hierarchy

As shown in this figure, an aggregate is a compound concept that explicitly includes multiple families, for example, HRQoL. A global concept includes one or more families that are implicitly defined and aggregated by the patient, for example, self-rating of health, and is outside the scope of a classification system that is based on clearly identified concepts and their explicit relationships.

The main organizing unit for specifying one or more concepts is the family. Each concept must belong to one family and, conversely, each family can have few or many singular concepts. In fact, a very simple depiction of the PRO Concept Taxonomy can contain one singular concept within a single family in a given application, for example, arthritis-specific pain within the HAQ. More complex, single-family concepts may have low-level singular concepts that are used to form singular concepts. Singular concepts may be used to form compound concepts if the instrument development process provides empiric evidence that the compound concept is defined by the singular concepts.

Procedures for identifying PRO concepts and their relationships are referenced in the FDA draft PRO guidance and documented in other publications [27–31]. These established methods reflect the importance of using both qualitative and quantitative techniques to assure that an instrument provides a suitable measure of the intended measurement goal. Instruments developed using such procedures are most likely to contain items and domains that adequately represent the concepts that are meaningful to both patient and health-care professional, and to incorporate an approach to measurement that creates scores appropriate for the intended use, for example, as clinical trial end points.

Consideration of these PRO Concept Taxonomy principles can assist in depicting an instrument’s conceptual framework. By comparing an instrument’s taxonomic structure with a product’s targeted labeling claims, the adequacy or an instrument can be assessed and researchers can gain insights into the additional instrument development work needed to support those claims. Insight into the complexity of a concept can also be useful when designing studies to support claims related to that concept.

The PRO Instrument Hierarchy

The second step in specifying an instrument’s conceptual framework is to formalize relationships between concepts through the identification of the measurement structure and scoring system and verify this against the measurement goals and the targeted claim. Our review of approved labeling indicated that, regardless of taxonomic structure, instruments could be grouped into nine categories, representing increasing orders of conceptual and measurement complexity. Table 1 shows the nine orders in the hierarchy in terms of their number of families and concepts, and measurement structure, along with examples to illustrate the type of PRO instrument in each order. As indicated in columns 2 and 3, multiple-family instruments may be made up of singular or compound concepts within the individual families. The number and type of families and concepts within an instrument varies depending on the intended use of the instrument. Some instruments with multiple families may also permit the formation of an aggregate concept that may support a claim of “health-related quality of life” (HRQoL) if the included concepts meet the FDA’s HRQoL definition [1]. A multifamily instrument may have a validated measurement structure that permits it to support end points of more than one order, depending on the concepts chosen as study end points (e.g., the 36-Item Short-Form Health Survey [SF-36]; see below).

Order 0 categorizes the simplest type of conceptual framework and Order 8 categorizes the most complex. All PRO instruments, whether generic, disease specific, treatment specific, or global, belong to at least one family and thus can be placed in at least one order in this hierarchy. Each order is also characterized by a measurement structure that indicates the degree to which scores for singular concepts can be combined to form higher-level scores. Thus, each instrument score (or set of scores) becomes a study end point and the concept represented by that score, or set of scores, determines the particular order in the hierarchy that score is assigned. PRO measures that are based on patient reports of frequencies or occurrences of disease- or treatment-related events are classified in Order 0. Instruments that record patients’ evaluative responses, for example, severity or bothersomeness, about symptoms, functions, or perceptions are placed in Orders 1–8.

Measures that assess a frequency count as a singular concept in one family, such as the number of stools observed in the past week, are classified into Order 0 and support very specific statements about treatment effect. Instruments that elicit a patient’s evaluation of a singular concept in one family are classified into Order 1; like instruments in Order 0, these also support very specific statements about treatment effect. The measure of ocular itching in ALAMAST labeling is an example of an Order 1 instrument.

Global item measures are placed in Order 2 as each assesses a compound, rather than a singular, concept. Global item measures provide general information that is difficult to use as the only evidence to support a clinical decision. They are included in the PRO Instrument Hierarchy, however, as they have frequently appeared in labeling, especially those for treatments of rheumatoid arthritis.

PRO measures in Order 3 assess singular concepts within one family measured as a battery. Order 3 instruments differ from those in Order 1 in that the singular concepts are clustered together in labeling in some explicit way, such as in the measurement of “time to symptom improvement” or in the need to “win” simultaneously on a cluster of symptoms. The battery of instruments measuring four migraine symptoms in IMITREX labeling (Table 1) is an example of an Order 3 measure. These measures support symptom-specific statements of treatment benefit and when taken into consideration altogether implicitly demonstrate, rather than explicitly measure, treatment benefit at the family level (e.g., migraine symptoms).

Order 4 measures support statements of treatment benefit based on both the singular concepts and the family, as illustrated by the excerpt from ARAVA labeling in Table 1. Instruments in Order 4 have a measurement structure that provides a profile of scores that allows for meaningful interpretation when comparing scores across domains throughout the duration of treatment. Order 5 measures have four levels within one family and can support statements of treatment benefit at three levels, namely, the singular and compound concept as well as the family levels. Although no instruments of this type were found in our review of approved labeling (see below), we include it for completeness.

Orders 6–8 instruments include two or more families with two or more concepts. Like Order 4 instruments, these instruments also generate profiles of scores that can support measurement of concepts at various levels and offer multiple study design and analysis options. Order 6 instruments, like those in Order 3, measure individual concepts, but unlike Order 3, the concepts in Order 6 instruments have a measurement approach, for example, summed ratings, that allows for comparisons between the family concepts; the SF-36 profile is an example of an Order 6 instrument [32]. Order 7 measures combine multiple singular or compound concepts into families or an aggregate that includes...
Table 1 PRO instrument hierarchy for classifying PRO instruments according to their taxonomic and measurement structures, with examples of PRO instruments and statements of treatment benefit from existing prescription drug labeling

<table>
<thead>
<tr>
<th>Order number</th>
<th>Families*</th>
<th>Concepts*</th>
<th>Taxonomic and measurement structure</th>
<th>Claim(s) supported by instrument</th>
</tr>
</thead>
</table>
| 0            | I         | 1 S       | I or more items in a singular concept that assess frequencies or occurrences that are disease or treatment related | Specific claim about the reported event
|              |           |           | Example: Number of stools per week | Example: “Patients on ZELNORM also experienced an increase in median number of stools from 3.8/week to 6.3/week at month 3...” |
| 1            | I         | 1 S       | I or more items eliciting patient evaluation of either a symptom, function, or perception | Specific claim about the evaluated singular concept
|              |           |           | Example: Ocular itching             | Example: “ALAMAST was significantly more effective than placebo after 28 days in preventing ocular itching associated with allergic conjunctivitis.” |
| 2            | I+        | 1 C       | A global, compound concept measured by a single item | General claim that reflects the content of the item
|              |           |           | Example: Overall rating of the condition of dry mouth now compared with before starting treatment | Example: “Statistically significant global improvement in the symptoms of dry mouth was seen...” (EVOKA) |
| 3            | I         | 2+ S      | Multiple singular concepts representing a cluster of disease-related concepts with one or more measurement approaches that allow for individual concept scores. There is no family score. | Concept-specific claims but no family-level claim. There are as many claimable end points as there are concepts.
|              |           |           | Example: Headache response defined in terms of severity of headache pain. Associated symptoms of nausea, photophobia and phonophobia were also assessed. | Example: “The percentage of patients achieving headache response 2 and 4 hours after treatment was significantly greater among patients receiving IMITREX for patients with migraine-associated nausea, photophobia and/or phonophobia at baseline, there was a lower incidence of these symptoms at 2 hours (Study 1) and at 4 hours (Studies 2, 2.5, and 3)” |
| 4            | I         | 2+ C      | Singular concepts are expressed in 2+ singular concepts with a measurement approach that allows for a compound family score. Concept and family scores are measured using a scoring system that allows direct comparison of concepts. | Both concept-specific, and family-level claims. There are at least three claimable end points. |
|              |           |           | Example: Health Assessment Questionnaire Disability Index (HAQ DI) | Example: “The mean change from baseline in functional ability as measured by the HAQ Disability Index (HAQ DI) in the 6 and 12 month placebo and active controlled trials is shown in Figure 4. ARAVA was statistically superior to placebo in improving physical function. Superiority to placebo was demonstrated consistently across all eight HAQ DI subscales (dressing, arising, eating, walking, hygiene, reach, grip and activities) in both placebo controlled studies.” |
| 5            | I         | 1+ C and 1+ S | Compound concepts each have at least one subconcept with a measurement approach that allows for the calculation of subconcept and concept scores as well as a family score. Both concept and family scores represent compound concepts. | One family, and concept and subconcept claims; there are as many claimable end points as there are end points in the three levels. |
|              |           |           | Example: None found | Example: None found in labeling 1997–2007 |
| 6            | 2+        | 2+ S      | Multiple singular concepts, each of which represents a family, with a measurement approach that allows comparison across concepts. There is no aggregate score. | Concept-level claims. There are as many claimable end points as there are concepts. |
|              |           |           | Example: Walking Impairment Questionnaire (WIQ) | Example: “The walking impairment questionnaire assesses the impact of a therapeutic intervention on walking ability. In a pooled analysis, patients reported improvement in their walking speed and walking distance.” (PLETAL).” |
| 7            | 2+        | 2+ S or C | Concepts are measured in terms of 2+ concepts and 2+ families with a measurement approach that allows calculation of concept and family scores that can be compared. There is an aggregate score that combines more than one family but omits at least one major family needed to support the HRQoL concept. | Family and concept-level claims, with as many claims as there are families and concepts. |
|              |           |           | Example: Asthma Quality of Life Questionnaire (AQLQ—Juniper) | Example: “The subjective impact of asthma on patient’s perception of health was evaluated through use of the AQLQ. Patients receiving ADVAIR DISKUS 100/50 had clinically meaningful improvements in overall asthma-specific quality of life as defined by a difference between groups of at least 0.5 points in change from baseline.” |
| 8†           | 3+        | 3+ C      | Family and concept scores measurement approach that allows comparison across families and concepts. There is an aggregate score that includes all families needed to support the HRQoL concept. | An overall (potentially HRQoL), as well as multiple family and concept claims; there are as many claimable end points in the three levels plus the aggregate score. |
|              |           |           | Example: Sickness Impact Profile (SIP) | Example: “The SIP, a multitem scale in 12 concepts designed to assess the patient’s functioning in multiple areas. Data for the overall SIP score at baseline and change from baseline at 3 months are presented in Table 2. For TASMAR, the change from baseline was statistically significant for the 200 mg tid treatment arm, with a p-value of 0.01.” |

*C: compound concept; HRQoL: health-related quality of life; PRO: patient-reported outcome; S: singular concept.

*Labeling statements are taken from: http://www.accessdata.fda.gov/scripts/cder/drugsatfda/ [46], the 2002 or the 2006 PDR [47].

Any instrument or battery of instruments that provides an overall score without documentation that supports an underlying theoretical model or justification for combining multiple families of concepts should not present the overall score for decision-making. If such a score is used, a caveat about the lack of an appropriate measurement structure should be stated in a footnote.

NOTE: The examples in this table are drawn from the review of new prescription drug labeling approved between 1997 and 2002. These examples illustrate relationships between statements of treatment benefit and the measurement structures of various instruments. They do not, however, provide assurance that the same relationships will be applied to future drug approvals.
more than one family; these instruments can be used to measure both the concepts and the families. In addition, Order 7 instruments may be used to measure the concept represented by the aggregate score.

Order 8 measures are the most “complex,” both conceptually and practically, because they: 1) measure three or more families, including all families needed to support the HRQoL concept as specified in FDA’s draft guidance, i.e., physical, psychological/emotional, and social functioning; 2) have multiple domain scores; and 3) incorporate measurement approaches that support the calculation of an aggregate score. Order 8 instruments can be used to measure singular concepts, family concepts, or aggregate concepts. A conclusion that a treatment impacts HRQoL would be based on an Order 8 instrument.

**Depicting the Conceptual Framework**

The conceptual framework of a battery of instruments proposed for evaluating the benefit of a new migraine treatment, that is, an Order 3 battery of instruments, is illustrated in Figure 3 using the taxonomy and hierarchy. The first step in developing this framework is to identify a set of signs and symptoms related to migraine headache that are recognized by patients and clinicians as being meaningful for defining migraine treatment response. The resulting specific family of migraine symptoms is represented by a cluster of five singular concepts, shown as the taxonomic structure in Figure 3. The dashed lines connecting the singular concepts to the family level indicate that relationships between the individual symptoms and the family, the measurement structure, are implied rather than explicit, that is, the scoring system for the five symptoms does not include a combined symptom score at the family level. In this example, a conclusion concerning a treatment benefit (migraine response) would be based on improvement in every symptom depicted in the conceptual framework.

Figure 4 shows the use of the taxonomy and hierarchy to depict the conceptual framework of the HAQ-DI for supporting labeling claims at both the family and compound concept levels, an instrument in Order 4. As shown in this figure, the HAQ-DI measures a specific family, defined by the eight singular concepts, which are, in turn, composed of low-level singular concepts. The solid lines indicate that the instrument’s measurement structure provides a rationale for combining low-level singular concepts to form explicit statements about patient performance of eight singular concepts as well as the compound concept of physical disability, which is expressed in a single score within the family of arthritis-related physical function.

In developing both the taxonomy and hierarchy, we started with the evaluation of a given instrument according to its content, measurement structure, and scoring system. This process produces a depiction of the conceptual framework as illustrated in Figures 3 and 4. The orders in the PRO Instrument Hierarchy also indicate the type of claim that the instrument can support.

**Evaluating the PRO Instrument Hierarchy Using Recently Approved Labeling**

The explicit relationships between the PRO instrument’s conceptual content, expressed in terms of the PRO Concept Taxonomy, and the treatment benefit statements in labeling, reflected in the PRO Instrument Hierarchy, were evaluated and validated in two separate stages. A previous analysis showed that labeling for 64 (30%) of the 215 new drugs approved from 1997 to 2002 included a treatment benefit statement (in the Clinical Studies section) about a concept measured by a PRO instrument [20]. We first reanalyzed the labeling for these 64 drugs, and classified the conceptual frameworks represented by the PRO statements therein into one of the nine categories described in Table 1. During this first stage, the PRO Instrument Hierarchy was adapted to better fit the actual labeling statements observed. To validate this hierarchy, we then analyzed the labeling of the 142 new drugs approved in 2003–2007 (following the same criteria used in the 1997–2002 study), of which 36 contained PRO-based
statements in their Clinical Studies section, to determine whether those statements and their implied conceptual frameworks mapped well into the hierarchy. This second mapping determined that no changes to the basic structure of the hierarchy were needed, but we felt it was appropriate to modify the description of Order 4, from “There are at least 3 claimable end points” to “This may allow 3 or more claimable end points.”

The percentage of times that each order occurred, for each of the two periods examined, is shown in Figure 5. Percentages add to more than 100% because the labeling for many drugs (38 of 100) contains more than one order of PRO statement. For some orders, the rate of use was similar between periods, in others it was not; some of the variation observed is due to differences in types of drugs approved between periods, as described below.

Simple event counts (Order 0) and singular PRO concepts measured with one or more items (Order 1) were the most commonly occurring orders, present in labeling for 40 and 52 of the 100 drugs, respectively. Some frequently used event counts

*Percentage of drugs with given orders in their labeling, among all approved drugs with PROs. Many drugs have more than 1 order in their labeling.
were cough (immunologic agents), partial seizure frequency (anti-epileptic agents), and use of rescue medications (antimigraine and respiratory agents). Frequently used singular PRO-concepts were: pain intensity, symptom assessments (several areas), ocular itching (ophthalmics), and dyspnea (cardiovascular).

PRO concepts of Order 2 (global concepts), Order 3 (a cluster of singular concepts), and Order 4 (1 family represented by one compound concept containing 2+ singular concepts) were the next most common, appearing in labeling for 25, 18, and 16 different drugs, respectively. Global concepts were most common for anti-inflammatory agents, as a patient global score is part of the American College of Rheumatology 20/50/70 criteria used in rheumatoid arthritis; these accounted for the labeling of 10 drugs out of the 26 with global scores [33,34]. Other statements classified as globals were: time spent in on-off states for Parkinson’s disease (five cases); ability to perform normal activities; and satisfaction with treatment. Interestingly, global items were rarely the only PRO concept in labeling (4 out of 26 cases). Global concepts were less common in labeling approved between 2003 and 2007, primarily due to only one drug for rheumatoid arthritis being approved during that period.

Order 3 PRO measures (which measure a cluster of singular concepts) were most common among gastrointestinal agents and antimigraine products, where different symptom concepts (e.g., phonophobia, photophobia, nausea) are clustered together as a single disease-specific family of concepts (migraine symptoms). Use of Order 3 instruments was much higher in the 1997–2002 than in 2003–2007 due to the approval of migraine drugs in the earlier period, all with Order 3 PRO measures, as opposed to no migraine drugs in the later period. In the earlier period, all but one of the approvals based on Order 4 measures referenced the HAQ Disability Index (or M-HAQ) for anti-inflammatory products [23,33]; the only other Order 4 instrument was the total nasal and non-nasal symptom score for a respiratory product. In the later period, however, there was more varied use of Order 4, including the Erectile Function domain of the International Index of Erectile Dysfunction, the Functional Living Index—Emesis, the Sheehan Disability Scale, and the Alzheimer’s Disease Cooperative Study-Activities of Daily Living Inventory (ADCS-ADL) [36–39].

More complex PRO measures were less common, with no examples of Order 5 instruments occurring in this set of labeling, and a total of 21 examples with Order 6–8 measures. Order 6 measures (2+ families represented by 2+ singular concepts with a profile of scores) included the Walking Impairment Questionnaire (cardiovascular agents), the SF-36 profile of scores (diagnostics), the Quality of Life in Narcolepsy [40], the Toronto Western Spasmodic Torticollis Rating Scale (TWSTRS) [41] (central nervous system agents), and the International Index of Erectile Function [36] (urologic agents). The SF-36 profile of scores was the only Order 6 PRO measure during 2003–2008, used for the lone antiarthritis drug approved in the later period.

Examples of Order 7 PRO measures (2+ families represented by 2+ singular or compound concepts, allowing for family or aggregate scores) also primarily occurred for anti-inflammatory products, based on the SF-36 Physical Component Score (PCS) and Mental Component Score (MCS) scores or the Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC) composite score [42]; one was based on the total score of the Fibromyalgia Impact Questionnaire [43]. The two examples of an Order 8 PRO measure (3+ families, 3+ concepts, with an aggregate score) were the sickness impact profile [44] total score, used for an anti-Parkinsonian product, and a claim for “improved health-related quality of life” for treatment for paroxysmal nocturnal hemoglobinuria, based on results of the European Organization for Research and Treatment of Cancer Quality-of-Life Questionnaire–30 items (EORTC-QLQ-C30) [45].

In reviewing these results across orders and time periods, there has been relatively less frequent use of Orders 2–8 since 2003. Of those drugs approved in 1997–2002 with PROs in their labeling, 59% (38/64) included at least one PRO of Order 2–8, while in 2003–2007, 42% (15/36) included at least one PRO of Order 2–8. Much of this difference is due to the nature of the drugs approved during these periods—in the earlier period, 15 arthritis or migraine drugs were approved, all having these higher-order PROs, while in the latter period, only one arthritis or migraine drug was approved. Not including those drugs in this comparison results in 47% (23/49) labels from 1997–2002 with Order 2–8 PRO’s, and 40% (14/35) in 2003–2007. This analysis indicates that instruments that have commonly been used in the drug approval process fit within the nine orders in the PRO Instrument Hierarchy, based on both an evaluation and a validation labeling sample. This finding provides evidence for the relevance of both the taxonomy and hierarchy in characterizing PRO instruments to be used in clinical trials. Most of the PRO data led to statements of treatment benefit within one family rather than multiple families, with over half being used to make narrow statements of treatment efficacy, that is, based on singular concepts that did not explicitly include a statement of family-level benefit.

Discussion

Specific terminology and the PRO Concept Taxonomy and PRO Instrument Hierarchy are proposed as approaches for more systematically establishing and evaluating conceptual frameworks for PRO instruments used in trials to assess clinical benefit. Beyond providing structures for characterizing PRO measures, they supply outcomes researchers with tools for evaluating and explaining an instrument’s conceptual framework within the context of a specific claim. With improved clarity of this structure, the linkage between the underlying diagnostic or conceptual terminology and the outcome of the health-care intervention becomes stronger and more transparent.

The drug-approval process is unique in that it explicitly links the use of a PRO instrument to medical decision-making through a statement of treatment benefit. The PRO Concept Taxonomy and PRO Instrument Hierarchy are proposed as structures for clarifying this linkage and for locating the use of well-established and relevant psychometric methods within this process. For example, use of these methods to demonstrate an instrument’s content validity within the context of the intended claim is part of the depiction of an instrument’s concept taxonomy. Similarly, depiction of an instrument’s measurement structure is determined by use of well-established quantitative psychometric methods which, in turn, locate the instrument within the PRO Instrument Hierarchy, thereby indicating its suitability for the intended claim.

The review of 1997–2002 new drug labeling illustrated that the PRO Instrument Hierarchy, incorporating the principles of the PRO Concept Taxonomy, is relevant across a wide range of both therapeutic products and the measures chosen to demonstrate their clinical benefit; this finding was confirmed by a subsequent review of 2003–2007 new drug labeling. For example, the predominance of the use of simple PRO instruments—event counts and singular concept PRO instruments (Orders 0 and 1)—along with global items and disease-specific, single-family PRO instruments (Orders 2 and 3) fits with the specific state-
ments about treatment benefit. Aside from the global PRO instruments, which are rarely used in isolation, the connection between the PRO instrument and the disease or its treatment is probably most transparent in these cases and the underlying conceptual framework of the instrument need not be complex.

Use of instruments with multiple concepts was much less common, particularly outside the antiinflammatory area, suggesting that establishing a clear relationship between treatment of a specific disease and broader PRO concepts can be more challenging, both in theory and in practice. Nevertheless, there are sufficient examples of measures with multiple concepts and families to indicate the relevance of the taxonomy and hierarchy and to establish the potential value of measures based on complex concepts. Use of the hierarchy along with the concept taxonomy, beyond simply allowing for a better understanding of the full spectrum of PRO statements allowed in labeling over this 11-year period, should assist in making the determination when to consider and justify the use of more comprehensive measures. Characterizing PRO instruments in a standardized way may improve not only the communication between industry and its regulators but also within the research community more broadly. For example, abstracts of clinical studies frequently use terms such as pain, physical function, and HRQoL to describe measures that may represent any of the orders in the hierarchy. Unless the abstract specifically names the instruments used, the reviewer must locate the article to fully understand both the concepts being measured and the conceptual framework of the instrument in order to interpret the findings. Even within an article, the exact concept(s) measured may be incompletely documented, leading to misinterpretation of findings. More careful attention to the naming of concepts with consideration for the PRO Concept Taxonomy and PRO Instrument Hierarchy will help to clarify the results of clinical studies using PRO instruments.

The work presented here is limited in several ways. First, our approach has been heavily influenced by use of PRO’s in new drug labeling and hence may not be as applicable to other areas using PRO’s. Second, it has been based on retrospective evaluation of instruments and labeling; prospective use may, and is in fact likely to, generate new considerations that could affect the proposed taxonomy and hierarchy. Third, while we have acknowledged the important role of measurement science, especially that of content validity, in the developing a conceptual framework, we have yet to explicitly incorporate this work into our specification of the two tools. And, perhaps most importantly, our approach has not yet been used, to the best of our knowledge, in any interactions between sponsors and regulators, nor has it been explicitly endorsed by any regulatory agency.

Finally, the terminology, taxonomy, and hierarchy described above are proposed as a way of improving clarity and consistency when studies intended to evaluate therapeutic impact are conceived, developed, evaluated, and communicated. It draws both from the existing theoretical literature and from what has been observed in approved labeling and in the regulatory setting. Nevertheless, refinements and extensions to improve the taxonomy and hierarchy to meet future needs are both encouraged and expected. The overriding goal is to better incorporate the most relevant and interpretable PRO measures into drug development, drug labeling, and ultimately, patient care.

The authors would like to acknowledge the helpful comments of three anonymous reviewers as well as those from a number of colleagues who reviewed earlier drafts. The views expressed in this article are those of the authors and do not reflect the positions of either Pfizer Inc. or the Food and Drug Administration.

Source of financial support: None.

References
12 Fries JF. The hierarchy of quality-of-life assessment, the Health Assessment Questionnaire (HAQ), and issues mandating development of a toxicity index. Control Clin Trials 1991;12(Suppl. 4):106S–17S.


31 Leidy NK, Vernon M. Perspectives on patient-reported outcomes: content validity and qualitative research in a changing clinical trial environment. Pharmacoeconomics 2008;26:363–70.


34 Felson DT, Anderson JJ, Lange ML, et al. Should improvement in rheumatoid arthritis clinical trials be defined as fifty percent or seventy percent improvement in core set measures, rather than twenty percent? Arthritis Rheum 1998;41:1564–70.


Psychometric Evaluation of the Diabetes Symptom Checklist-Revised (DSC-R)—A Measure of Symptom Distress

Robert A. Arbuckle, MA,1 Louise Humphrey, MSc,1 Kawiitha Vardeva, MSc,1 Bhakti Arondekar, PhD, MBA,2 Muriel Danten-Viala, MSc,4 Jane A. Scott, PhD,1 Frank J. Snoek, PhD5

1Mapi Values, Bollington, UK; 2Amgen Ltd, Cambridge, UK; 3GlaxoSmithKline, Philadelphia, PA, USA; 4Mapi Values, Lyon, France; 5VU University Medical Center, Amsterdam, The Netherlands

ABSTRACT

Objective: To assess the psychometric validity, reliability, responsiveness, and minimal important differences of the Diabetes Symptoms Checklist-Revised (DSC-R), a widely used patient-reported outcome measure of diabetes symptom distress.

Research Design and Methods: Psychometric validity of the DSC-R was assessed using blinded data from a large-scale trial of approximately 4000 type 2 diabetes patients. Confirmatory factorial analysis (CFA) and multitrait analysis were used to examine the construct validity of the structure of DSC-R. DSC-R internal consistency, discriminative validity, and responsiveness were also assessed. Distribution and anchor-based methods were used to estimate minimal important differences for DSC-R domains.

Results: Mean age of the sample was 56 years, 42% were female, 88% were Caucasian. Patients had a mean body mass index (BMI) of 32.2 and mean glucose-fasting level of 151.7 mg/dl. CFA and multitrait analysis indicated that the wording of the DSC-R has acceptable construct validity. Item-scale correlations ranged from 0.44 to 0.78. Cronbach’s alpha coefficients ranged from 0.69 to 0.87. At baseline, DSC-R scores were higher among patients with higher BMI scores (P < 0.0001), supporting the discriminative validity of the DSC-R. Minimal important difference estimates ranged from 0.39 to 0.60 points when using distribution methods and from 0.00 to 0.33 when estimated using anchor-based methods.

Conclusions: The DSC-R demonstrated excellent psychometric properties when tested in a large-scale diabetes clinical trial. Responsiveness and test–retest reliability of the DSC-R warrant further evaluation.

Keywords: diabetes, patient reported outcome, psychometric validation, quality of life, symptom distress.

Introduction

There is general consensus that quality of life is an important outcome of diabetes care, requiring reliable patient-reported measures, pertaining to the physical, social, and psychological domains [1]. The Diabetes Symptoms Checklist (DSC) was developed to capture the subjective experience of diabetes-related symptoms and changes therein as a result of medical treatment [2]. Symptoms associated with type 2 diabetes may be directly related to hyperglycemia (e.g., excessive thirst, dryness of the mouth, fatigue, frequent urination), complications associated with diabetes (e.g., loss of sensation in the extremities), and the treatment of diabetes (e.g., hypoglycemia) [3].

The DSC items were derived from a review of the literature and discussions with experienced physicians in the field [2]. Guided by their clinical knowledge, the developers identified eight domains of importance for diabetes symptom distress. A final selection of 34 items measuring these eight domains was made based on psychometric criteria. Based on research findings [4–6], the developers of the instrument later sought to improve the DSC in two ways. First, the frequency scale was replaced by a dichotomous “yes” or “no” response for the presence or absence of each symptom. This change was made because at times patients found the dual response format confusing to answer and missing data could be a problem. Moreover, reported frequency and burden were generally highly intercorrelated (>0.80), suggesting redundancy. The second change was to change the scaling from a 4-point to a 5-point Likert scale. No changes to item content were made. The resulting instrument is known as the DSC-Revised (DSC-R).

Validity and reliability for the DSC-R have not been reported in the literature, although recent studies would suggest DSC-R has satisfactory reliability in both newly diagnosed and insulin-treated type 2 diabetes patients [7–11]. To address the need for evidence of psychometric validity, blinded, post hoc analysis of data from a large scale, multicenter, randomized controlled clinical trial was performed to assess the psychometric properties of DSC-R.

Subjects and Methods

Post hoc psychometric validation analyses were performed using blinded data from a Diabetes Outcome Progression Trial (ADOPT) [12]. The ADOPT study was a double-blind, randomized, parallel group study comparing rosiglitazone, metformin, and glyburide as an initial treatment for 4360 recently diagnosed patients with type 2 diabetes. Details of the study design are reported elsewhere [12]. The sample included patients from the United States (n = 1644), Canada (n = 612), France (n = 388), Germany (n = 466), Spain (n = 397), UK (n = 313), and other countries (n = 466). The primary outcome was the time to monotherapy failure defined as fasting plasma glucose of more than 180 mg/dl on consecutive testing after at least 6 weeks of treatment at the maximum tolerated dose of study medication. Cross-sectional psychometric validation analyses (all analyses except responsiveness and minimal important differences [MID]) were performed on 4286 patients who completed the DSC-R at baseline. The 3594 patients who completed the DSC-R and the Short-Form 36 (SF-36) at 1-year follow-up were included in the longitudinal analyses.
Psychometric Validity of the DSC-R

Questionnaires

**Diabetes Symptom Checklist-Revised**

The 34 items of the DSC-R are grouped into eight symptom clusters or domains, each measuring a different aspect of diabetes symptomatology: hyperglycemic, hypoglycemic, psychological-cognitive, psychological-fatigue, cardiovascular, neurological-pain, neurological-sensory, and ophthalmologic. A conceptual framework detailing the items included in each domain is presented in Figure 1. For each item, participants are asked if they have experienced the symptom in the past 4 weeks, and if yes, how troublesome that particular symptom is for them. Two example items demonstrating the format of the questionnaire are provided in Figure 2. Items are summed to form domain scores and all items of the DSC-R can be summed together to form a total score. Higher scores indicate greater symptom burden.

**Short-Form 36**

The SF-36 is a 36-item measure of perceived health status with established validity in both healthy subjects and somatic patients [13–16]. A total score can be calculated as well as a Physical Component and Mental Component summary score. Domain scores range from 0 to 100; higher scores indicate better health status. The recall period is the past 4 weeks.

**Clinical Measures**

Body mass index (BMI) and HbA1c levels were recorded at both study visits reported here.

**Analysis**

All analyses were performed post hoc on blinded data from the ADOPT clinical trial. With the exception of the analyses of responsiveness and MID, all analyses were performed on baseline cross-sectional data. Responsiveness and MID analyses were assessed using baseline and 1 year (visit 9) data.

Confirmatory factorial analysis (CFA) was used to evaluate the overall structure of the DSC-R, as shown in Figure 1. The goodness of fit of the model was assessed based on the Goodness of Fit Index (GFI) (good fit if >0.90), the Root Mean Square Residual (RMR) (good fit if <0.05), the Comparative Fit Index (CFI) (good fit if >0.95), and the Root Mean Square Error of Approximation (RMSEA) fit indices (good fit if <0.05) [17].

Multitrait analysis was performed to examine the validity of the DSC-R item-scale structure. This included tests of item convergent validity (the correlation between each item and its own scale should be ≥0.40), [18] and internal consistency reliability (alpha coefficients ≥0.70 for acceptable internal consistency) [19]. Percentages scoring at floor and ceiling (lowest and highest possible scores) were also examined to check for the presence of floor or ceiling effects, and the DSC-R interscale correlations were examined.

Concurrent validity (also known as convergent/divergent validity) involves analyzing correlation levels between the scales of the studied questionnaire and scales of well-established and validated questionnaires measuring similar concepts [18]. Scales measuring similar concepts are expected to correlate more highly than scales measuring unrelated concepts. Concurrent validity was assessed by examining Spearman’s correlations between the DSC-R scores and the SF-36 scores.

The known groups validity of the DSC-R (ability to distinguish among groups of patients that would be expected to differ [18]) was evaluated by examining differences in DSC-R scores according to HbA1c levels (<6%; 6–6.9%; 7–7.9%; ≥8%) and BMI levels (<25, 25–29.9, 30–39.9, ≥40). It was hypothesized that there would be statistically significant differences among the groups compared (P < 0.05), as assessed using analysis of variance (ANOVA).

Responsiveness was evaluated by comparing changes in DSC-R scores from baseline to 1 year later among subgroups of patients defined by the change score on SF-36 item 2 “compared to one year ago, how would you rate your health in general now?” Only patients who completed the DSC-R at both visits were included in this analysis. Effect sizes were calculated. The effect size is calculated as the mean difference (change score) in scores from baseline to 1 year later divided by the standard deviation of the baseline score. Effect sizes of 0.2, 0.5, and 0.8 were considered small, moderate, and large, respectively, as defined by Cohen [20]. Effect sizes were expected to be moderate or large (≥0.50) in the expected direction for patients who
reported their health to have improved or worsened, and small or negligible (<0.25) for patients who reported having experienced no change in their health.

An MID for a patient reported outcome (PRO) has been defined as the “smallest difference in score in the domain of interest which patients perceive as beneficial, and which would mandate, in the absence of troublesome side effects and excessive cost, a change in patient’s management” [21]. It is recommended that more than one method of estimating MID is employed—thus in the present study both distribution-based and anchor-based methods were used. First, an effect size of 0.5 (a change of 0.5 of a standard deviation) was considered to be clinically significant (distribution method) [22]. Second, the MID was considered to be the change in DSC-R scores for patients who considered their health to be “somewhat better” or “somewhat worse” in response to the SF-36 item 2 “compared to one year ago, how would you rate your health in general now?” (anchor-based method).

For all analyses, the threshold for significance was \( P < 0.05 \). Where correlations were evaluated, Pearson’s correlation coefficient was used. ANOVA was used in the comparison of groups. All data processing and analyses were performed using SAS software (Statistical Analysis System, Version 9, SAS Institute, Cary, NC).

**Results**

Cross-sectional analyses were performed on 4286 randomized patients. Mean age for the overall population was 56 years. Forty-two percent of the sample was female; 88% were Caucasian, 2% were Asian, 4% were black, 5% were Hispanic, and 1% other. Mean BMI was 32.2; mean waist-hip ratio was 0.95; mean number of years with diabetes was 0.8 years; mean HbA1c was 7.36, and the mean glucose-fasting level was 151.7 mg/dl.

The number of missing data was very low for all items, with less than 3% of missing data for all items, at all study visits. At
The percentage of missing data for the DSC-R items ranged from 0.37% (n = 16) missing for item 15 “Pain in calves at night” to 1.63% missing for item 23 “Frequent voiding.” Thus, there was considered to be an excellent quality of completion.

The model fit indices for the CFA all narrowly failed to meet the criteria for acceptability: the GFI was 0.9022, the RMR was 0.0522, the CFI was 0.9029, and the RMSEA was 0.055 (90% lower confidence limit: 0.0538). Standardized factor loadings ranged from 0.50 to 0.85—all were well above the 0.40 threshold used as a rule of thumb for acceptable factor loadings. Scaling test results are summarized in Table 1. Item-domain correlations ranged from 0.44–0.78; thus all items met the test of item convergent validity (a correlation of >0.40 between the item and the other items in its domain).

Almost all of the scales surpassed the 0.70 threshold for acceptable internal consistency reliability (Cronbach’s alpha range: 0.69–0.87). “Cardiovascular” was the single exception, with an alpha coefficient of 0.69, just below the alpha threshold. Across all of the domains there was a high floor effect (for each domain the percentages of patients with lowest possible score ranged from 27.8% to 60.2%), reflecting the fact that many patients gave the response “Did not occur” for all the items in those domains. Zero-order interdomain correlations ranged from $r = 0.39$ (between Neuropathic Sensory and Hyperglycemic domains) to $r = 0.71$ (between Psychological Fatigue and Psychological Cognitive domains). Neuropathic Sensory and Neuropathic Pain domains showed a zero-order correlation of 0.62. Thus, we believe that the domains are related but not redundant.

Correlations between the DSC-R scale scores and the SF-36 scale scores ranged from $-0.22$ to $-0.69$ and were mostly small to moderate (Table 2). Scales measuring concepts that would be expected to be related correlated the most highly. For example, the highest correlation was between the DSC-R Psychological Fatigue domain and the SF-36 Vitality scale ($r = -0.69$), both of which measure fatigue/vitality. Similarly, the DSC-R Neuropathic Pain domain correlated most highly with the SF-36 Bodily Pain scale ($r = -0.38$) and the DSC-R Hypoglycemic domain (which includes items asking about mood and irritability) correlated most highly with the SF-36 Mental Health scale ($r = -0.45$) and the Mental Component Scale ($r = -0.45$).

Known groups validity results are summarized in Table 3. There were statistically significant differences in all DSC-R scale scores among the four groups determined by HbA1c level ($P < 0.005$). For all DSC-R scales among the three groups <6%, 6% to 6.9% and 7% to 7.9%, there was a clear linear progression of higher DSC-R scores for higher HbA1c levels. However, with the exceptions of the Neuropathic pain and Ophthalmologic scales, the ≥8% group did not continue this progression—for all other DSC-R scales scores for this group were slightly lower than for the 7% to 7.9% group. For all DSC-R scales, there was a linear pattern of higher DSC-R scores (indicating greater symptom bother) for patients with higher BMI, with statistically significant differences among the groups ($P < 0.0001$).

Responsiveness results are summarized in Table 4. Analysis of changes in DSC-R scale scores from baseline to year one revealed the expected pattern, but changes were very small for all groups. Patients who believed they had “much worse health” compared to baseline had worse scores on the DSC-R scales compared to baseline (effect sizes ranged from 0.18 to 0.51). Patients who reported their health was “about the same” had negligible changes in their DSC-R scores between baseline and 1-year follow up (effect sizes ranged from −0.12 to 0.10), and there were small or negligible improvements in DSC-R scores in patients with better health at baseline compared to those whose health remained unchanged or improved (effect sizes ranged from −0.12 to 0.10)
who believed their health was much better (effect sizes ranged from −0.32 to −0.08). This pattern was observed across all scales, except the “Hyperglycemic” scale.

MID were estimated to range from 0.39 to 0.60 points across domains when estimated using distribution methods (0.5 of a standard deviation) and from 0.00 to 0.33 points when estimated using anchor-based methods (change for patients who reported that they were “somewhat better” or “somewhat worse” on the SF-36 item 2). Therefore, a conservative approach would be to adopt the distribution-based MID for all domains.

**Conclusions**

The ADOPT study provided an excellent opportunity to examine the psychometric properties of the DSC-R in a large sample of type 2 diabetes patients. Analysis of data gathered from this study confirmed that the DSC-R has good validity and reliability. Construct validity testing through CFA and multitrait analyses confirmed the validity of the DSC-R item-domain structure, with all items loading on and correlating with the domain in which they were included. A priori criteria for “goodness of fit” were not quite met, but in all cases the criteria were only narrowly missed. Furthermore, it could be argued that different symptom clusters should not necessarily be expected to be (strongly) related. For example, a patient experiencing ophthalmologic symptoms (and associated distress) would not necessarily be expected to be experiencing cardiovascular symptoms. The Cardiovascular domain was the only domain that failed (narrowly) to meet the criterion for internal consistency, but this is unsurprising given that the scale includes items relating to shortness of breath and pain/palpitations in the heart, which would not necessarily be expected to be closely related.

Both correlations among the factors of the CFA (shown in Fig. 2) and zero-order Pearson’s correlations between the domains (paragraph 4 of the results) were conducted. For the correlations between the “psychological fatigue” and “psychological cognitive” domains, there were considerable differences between the zero-order or “raw” correlations between the domains (0.71), and the “adjusted” interdomain correlation in the CFA model (0.33). This difference is most likely explained by the fact that these two factors are highly related to the Total score (factor loading with Total score = 0.81 for “psychological fatigue” and 0.84 for “psychological cognitive”). As the Total score explains a large part of the variance between these two factors, it only leaves a “residual” correlation of only 0.33 between the “psychological” domains in the model. By contrast, the raw correlation between the “neuropathic pain” and “neuropathic sensoric” domains is close to the correlation given by the model (0.62 and 0.60 respectively), and the loading of these “neuropathic” domains on the Total score are lower than those of the “psychological” domains (0.68 and 0.66 vs. 0.81 and 0.84). This shows that the two “neuropathic” domains may have a strong “residual” correlation independent from the Total score.

The relatively high correlation between the two “neuropathic” domains could indicate the presence of a higher-order factor. However, it was deliberately decided not to evaluate this as part of the model, because the objective was to evaluate the fit of the original model of the DSC-R and not to try to find the best structure. Furthermore, it also could be argued that the two domains should be left separate as they are distinct in terms of the item content: one focuses on neuropathic pain while the other on neuropathic sensoric. As the model fit was acceptable with this specification and the other psychometric results of the two separate domains were good, it was decided to keep these two distinct (but correlated) domains separate. Future validation of the DSC-R could further explore whether these domains are best kept separate or if they should be combined into a single higher-order domain.

An examination of correlations between DSC-R scores and SF-36 scale scores suggested that the DSC-R has acceptable concurrent validity. Domains that one would expect to be related correlated more highly than domains measuring less closely related concepts. Correlations were generally small or moderate, but this is unsurprising given that the DSC-R is a disease-specific measure of symptom distress and the SF-36 is a generic measure

---

**Table 2**  Concurrent validity: Correlation coefficients between DSC-R scores and SF-36 scale scores (N = 4286)

<table>
<thead>
<tr>
<th>SF-36 scales</th>
<th>Bodily pain</th>
<th>General health</th>
<th>Mental health</th>
<th>Physical functioning</th>
<th>Role emotional</th>
<th>Role physical</th>
<th>Social functioning</th>
<th>Vitality</th>
<th>Mental component summary</th>
<th>Physical component summary</th>
</tr>
</thead>
<tbody>
<tr>
<td>DSC-R domains</td>
<td>Psychological—fatigue</td>
<td>−0.43*</td>
<td>−0.46</td>
<td>−0.45</td>
<td>−0.41</td>
<td>−0.41</td>
<td>−0.45</td>
<td>−0.46</td>
<td>−0.69</td>
<td>−0.48</td>
</tr>
<tr>
<td></td>
<td>Psychological—cognitive</td>
<td>−0.37</td>
<td>−0.37</td>
<td>−0.39</td>
<td>−0.32</td>
<td>−0.38</td>
<td>−0.37</td>
<td>−0.42</td>
<td>−0.52</td>
<td>−0.42</td>
</tr>
<tr>
<td></td>
<td>Neuropathic pain</td>
<td>−0.38</td>
<td>−0.28</td>
<td>−0.25</td>
<td>−0.34</td>
<td>−0.23</td>
<td>−0.30</td>
<td>−0.32</td>
<td>−0.33</td>
<td>−0.22</td>
</tr>
<tr>
<td></td>
<td>Neuropathic sensory</td>
<td>−0.37</td>
<td>−0.25</td>
<td>−0.23</td>
<td>−0.28</td>
<td>−0.26</td>
<td>−0.28</td>
<td>−0.29</td>
<td>−0.30</td>
<td>−0.22</td>
</tr>
<tr>
<td></td>
<td>Cardiovascular</td>
<td>−0.37</td>
<td>−0.34</td>
<td>−0.30</td>
<td>−0.41</td>
<td>−0.29</td>
<td>−0.32</td>
<td>−0.35</td>
<td>−0.44</td>
<td>−0.29</td>
</tr>
<tr>
<td></td>
<td>Ophthalmologic</td>
<td>−0.27</td>
<td>−0.23</td>
<td>−0.23</td>
<td>−0.23</td>
<td>−0.25</td>
<td>−0.25</td>
<td>−0.27</td>
<td>−0.29</td>
<td>−0.24</td>
</tr>
<tr>
<td></td>
<td>Hypoglycemic</td>
<td>−0.31</td>
<td>−0.32</td>
<td>−0.45</td>
<td>−0.25</td>
<td>−0.35</td>
<td>−0.30</td>
<td>−0.42</td>
<td>−0.42</td>
<td>−0.45</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>−0.47</td>
<td>−0.45</td>
<td>−0.46</td>
<td>−0.43</td>
<td>−0.41</td>
<td>−0.44</td>
<td>−0.49</td>
<td>−0.61</td>
<td>−0.47</td>
</tr>
</tbody>
</table>

* Spearman’s correlation, P < 0.0001 for all correlations examined

† Number of patients

DSC-R, Diabetes Symptoms Checklist-Revised; SF-36, Short-Form 36.
Table 3  Known groups validity of Dsc–R total score and by domain

<table>
<thead>
<tr>
<th>Known groups validity</th>
<th>Groups</th>
<th>Psychological–fatigue Mean (SD)</th>
<th>Psychological–cognitive Mean (SD)</th>
<th>Neuropathic pain Mean (SD)</th>
<th>Neuropathic sensory Mean (SD)</th>
<th>Cardiovascular Mean (SD)</th>
<th>Ophthalmologic Mean (SD)</th>
<th>Hypoglycemic Mean (SD)</th>
<th>Hypertainic Mean (SD)</th>
<th>Total score Mean (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BMI</td>
<td>&lt;25</td>
<td>1.04 (1.11)</td>
<td>0.69 (0.91)</td>
<td>0.37 (0.69)</td>
<td>0.45 (0.76)</td>
<td>0.41 (0.65)</td>
<td>0.50 (0.79)</td>
<td>0.59 (0.89)</td>
<td>0.96 (1.04)</td>
<td>0.63 (0.64)</td>
</tr>
<tr>
<td></td>
<td>25–29.9</td>
<td>1.05 (1.12)</td>
<td>0.70 (0.87)</td>
<td>0.38 (0.70)</td>
<td>0.47 (0.75)</td>
<td>0.50 (0.71)</td>
<td>0.50 (0.77)</td>
<td>0.63 (0.91)</td>
<td>1.02 (1.06)</td>
<td>0.65 (0.63)</td>
</tr>
<tr>
<td></td>
<td>&gt;30</td>
<td>1.27 (1.21)</td>
<td>0.87 (0.98)</td>
<td>0.49 (0.78)</td>
<td>0.60 (0.82)</td>
<td>0.67 (0.79)</td>
<td>0.61 (0.87)</td>
<td>0.80 (1.00)</td>
<td>1.24 (1.15)</td>
<td>0.82 (0.71)</td>
</tr>
<tr>
<td></td>
<td>30–39.9</td>
<td>1.62 (1.34)</td>
<td>1.04 (1.04)</td>
<td>0.67 (0.91)</td>
<td>0.85 (1.00)</td>
<td>0.92 (0.90)</td>
<td>0.73 (0.92)</td>
<td>1.04 (1.16)</td>
<td>1.59 (1.30)</td>
<td>1.06 (0.81)</td>
</tr>
<tr>
<td></td>
<td>n=332</td>
<td>n=335</td>
<td>n=338</td>
<td>n=337</td>
<td>n=338</td>
<td>n=338</td>
<td>n=337</td>
<td>n=336</td>
<td>n=338</td>
<td>n=330</td>
</tr>
<tr>
<td></td>
<td>n=1428</td>
<td>n=1429</td>
<td>n=1435</td>
<td>n=1433</td>
<td>n=1435</td>
<td>n=1435</td>
<td>n=1433</td>
<td>n=1430</td>
<td>n=1415</td>
<td>n=1400</td>
</tr>
<tr>
<td></td>
<td>n=2039</td>
<td>n=2036</td>
<td>n=2042</td>
<td>n=2040</td>
<td>n=2040</td>
<td>n=2045</td>
<td>n=2049</td>
<td>n=2032</td>
<td>n=2023</td>
<td>n=2032</td>
</tr>
<tr>
<td></td>
<td>n=451</td>
<td>n=453</td>
<td>n=454</td>
<td>n=454</td>
<td>n=455</td>
<td>n=453</td>
<td>n=453</td>
<td>n=451</td>
<td>n=448</td>
<td>n=448</td>
</tr>
</tbody>
</table>

P-value (ANOVA) <0.0001 <0.0001 <0.0001 <0.0001 <0.0001 <0.0001 <0.0001 <0.0001 <0.0001 <0.0001

ANOVA, analysis of variance; BMI, body mass index; Dsc–R, Diabetes Symptoms Checklist-Revised.

Table 4  Responsiveness of Dsc–R total and subscale scores

<table>
<thead>
<tr>
<th>Change* groups</th>
<th>Psychological–fatigue Effect size†</th>
<th>Psychological–cognitive Effect size†</th>
<th>Neuropathic pain Effect size†</th>
<th>Neuropathic sensory Effect size†</th>
<th>Cardiovascular Effect size†</th>
<th>Ophthalmologic Effect size†</th>
<th>Hypoglycemic Effect size†</th>
<th>Hypertainic Effect size†</th>
<th>Total score Effect size†</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;Much better&quot;</td>
<td>-0.21</td>
<td>-0.13</td>
<td>-0.10</td>
<td>-0.08</td>
<td>-0.18</td>
<td>-0.16</td>
<td>-0.13</td>
<td>-0.32</td>
<td>-0.22</td>
</tr>
<tr>
<td>n=535</td>
<td>n=543</td>
<td>n=539</td>
<td>n=541</td>
<td>n=542</td>
<td>n=540</td>
<td>n=543</td>
<td>n=538</td>
<td>n=526</td>
<td></td>
</tr>
<tr>
<td>&quot;Somewhat better&quot;</td>
<td>-0.10</td>
<td>-0.08</td>
<td>-0.02</td>
<td>-0.00</td>
<td>-0.08</td>
<td>-0.07</td>
<td>-0.06</td>
<td>-0.20</td>
<td>-0.11</td>
</tr>
<tr>
<td>n=1001</td>
<td>n=1013</td>
<td>n=1013</td>
<td>n=1010</td>
<td>n=1010</td>
<td>n=1013</td>
<td>n=1013</td>
<td>n=1015</td>
<td>n=986</td>
<td></td>
</tr>
<tr>
<td>&quot;About the same&quot;</td>
<td>0.03</td>
<td>-0.01</td>
<td>0.10</td>
<td>0.06</td>
<td>0.06</td>
<td>-0.01</td>
<td>0.03</td>
<td>-0.12</td>
<td>0.02</td>
</tr>
<tr>
<td>n=1664</td>
<td>n=1683</td>
<td>n=1678</td>
<td>n=1680</td>
<td>n=1681</td>
<td>n=1681</td>
<td>n=1682</td>
<td>n=1686</td>
<td>n=1645</td>
<td></td>
</tr>
<tr>
<td>&quot;Somewhat worse&quot;</td>
<td>0.33</td>
<td>0.22</td>
<td>0.21</td>
<td>0.21</td>
<td>0.29</td>
<td>0.06</td>
<td>0.22</td>
<td>-0.07</td>
<td>0.25</td>
</tr>
<tr>
<td>n=211</td>
<td>n=215</td>
<td>n=214</td>
<td>n=216</td>
<td>n=216</td>
<td>n=214</td>
<td>n=215</td>
<td>n=214</td>
<td>n=209</td>
<td></td>
</tr>
<tr>
<td>&quot;Much worse&quot;</td>
<td>0.51</td>
<td>0.25</td>
<td>0.19</td>
<td>0.19</td>
<td>0.40</td>
<td>0.39</td>
<td>0.24</td>
<td>0.24</td>
<td>0.19</td>
</tr>
<tr>
<td>n=19</td>
<td>n=20</td>
<td>n=20</td>
<td>n=20</td>
<td>n=20</td>
<td>n=20</td>
<td>n=20</td>
<td>n=20</td>
<td>n=20</td>
<td></td>
</tr>
</tbody>
</table>

*Defined by responses to SF–36 item 2 "Compared to one year ago, how would you rate your health in general now?"  
†Mean Dsc–R change score between baseline and year by the standard deviation at baseline.
of health status. Indeed, the relatively small correlations provide support for the value of a disease-specific measure such as the DSC-R.

Analysis of DSC-R scores according to differences in clinical variables supported the discriminative or “known groups” validity of the instrument. DSC-R scores were higher among patients with higher BMI levels. Interestingly, a recent study in non-diabetic obese and nonobese participants also found significant higher DSC-R scores in the obese subjects, further confirming the discriminative validity of the measure [23]. The finding that a linear relationship between DSC-R scores and HbA1c levels was not observed above HbA1c values of 7% to 8% is intriguing and deserves further investigation.

MID was estimated using both anchor and distribution methods, with anchor-based estimate resulting in very much smaller MIDs than the distribution-based estimate. The applicability of the anchor-based results may be limited by the appropriateness of the anchors used. Ideally, an anchor more closely related to diabetes symptom-distress would have been used. In the absence of such an anchor, the SF-36 change item was used; however, the results should be interpreted with caution due to the questionable appropriateness of this anchor. A conservative approach would be to accept the MID estimated based on the distribution approach of 0.5 of the baseline standard deviation. This is more conservative because this distribution approach resulted in the largest MID and would therefore require a greater change in score in order to conclude a meaningful difference had been achieved.

The study has a number of limitations that need to be mentioned. First, the study was performed in recently diagnosed participants (less than 3 years) who had still few diabetes-related symptoms. This was reflected in the high percentages of patients who scored at floor. Despite the changes being very small, there were still clear differences in DSC-R change scores between subjects who reported changes in their health and those who reported their health to be unchanged. Further evaluation of changes over time in a sample of patients who have more severe diabetes symptoms is recommended. Second, because the psychometric validation was performed post hoc using clinical trial data, test–retest reliability could not be assessed and should be evaluated in future studies. Third, due to the post hoc nature of the study, we were constrained in the other PRO and clinical measures available for validation of the DSC-R. In particular, the criterion used to define change groups in the responsiveness analysis was less than optimal—further testing of responsiveness using a change criterion which is focused on changes in diabetes symptoms (as opposed to general health) is recommended as a priority.

In summary, the evidence reported here suggests that the DSC-R has acceptable reliability, validity, and sensitivity to changes over time, thus making it a suitable measure of diabetes symptom burden for use in clinical research involving patients with type 2 diabetes.

Acknowledgments

The authors would like to acknowledge the contribution of the ADOP'T steering committee in designing the ADOP'T trial and collecting the data as well as commenting on a draft of the article.

Source of financial support: This article was supported by a grant from GlaxoSmithKline Pharmaceuticals. Competing interests: KV was and BA is an employee of GlaxoSmithKline. LH, RA, JS, and MV are all employees of Mapi Values and were contracted by GlaxoSmithKline to conduct the study. FS has received financial compensation from GSK to consult on this validation project.

Authors’ Contributions

RA and JS designed the study, interpreted the results, and participated in the writing of the article. LH wrote the first draft of the article. MV wrote the first draft of the analysis plan, performed the psychometric validation analysis, and reviewed the article. KV and BA were involved in the design of the study, reviewed the analysis plan, interpreted the results, and reviewed the article. FS reviewed the analysis plan, interpreted the results, and reviewed and participated in the writing of the article.

References

Patient Experiences of Continuity of Cancer Care: Development of a New Medical Care Questionnaire (MCQ) for Oncology Outpatients

Clare Harley, PhD,1 Jacqui Adams, PhD,1 Laura Booth, MA,1 Peter Selby, MRCP, MD,1 Julia Brown, MSc,2 Galina Velikova, BMBS (MD), PhD1

1Cancer Research UK Clinical Centre Leeds, Cancer Medicine Research Unit, St James’s University Hospital, Leeds, UK; 2Clinical Trials and Research Unit, University of Leeds, Leeds, UK

ABSTRACT

Objectives: To adapt the Components of Primary Care Index (CPCI) to be applicable to oncology outpatients and to assess the reliability and validity of the adapted instrument (renamed the Medical Care Questionnaire [MCQ]).

Methods: The development and validation of the MCQ took place in four phases. Phase 1 reviewed the literature and examined existing measures. In Phase 2, the selected instrument (CPCI) was reviewed by a panel of experts using a stepwise consensus procedure. In Phase 3, the adapted 21-item MCQ was administered to 200 outpatients attending oncology appointments. The instrument was refined to 15 items and in Phase 4, it was completed by 477 oncology outpatients. The psychometric properties of the new instrument were assessed using exploratory factor analysis (EFA), confirmatory factor analysis, multitrait scaling analysis, and by comparing MCQ scores between known groups.

Results: EFA of the 15-item MCQ suggested three subscales with acceptable to good reliability: “Communication” α = 0.69; “Coordination” α = 0.84; and “Preferences” α = 0.75. Comparing known groups showed that patients who saw fewer doctors during their clinic visits reported stronger “Preferences” to see their usual doctor and rated “Communication” with their doctors as better than patients who saw more doctors during their clinic visits.

Conclusion: The MCQ demonstrates good psychometric properties in the target population. It is a brief and simple-to-use instrument, which provides a valid perspective on patients’ experiences of communicating with doctors and their perceptions of the continuity and coordination of their cancer care.

Keywords: cancer, communication, patient-reported outcomes, psychometric properties, questionnaire development.

Introduction

In the late 1990s, hospital-based oncology practices began to change with the development of new and effective systemic cancer treatments. The delivery of cancer care became more complex with increasing number of patients surviving for longer and increasing number of oncologists and nurses being involved in the care delivery. Multidisciplinary teams were formed to ensure involvement of the necessary experts in diagnosis, treatment modalities, and patient care, so that all patients received consistently high quality and timely treatment. Such multidisciplinary and team-based structures are common within UK hospitals for the delivery of a variety of medical interventions. Nevertheless, the involvement of a large number of medical staff for each patient can have a negative impact on the continuity of care that patients receive if medical staff vary in their ability to elicit important symptoms or functional limitations, to assess change over time, or to make an objective medical record of problems [1–3]. Continuity of care is an important issue for modern health service provision, yet assessing continuity is not always straightforward, in part because it has been a difficult subject to define.

Early definitions described good continuity of care quantitatively as a succession of visits by a patient to the same health-care provider [4]. More recent definitions have made attempts to evaluate continuity of care within the context of a multidisciplinary and multiservice health system. As part of a National Health Service (NHS) scoping exercise, Freeman et al. [5] identified three aspects of health care that were considered important to continuity of care: seeing the same health-care provider over time; having continuity when care is shared or transferred between health-care providers; and having continuity of information across medical records and providers. Continuity of care is expected to have an impact on the quality of care that patients receive and may improve patient outcomes. For example, higher experienced continuity may be associated with lower health-care needs in the future [6]. Nevertheless, it can be difficult to elicit reliable self-reports of patients’ perceptions of the care they receive. For example, in oncology, patients tend to report high levels of satisfaction with their care and appear reluctant to rate their medical team negatively [7]. Therefore, measuring satisfaction with care may not offer a true reflection of patients’ experiences of the continuity of their care.

Within the context of changes to patient care and management in oncology during the late 1990s, we wished to examine patients’ perceptions of the continuity of their care. Nevertheless, at this time there were no cancer-specific instruments suitable for measuring continuity of care in secondary/tertiary health services. As such, we adapted an existing instrument that assessed continuity of care in the primary health-care setting. Over several years, we have continued to develop this instrument and have used it in randomized trials to document patients’ experiences of the continuity of their care. In this article, we present data showing the development and psychometric validity of the Medical Care Questionnaire (MCQ).
Table 1  Aims and methods for each study phase

<table>
<thead>
<tr>
<th>Phase</th>
<th>Year</th>
<th>Aim</th>
<th>Procedure</th>
</tr>
</thead>
</table>
| Phase 1 | 1999   | Review literature relevant to continuity of care issues for outpatient oncology and identify relevant instruments | 1. Literature search  
2. Review existing instruments  
3. Expert review of CPCI  
4. Removal of incompatible items  
5. Rewording existing items |
| Phase 2 | 1999   | Obtain expert opinion on the relevance of the CPCI and to modify the instrument to be applicable to outpatient oncology practice | 1. Patient completion of questionnaire  
2. Descriptive analysis and modification of questionnaire  
3. Psychometric exploration of factor structure  
4. Patient completion of secondary instruments  
5. Psychometric exploration of factor structure, reliability, validity, and known groups |
| Phase 3 | 1999–2000 | Test the acceptability and relevance of the adapted questionnaire and explore its measurement properties | 1. Literature search  
2. Review existing instruments  
3. Expert review of CPCI  
4. Removal of incompatible items  
5. Rewording existing items |
| Phase 4 | 2000–2003 | Examine the hypothesized subscales in a new patient population. | 1. Literature search  
2. Review existing instruments  
3. Expert review of CPCI  
4. Removal of incompatible items  
5. Rewording existing items |

CPCI, Components of Primary Care Index.

Methods

The development of the MCQ was carried out in four phases. Phase 1 was a literature review to determine whether existing instruments could be used or adapted for outpatient oncology. Phase 2 included modification of an existing instrument (Components of Primary Care Index [CPCI]) [8] by expert review. Phase 3 was a pilot study to explore the psychometric properties of the refined instrument in a patient population. The results of Phase 3 suggested further modification of the instrument, so Phase 4 examined the validity of the instrument in a larger patient sample. Each phase was carried out sequentially and data for Phases 3 and 4 were collected from the same medical oncology outpatient clinic. Table 1 summarizes the aims and methods for each phase of instrument development.

Phase 1: Literature Review

A literature search was performed in Medline, using the key words “continuity of care,” “co-ordination of care,” “patient satisfaction,” and “cancer.” The purpose of the review was twofold: 1) to identify definitions of “continuity and co-ordination of care” applicable to secondary/tertiary hospital care, and 2) to find instruments that measure coordination and continuity of medical care from patients’ perspective. The literature review did not identify any self-reported instruments suitable for hospital-based oncology practices. One instrument, the CPCI, which was designed for use in primary health care, was found to employ a useful taxonomy and included a number of items and subscales that were of relevance to the cancer care setting. This questionnaire consists of 19 items, organized into four domains: “Patient preference to see usual doctor,” “Interpersonal communication,” “Physician’s accumulated knowledge of the patient,” and “Co-ordination of care.” The internal consistency reliability of the four subscales ranged between 0.68 and 0.79. The instrument demonstrates good psychometric properties and was originally developed and evaluated in a sample of 2899 primary care patients visiting 138 family physicians’ offices in the United States [8,9]. All items have a 5-point Likert scale response format anchored by strongly agree and strongly disagree. The way in which items are phrased requires patients to report rather than rate their interaction with the physician. Because cancer patients are typically reluctant to rate their physicians poorly [6], the less judgemental reporting style may serve to reduce ceiling effects from responses.

Phase 2: Expert Review

While the CPCI provides a valuable scale structure and taxonomy, many items are phrased in a manner unsuitable for the purposes of team-based hospital care. The CPCI was reviewed for applicability to outpatient oncology by an expert panel of three consultant medical oncologists and the experimenter (GV), an oncologist in training. The experts were selected from medical oncology and were chosen because they had experience in managing team-based patient care across different cancer specialities. A stepwise procedure (similar to the Delphi technique) was used to adapt the original CPCI and consensus was reached for each decision to alter, remove, or add an item. The first step included a review of item content to determine applicability to cancer patients. The second step examined the wording of the remaining items and the final step was item generation to replace items that had been removed. See Table 2 for each modification step. The modified questionnaire was renamed the MCQ.

Phase 3 and 4: Evaluating the Psychometric Properties of the New Instrument

In Phase 3, as part of an outpatient audit, 285 cancer patients were invited to complete the new MCQ instrument during their visit to the hospital or by post. Of those contacted, 200 (70%) patients returned completed questionnaires. For Phase 4, MCQ responses were collected from patients taking part in two separate studies. The first study was a postal audit to determine patient experiences of their care; 313 cancer patients were contacted by post and asked to complete and return the MCQ. Two hundred fifteen (69%) completed questionnaires were returned. The second study was a randomized controlled trial (RCT) examining the impact of routine quality of life assessment on patient–doctor communication [10]. In this study, patients were asked to complete the MCQ at baseline. Of the 286 cancer patients who took part in the RCT, 262 (92%) completed the MCQ. In total, 477 patients completed the MCQ questionnaire in Phase 4.

The MCQ was administered at a regional hospital (North England) with a specialist cancer service (Medical Oncology Unit). The audits carried out in Phases 3 and 4 were performed as part of a service improvement and as such were not subject to NHS ethical approval. Adult patients from all tumor groups attending the Medical Oncology Unit were eligible to take part in Phases 3 and 4, provided they could read and understand English and in the opinion of the investigator they were not exhibiting overt cognitive dysfunction or signs of distress. The Phase 4 RCT received NHS ethical approval and all patients gave written informed consent before data collection in accordance with the Declaration of Helsinki.
This doctor always follow up on a problem I've had. This doctor knows the results of my visits to other health care providers I see. My medical care improves when I see the same doctor that I have seen before. It is very important to me to see my regular doctor. This doctor and I have been through a lot together. The doctors clearly understand my health needs. If I am sick, I would always contact a doctor in this clinic. I go to this doctor for almost all of my medical care. How many years have you been a patient of this physician? I go to this doctor for almost all of my medical care. I rarely see the same doctor when I come to this clinic. Sometimes the doctors do not listen to me. This doctor communicates with the other health care providers I see. The doctors I see in this clinic communicate with each other. The doctors do not always know my medical history and problems very well. My medical care improves when I see the same doctor that I have seen before. It is very important to me to see my regular doctor. The doctors know how I feel emotionally while they are treating me. This doctor does not always know about care I have received at other places. I don’t always feel comfortable asking questions of this doctor. This doctor knows a lot about the rest of my family. The doctors know about non-medical things in my life (family, job, hobbies, social life). I sometimes have to repeat my problems to the different doctors I see in this clinic. I would rather wait for the doctor who saw me last than be seen by the next available doctor in clinic. The doctors usually know about the problems that have bothered me at the previous visits. I can easily talk about personal things with the doctors. Demographic Details

For all studies in Phases 3 and 4, patient medical details, such as the primary tumor site, were recorded from medical notes. Patients completed a sociodemographic survey that included details on the patients’ age, sex, marital status, and employment status. Medical and social demographic details are summarized in Table 3. Patients were predominantly female (81% in Phase 3; 74% in Phase 4) and diagnoses of gynecological, breast, and genitourinary cancers were most common. The biases in distribution of sex and diagnosis reflect the demographics of the unit, with three specialized clinics in breast and gynecological cancers, and one general oncology clinic.

Sample Size

For factor and multitrait analyses, sample size is typically recommended to require 5–10 times the number of participants as the number of items included in the instrument [11]. In Phase 3, the MCQ included 21 items and was completed by 200 patients, giving a subject to item ratio of 9.5:1. In Phase 4, the MCQ contained 15 items and was completed by 477 patients, giving a subject to item ratio of 31.8:1.

Descriptive Analysis

In Phase 3, descriptive data were examined to assess the acceptability of each item to patients and to evaluate the contribution of each item to the scale. Positively worded items on the MCQ were reversed scored to be consistent with the remaining items. Criteria for retaining items included: 1) response ranges spanned three or more response categories (i.e., categories 1 through to 4, or 2 through to 5 were selected); 2) mean values ≤; and 3) no ceiling effect, i.e., frequency of responses for less favorable response categories should be >20%. Items not meeting these criteria were removed before exploratory factor analysis (EFA) because they were deemed likely to contribute to a ceiling effect.

Table 2 Adaptation process of CPCI to MCQ showing item wording and factor loading of final MCQ items

<table>
<thead>
<tr>
<th>Original 19-item CPCI</th>
<th>Phase 2 expert review</th>
<th>Phase 3 psychometric evaluation</th>
</tr>
</thead>
<tbody>
<tr>
<td>I rarely see the same doctor when I go for medical care</td>
<td>I rarely see the same doctor when I come to this clinic</td>
<td>1 0.403 1</td>
</tr>
<tr>
<td>Sometimes this doctor does not listen to me</td>
<td>Sometimes the doctors do not listen to me</td>
<td>1 0.506 2</td>
</tr>
<tr>
<td>I want one doctor to coordinate all the health care I receive</td>
<td>I want one doctor to coordinate all the care I receive</td>
<td>2 0.639 3r</td>
</tr>
<tr>
<td>This doctor communicates with the other health care providers I see</td>
<td>The doctors I see in this clinic communicate with each other</td>
<td>3 0.555 4r</td>
</tr>
<tr>
<td>This doctor do not always know my medical history very well</td>
<td>The doctors do not always know my medical history</td>
<td>1 0.678 5</td>
</tr>
<tr>
<td>My medical care improves when I see the same doctor</td>
<td>My medical care improves when I see the same doctor</td>
<td>2 0.620 6r</td>
</tr>
<tr>
<td>that I have seen before</td>
<td>that I have seen before</td>
<td></td>
</tr>
<tr>
<td>It is very important to me to see my regular doctor</td>
<td>It is very important to me to see my regular doctor</td>
<td>2 0.878 7r</td>
</tr>
<tr>
<td>This doctor and I have been through a lot together</td>
<td>The doctors know how I feel emotionally while they are treating me</td>
<td>3 0.785 8r</td>
</tr>
<tr>
<td>This doctor does not always know about care I have received at other places</td>
<td>The doctors do not always know about the care and treatment I have received previously in this clinic</td>
<td>1 0.677 9</td>
</tr>
<tr>
<td>I don’t always feel comfortable asking questions of this doctor</td>
<td>I don’t always feel comfortable asking the doctors</td>
<td>1 0.513 10</td>
</tr>
<tr>
<td>This doctor knows a lot about the rest of my family</td>
<td>The doctors know about non-medical things in my life (family, job, hobbies, social life)</td>
<td>3 0.642 11r</td>
</tr>
<tr>
<td>NEW</td>
<td>I sometimes have to repeat my problems to the different doctors I see in this clinic</td>
<td>1 0.506 12</td>
</tr>
<tr>
<td>NEW</td>
<td>I would rather wait for the doctor who saw me last than be seen by the next available doctor in clinic</td>
<td>2 0.736 13r</td>
</tr>
<tr>
<td>NEW</td>
<td>The doctors usually know about the problems that have bothered me at the previous visits</td>
<td>3 0.591 14r</td>
</tr>
<tr>
<td>I can easily talk about personal things with this doctor</td>
<td>I can easily talk about personal things with the doctors</td>
<td>3 0.458 15r</td>
</tr>
</tbody>
</table>

*The three factors were labeled as follows: factor 1, “Coordination” of medical information and doctor’s accumulated knowledge about patient; factor 2, “Preferences” to see usual doctor; and 3, “Communication” with doctor. r represents that the item has been reversed scored.

CPCI, Components of Primary Care Index; MCQ, Medical Care Questionnaire.
**Medical Care Questionnaire (MCQ)**

**Table 3** Patient demographic and clinical details for study phases 3 and 4

<table>
<thead>
<tr>
<th></th>
<th>Phase 3 n = 200</th>
<th>Phase 4 n = 477</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sex, n (%)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>162 (81)</td>
<td>354 (74.2)</td>
</tr>
<tr>
<td>Male</td>
<td>38 (19)</td>
<td>123 (25.8)</td>
</tr>
<tr>
<td><strong>Age group (years), n (%)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15–29</td>
<td>8 (4)</td>
<td>14 (2.9)</td>
</tr>
<tr>
<td>30–44</td>
<td>27 (13.5)</td>
<td>74 (15.5)</td>
</tr>
<tr>
<td>45–59</td>
<td>84 (42)</td>
<td>195 (40.9)</td>
</tr>
<tr>
<td>60–74</td>
<td>67 (33.5)</td>
<td>160 (33.5)</td>
</tr>
<tr>
<td>75+</td>
<td>14 (7)</td>
<td>34 (7.1)</td>
</tr>
<tr>
<td><strong>K index, median (range)</strong></td>
<td>0.3 (0–1)</td>
<td>0.50 (0–1)</td>
</tr>
<tr>
<td><strong>Marital status, n (%)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Single</td>
<td>12 (6)</td>
<td>30 (6.3)</td>
</tr>
<tr>
<td>Married/cohabiting</td>
<td>148 (74)</td>
<td>358 (75.1)</td>
</tr>
<tr>
<td>Divorced/widowed</td>
<td>37 (18.5)</td>
<td>84 (17.6)</td>
</tr>
<tr>
<td>Missing</td>
<td>3 (1.5)</td>
<td>5 (1.0)</td>
</tr>
<tr>
<td><strong>Employment status, n (%)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Working full time</td>
<td>26 (13)</td>
<td>211 (44.2)</td>
</tr>
<tr>
<td>Working part time</td>
<td>23 (11.5)</td>
<td>43 (9.0)</td>
</tr>
<tr>
<td>On sick leave</td>
<td>51 (25.5)</td>
<td>58 (12.2)</td>
</tr>
<tr>
<td>Homemaker</td>
<td>16 (8)</td>
<td>40 (8.4)</td>
</tr>
<tr>
<td>Retired</td>
<td>76 (38)</td>
<td>112 (23.5)</td>
</tr>
<tr>
<td>Other</td>
<td>4 (2)</td>
<td>5 (1.0)</td>
</tr>
<tr>
<td>Missing</td>
<td>4 (2)</td>
<td>8 (1.7)</td>
</tr>
<tr>
<td><strong>Diagnosis, n (%)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Breast cancer</td>
<td>53 (26.5)</td>
<td>112 (23.5)</td>
</tr>
<tr>
<td>Gastrointestinal</td>
<td>9 (4.5)</td>
<td>0</td>
</tr>
<tr>
<td>Genitourinary</td>
<td>33 (16.5%)</td>
<td>102 (21.4)</td>
</tr>
<tr>
<td>Gynecological</td>
<td>76 (38)</td>
<td>161 (33.8)</td>
</tr>
<tr>
<td>Melanoma</td>
<td>2 (1)</td>
<td>40 (8.4)</td>
</tr>
<tr>
<td>Sarcoma</td>
<td>11 (5.5)</td>
<td>36 (7.5)</td>
</tr>
<tr>
<td>Other</td>
<td>16 (8)</td>
<td>26 (5.5)</td>
</tr>
</tbody>
</table>

**K index** = (Number of Visits – Number of Doctors) / (Number of visits – 1)

**Exploratory Factor Analysis**

Kolmogorov–Smirnov tests showed the MCQ data in both Phases 3 and 4 were not normally distributed (P < 0.05), therefore the latent structure of the instrument was examined using principle axis factoring. Oblique (direct oblimin) rotation was applied because the original CPCI reported that the factors “coordination” and “accumulated knowledge” were correlated [8] and correlations between factors were expected for current data. The criteria for factor extraction were a minimum eigenvalue of 1.00 and that each component accounted for at least 5% of the variance among items. Scree plots assisted the decision to retain factors. Data with more than 40% missing values were removed before analysis and remaining missing data were replaced by mean values for the item.

**Confirmatory Factor Analysis**

The suggested factor structure of the MCQ (from Phase 3 exploratory factor analysis) was examined using confirmatory factor analysis (CFA) with data in Phase 4. Goodness of fit was deemed acceptable if the chi-square value was low with a nonsignificant P-value, and if the root mean square error of approximation (RMSEA) was below 0.080.

**Reliability**

In Phase 4, the reliability of each subscale was examined using multitrait analysis. This analysis examined the item-convergent and item-discriminant validity of the subscales that were derived from Phase 3 exploratory factor analysis and supported by Phase 4 CFA. Item-convergent validity was supported if items had correlations >0.40 with their own hypothesized subscale. Item-discriminant validity was supported if items correlated more highly with their own hypothesized subscale than they did with other subscales. The internal consistency reliability of each subscale and the total scale was assessed by calculating Cronbach’s alpha (α) coefficients. Values above 0.70 were accepted as moderate, while values above 0.80 were accepted as showing good internal consistency.

**Validity**

In Phase 4, an objective measure of “continuity of care” was derived to explore the external validity of the MCQ. The literature describes several indexes for continuity of care developed mainly for family practice [12]. The simplest measure considers the number of visits each patient has made and the number of care providers seen, this is called the “K index” [13,14]. The K index can be applied to a team-based hospital oncology practice by recording the number of doctors each patient has seen and the total number of clinic visits over time.

K index = (Number of Visits – Number of Doctors) / (Number of visits – 1)

The K index has a value between 0 and 1. When a patient has seen only one doctor over time, the K index will be 1. When a patient has seen different doctors at each visit, the K index will be 0.

The validity of the MCQ subscales was explored against medical and demographic known groups. MCQ subscale scores were derived by computing the mean of subscale items and linearly transforming the data to a 0–100 scale. One-way analysis of variance (ANOVA) tests (with Bonferroni corrections for post hoc analyses) were carried out to determine any differences in subscale scores for the following groups: diagnosis (breast, genitourinary, gynecological, melanoma, sarcoma, or other); and K index quartiles (quartiles were calculated using SPSS to identify the score boundaries for the 25th, 50th, and 75th percentiles: 1st = 0.0–0.24; 2nd = 0.25–0.49; 3rd = 0.50–0.59; 4th = 0.60–1.00). Independent samples t tests compared subscale scores between age groups (less than or more than 60 years) and between sexes.

Data were analyzed using SPSS version 16.0 for Windows (SPSS Inc., Chicago, IL) and LISREL 8.80 Student (Scientific Software International, Inc, Lincolnwood IL). The threshold for statistical significance was set at P < 0.05. Effect sizes for ANOVAs (Cohen’s f) were calculated using GPower 3.0.10 [15]. Cohen’s f values are interpreted as small = 0.10, medium = 0.25, and large = 0.40 [16].

**Results**

**Phase 2: Expert Review**

Five items were removed from the original CPCI instrument because they were not considered applicable to the cancer outpatient population. Minor changes were made to eight items, such as the replacement of “this doctor” with “the doctors” or the addition of a few words to specify the setting, i.e., “this clinic,” and major changes were made to two items. Seven new items were added to the instrument, which covered aspects of medical care specific to oncology and the system of delivery of cancer care. The expert review resulted in a 21-item instrument renamed the MCQ. See Table 2 for the expert review stages including the original CPCI items and the adapted MCQ items.
Phase 3: Descriptive Analysis

The proportion of missing responses to the 21-item MCQ was low (1–5%). Five items did not meet the criteria for retention because they had high mean scores (range 4.4–4.5 across items) and had a low cumulative frequency of less favourable responses (range 9–12% across items). As such, these five items were removed. One item (“I don’t mind seeing different doctors because everyone in the team knows my case”) was deleted despite meeting the criteria because it was a double statement with ambiguous meaning. After descriptive analysis, the MCQ instrument was reduced to 15-items. These remaining items were subject to exploratory factor analysis.

Phase 3: Exploratory Factor Analysis and Reliability

Three factors were extracted with eigenvalues greater than 1 and which accounted for at least 5% of variance in the data. Examination of the inflexion point of the Scree plot confirmed the retention of three factors. The 3 factors accounted for 45.47% of the common variance and were labeled: “Coordination,” which included items on the coordination of patient information and accumulated physician knowledge about the patient; “Preferences,” which included items on patient preferences to see their usual doctor; and “Communication,” which included items on communication with doctors and knowledge about non-medical issues. Each of the three subscales showed satisfactory internal consistency (Cronbach’s alpha): Coordination α = 0.76; Preferences α = 0.83; and Communication α = 0.80. The subscale scores were interpreted as follows: patients with higher “Communication” and “Coordination” scores on the MCQ rated their communication with doctors and coordination of their medical information as better than patients with lower scores; patients with higher “Preferences” subscale scores had a stronger preference for seeing their usual doctor (or fewer doctors) during clinic visits than patients with lower scores.

Phase 4: CFA

The three-factor model derived by EFA of the Phase 3 data was examined in the Phase 4 data with CFA. We were concerned that the Coordination subscale derived by EFA could have been an artefact because the items contributing to this subscale were all negatively worded. To determine whether the Coordination subscale should be kept as an independent subscale or merged with the Communication subscale, we compared the goodness of fit of two models. The first model contained two factors: factor 1 combined all items from the Coordination and Communication subscales and factor 2 contained the items from the preferences subscale. The second model contained three factors, with the items remaining within the three factors described in the Phase 3 EFA.

The two-factor model had poorer fit than the three-factor model. Goodness of fit for two-factor model: χ² = 405.04; df = 89; P = 0.000; RMSEA = 0.086; confidence interval (CI) of RMSEA = 0.078–0.095. The modification indices suggested adding paths between factor 1 (combined Coordination/ Communication subscale) and item 13; and paths between the Preferences subscale and items 1, 2, 8, and 14. The three-factor model showed improvement in goodness of fit compared with the 2 factor model: χ² = 269.15; df = 87; P = 0.000; RMSEA = 0.066; CI of RMSEA = 0.057–0.075. The modification indices suggest adding a path between the Preferences subscale and item 1, and adding paths between the Communication subscale and items 1 and 13.

Despite the improvement in fit between the two-factor and the three-factor models, the chi-square value remained high and significant. Nevertheless, the chi-square is often reported to be inflated by large sample sizes, and the acceptable RMSEA score for the three-factor model suggested adequate fit of the 3 factor model. The reliability of the three-factor model was explored further with multitrait analyses, to determine whether any items should be removed or moved from the three subscales.

Phase 4: Multitrait Item-Subscale Correlations

Data from Phase 4 were used to examine item-convergent and item-discriminant validity of the 3 factor domain structure using multitrait correlation analyses (Table 4). Items 1 and 10 showed low item-convergent validity (0.36 for both items) with the Coordination subscale but did not show higher correlations with other subscales. Item 11 showed low item-convergent validity (0.38) with the Communication subscale, but did not have a higher correlation with any other subscale. Item 13 showed good item-convergent validity with the Preferences subscale. The Coordination and Communication subscales were positively correlated (r[469] = 0.45, P < 0.001). The internal consistency (Cronbach’s alpha) for the three subscales was: Communication = 0.69; Preferences = 0.84; and Coordination = 0.75. Cronbach’s alpha “if item deleted” values were examined to determine whether the subscales would be improved with the removal of items 1, 10, and 11. Cronbach’s alpha for the Coordination subscale showed no improvement for removing item 10 and showed only a small improvement of 0.01 with the removal of item 1. Cronbach’s alpha for the Communication subscale showed no improvement with the removal of item 11. We decided to retain items 1, 10, 11, and 13 in the original subscales, as suggested by Phase 3 EFA.

Phase 4: Known Groups Comparisons

Patients were divided into groups based on demographic and medical details and their scores on the MCQ subscales were compared (Table 5). Patients with breast cancer had lower Coordination subscale scores (F[546] = 2.53, P = 0.028, f = 0.16) than patients with melanoma cancer but had higher Preferences subscale scores (F[545] = 3.75, P = 0.002, f = 0.20) than patients with gynecological cancer. Breast cancer patients also showed item-discriminant validity for removing items 10 and 11. It is important to note that Cronbach’s alpha for the Communication subscale showed no improvement with the removal of item 11. We decided to retain items 1, 10, 11, and 13 in the original subscales, as suggested by Phase 3 EFA.

Table 4 Multitrait item-subscale correlations (phase 4)

<table>
<thead>
<tr>
<th></th>
<th>Coordination</th>
<th>Preferences</th>
<th>Communication</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coordination</td>
<td>—</td>
<td>-0.292*</td>
<td>0.450*</td>
</tr>
<tr>
<td>1</td>
<td>0.362†</td>
<td>0.130</td>
<td>0.298</td>
</tr>
<tr>
<td>2</td>
<td>0.533†</td>
<td>-0.242</td>
<td>0.302</td>
</tr>
<tr>
<td>3</td>
<td>0.562†</td>
<td>-0.183</td>
<td>0.376</td>
</tr>
<tr>
<td>4</td>
<td>0.591†</td>
<td>-0.231</td>
<td>0.353</td>
</tr>
<tr>
<td>5</td>
<td>0.364†</td>
<td>-0.166</td>
<td>0.225</td>
</tr>
<tr>
<td>6</td>
<td>0.556†</td>
<td>-0.225</td>
<td>0.356</td>
</tr>
<tr>
<td>7</td>
<td>-0.292*</td>
<td>—</td>
<td>-0.066*</td>
</tr>
<tr>
<td>8</td>
<td>-0.202</td>
<td>0.683†</td>
<td>-0.034</td>
</tr>
<tr>
<td>9</td>
<td>-0.218</td>
<td>0.643†</td>
<td>-0.009</td>
</tr>
<tr>
<td>10</td>
<td>-0.246</td>
<td>0.769†</td>
<td>0.009</td>
</tr>
<tr>
<td>11</td>
<td>-0.276</td>
<td>0.645†</td>
<td>-0.139</td>
</tr>
<tr>
<td>12</td>
<td>0.450†</td>
<td>-0.066*</td>
<td>—</td>
</tr>
<tr>
<td>13</td>
<td>0.277</td>
<td>-0.118</td>
<td>0.418†</td>
</tr>
<tr>
<td>14</td>
<td>0.266</td>
<td>0.028</td>
<td>0.503†</td>
</tr>
<tr>
<td>15</td>
<td>0.344</td>
<td>0.019</td>
<td>0.504†</td>
</tr>
<tr>
<td>16</td>
<td>0.409</td>
<td>-0.102</td>
<td>0.472†</td>
</tr>
</tbody>
</table>

*Correlation between subscales (subscale values derived by computing the mean of subscale items and linearly transforming the data to a 0–100 scale).
†Item correlation with own scale, corrected for overlap.
*Correlation between subscales (subscale values derived by computing the mean of subscale items and linearly transforming the data to a 0–100 scale).
†Item correlation with own scale, corrected for overlap.
MCQ, Medical Care Questionnaire.
had lower Communication subscale scores (F[5465]= 3.09, P = 0.009, f = 0.18) than patients with gynecological (P = 0.027) or melanoma (P = 0.050) cancers. Individuals with the highest K index (4th quartile) had higher Preferences subscale scores (F[3435]= 6.46, P = 0.000, f = 0.21) than patients from lower K index quartile groups (1st quartile P = 0.083; 2nd quartile P < 0.001; 3rd quartile P = 0.049). There were no between group differences by K index quartile for Coordination or Communication subscale scores. There were no between group differences in MCQ subscale scores for age group or sex.

Discussion

We have presented the various stages of development and validation of the MCQ, to measure oncology patients’ perceptions of the continuity and coordination of their medical care and communication with their doctors. The MCQ was adapted from the CPCI by a process of expert review and psychometric evaluation. This process led to a number of changes being made to the original 19-item CPCI to make it applicable to an oncology setting. This included removing or rewording items and generating new items. Although the item adaptation process was based on consensus methods using expert reviewers, it could have been improved by including patient opinions and feedback. Although not reported in this study, patient feedback was elicited during Phase 3. During this phase, patients were encouraged to comment on the items and give feedback on the questionnaire. Patient feedback was analyzed qualitatively and was taken into consideration alongside the descriptive analysis. In summary, most patients confirmed the importance of the identified subscales: many patients reported that it was important for them to see the same doctor at each visit and that the coordination of their medical information between individual doctors and the wider medical team was very important to their care. The adapted instrument was renamed the MCQ and contained 21 items.

Initial psychometric evaluation of the 21-item MCQ suggested removal of five items that contributed to a ceiling effect in responses. Of these items, two were from the original CPCI questionnaire, two were adapted from the original CPCI, and one was a new item. One additional new item was removed because it was considered a double statement and was ambiguous to interpret. The psychometric evaluation of the remaining 15 items suggested the MCQ measured three domains of continuity of care: “Communication” with doctors; “Coordination” of medical information and physicians’ accumulated knowledge about the patient; and “Preferences” to see usual doctor. The “Preferences” subscale remained from the original CPCI, with the addition of a new item (item 1) “I rarely see the same doctor when I come to this clinic.” Subsequent evaluation of the hypothesized domains in a new patient population showed that the three subscales had reasonable internal and external reliability and validity in the target population. While the item-factor structure of the MCQ differs from the CPCI, the two instruments remain conceptually similar in the measurement of patients’ preferences to see their usual doctor, their evaluation of communication, and their perception of the coordination of their medical information between doctors. The differences in factor structure between the two instruments might be expected given the removal of seven original CPCI items and the addition of three new items to the MCQ. The differences in factor structure may also be because of differences in medical setting (primary versus secondary/tertiary care) and the different patient population sampled in the current study.

After item deletion in Phase 3, the 15-item MCQ was administered to a new oncology outpatient population in Phase 4. This data were used to reexamine the hypothesized domain structure and internal validity of the MCQ and examined its external validity by comparing known groups. Although this second administration of the 15-item MCQ showed slightly poorer internal validity of the subscales than in the previous sample, each subscale showed reasonable internal consistency and reliability, and appears suitable for use in a mixed oncology outpatient population.

The Communication domain of the MCQ is an element of continuity of care that was not identified by Freeman et al. [5] as being important to continuity of care. Nevertheless, in oncology, it is important that the patients and doctors maintain good levels of communication to enable the identification of symptoms and toxicities during treatment and to monitor the impact of disease and treatment on broader social and psychological well-being. The items in the Communication domain of the MCQ reflect the importance of communication about nonmedical issues with items such as: "The doctors know how I feel emotionally while I am in the clinic." Subsequent evaluation of the hypothesized domains in a new patient population showed that the three subscales had reasonable internal and external reliability and validity in the target population. While the item-factor structure of the MCQ differs from the CPCI, the two instruments remain conceptually similar in the measurement of patients’ preferences to see their usual doctor, their evaluation of communication, and their perception of the coordination of their medical information between doctors. The differences in factor structure between the two instruments might be expected given the removal of seven original CPCI items and the addition of three new items to the MCQ. The differences in factor structure may also be because of differences in medical setting (primary versus secondary/tertiary care) and the different patient population sampled in the current study.

After item deletion in Phase 3, the 15-item MCQ was administered to a new oncology outpatient population in Phase 4. This data were used to reexamine the hypothesized domain structure and internal validity of the MCQ and examined its external validity by comparing known groups. Although this second administration of the 15-item MCQ showed slightly poorer internal validity of the subscales than in the previous sample, each subscale showed reasonable internal consistency and reliability, and appears suitable for use in a mixed oncology outpatient population.

The Communication domain of the MCQ is an element of continuity of care that was not identified by Freeman et al. [5] as being important to continuity of care. Nevertheless, in oncology, it is important that the patients and doctors maintain good levels of communication to enable the identification of symptoms and toxicities during treatment and to monitor the impact of disease and treatment on broader social and psychological well-being. The items in the Communication domain of the MCQ reflect the importance of communication about nonmedical issues with items such as: “The doctors know how I feel emotionally while

---

Table 5  Known group comparisons (phase 4)

<table>
<thead>
<tr>
<th>Groups</th>
<th>Communication</th>
<th>Coordination</th>
<th>Preferences</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n</td>
<td>Mean</td>
<td>SD</td>
</tr>
<tr>
<td>Sex</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>121</td>
<td>70.80</td>
<td>16.86</td>
</tr>
<tr>
<td>Female</td>
<td>350</td>
<td>67.79</td>
<td>17.72</td>
</tr>
<tr>
<td>Age</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;60</td>
<td>280</td>
<td>67.74</td>
<td>17.61</td>
</tr>
<tr>
<td>&gt;60</td>
<td>191</td>
<td>69.78</td>
<td>17.42</td>
</tr>
<tr>
<td>Tumor group</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Breast</td>
<td>110</td>
<td>63.17</td>
<td>18.66</td>
</tr>
<tr>
<td>Gastrointestinal</td>
<td>102</td>
<td>69.32</td>
<td>17.34</td>
</tr>
<tr>
<td>Gynecological</td>
<td>159</td>
<td>69.93</td>
<td>16.41</td>
</tr>
<tr>
<td>Melanoma</td>
<td>40</td>
<td>72.63</td>
<td>18.22</td>
</tr>
<tr>
<td>Sarcoma</td>
<td>34</td>
<td>69.60</td>
<td>17.87</td>
</tr>
<tr>
<td>Other</td>
<td>26</td>
<td>72.45</td>
<td>14.91</td>
</tr>
<tr>
<td>K index quartiles</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0–0.24</td>
<td>103</td>
<td>64.94</td>
<td>17.04</td>
</tr>
<tr>
<td>0.25–0.49</td>
<td>117</td>
<td>69.46</td>
<td>18.36</td>
</tr>
<tr>
<td>0.50–0.75</td>
<td>72</td>
<td>67.88</td>
<td>17.70</td>
</tr>
<tr>
<td>0.80–1.00</td>
<td>161</td>
<td>70.53</td>
<td>17.19</td>
</tr>
</tbody>
</table>

P-value from independent samples t test.

P-value from one-way analysis of variance showing overall group effect.
they are treating me,” “The doctors know about non-medical things in my life,” and “I can easily talk about person things with the doctors.” Among the patient groups, we found that patients with breast cancer reported lower Communication scores than patients with gynecological and melanoma cancers.

The Coordination domain of the MCQ was considered to reflect patients’ experiences of continuity when they saw different doctors for their medical care. Although we have given this subscale the label Coordination, it is clear from the items included in this domain that patients’ experiences of coordination are dependent on the quality of communication and the flow of information between health professionals and across clinic visits. As may be expected, we found the Communication and Coordination subscales were correlated. We explored combining the two subscales, using CFA of phase 4 data, but this analysis showed better model fit when the factors Communication and Coordination remained separate. In line with the correlation between the two factors, we found that patients with breast cancer reported lower Communication scores than patients with gynecological and melanoma patients also reported lower Coor-

dination subscale scores than patients with melanoma cancer.

The Preferences subscale was considered to reflect the impor-
tance that patients place in seeing the same health professional at each hospital visit. This has been identified in previous research as an important component of continuity of care [4,5]. We found a small negative correlation between the Preferences subscale and the Coordination subscale, suggesting that those patients who rated the coordination of their medical information between doctors as poor might be more likely to place greater value in seeing fewer health professionals for their medical care. In this study, patients with breast cancer were more likely than patients with gynecological cancer to endorse items from the Preferences subscale. Perhaps counter intuitively we found that patients with high K index values (who saw a fewer doctors per hospital visit) were more likely to endorse items from the Preferences subscale. This result may reflect that some clinics within the Medical Oncology Unit endeavor to accommodate patients who have strong preferences to see a particular doctor for their clinic visit. It could be that a number of patients with higher scores on the Preferences subscale were more active in ensuring their care was delivered by particular health professionals.

There were several limitations to this study. Although we believe our sample provided good representation of the patient population seen in the Medical Oncology Unit, the majority of the patients were female with breast or gynecological cancers and the results may be biased toward female opinion. While the comparison of MCQ subscales scores between males and females did not show any significant differences, further validation of the questionnaire to include a larger number of men with cancer would be desirable. Another limitation to generalized interpretation of the results is that the study phases 3 and 4 were carried out in a single Medical Oncology Unit, part of a tertiary referral cancer center. A typical feature of this setting is the large number of doctors looking after the patients (teams of approximately 4–8 doctors), which was reflected in the relatively low K-index in our study populations. Thus, further validation of the MCQ may be required before it is applicable to hospitals where the oncology care is delivered by a smaller team of doctors.

While the validity of the MCQ has been shown to be good in a general cancer population, it is important that further work is carried out to establish test–retest validity and to gather stronger data on the relationship between patient scores and indicators of clinical practice that are predicted to affect continuity of medical care. Until the psychometric properties of the MCQ have been validated further, we recommend that patient responses to the MCQ are interpreted at the level of the three domains rather than calculating a 15-item total score.

We have provided preliminary evidence that the MCQ instru-
ment can provide valuable information on patients’ experiences of communicating with doctors and their perceptions of the continuity and coordination of their medical care. The MCQ instrument is brief (5–10 minutes to complete), easy to adminis-
ter, and is simple to score, therefore we feel it would be a valuable and suitable patient-reported measure to be used in busy oncol-
yology practice, clinical trials, and service improvement programs.

Source of financial support: This study was funded by Cancer Research UK C7775/A294. Data from this article have been presented in part at the 8th Annual Conference of the International Society for Quality of Life Research (ISOQOL), November 7–10, Amsterdam, The Netherlands. Booth, L, Selby, P, Lynch, P, Brown, J, Velikova, G. Measurement of continuity and coordination of care in a cancer centre. Quality of Life Research, 2001, 10 (3), 246 (Abstract No 212).

References


South Korean Time Trade-Off Values for EQ-5D Health States: Modeling with Observed Values for 101 Health States

Yeon-Kyung Lee, PhD,1 Hae-Sung Nam, MD, PhD,2 Ling-Hsiang Chuang, MSc,3 Keon-Yeop Kim, MD, PhD,4 Hae-Kyung Yang, MPH,2 In-Sun Kwon, MS,2 Paul Kind, MPhil,3 Sun-Seog Kweon, MD, PhD,5 Young-Tack Kim, MD, PhD1

1Division of Chronic Disease Surveillance, Korea Centers for Disease Control and Prevention, Seoul, Korea; 2Department of Preventive Medicine, Research Institute for Medical Sciences, College of Medicine, Chungnam National University, Daejeon, Korea; 3Centre for Health Economics, University of York, York, UK; 4Department of Preventive Medicine, School of Medicine, KyungPook National University, Daegu, Korea; 5Jeonnam Regional Cancer Center, Chonnam National University Hwasun Hospital, Jeonnam, Korea

ABSTRACT

Objectives: This study establishes the South Korean population-based preference weights for EQ-5D based on values elicited from a representative national sample using the time trade-off (TTO) method.

Methods: The data for this paper came from a South Korean EQ-5D valuation study where 1307 representative respondents were invited to participate and a total of 101 health states defined by the EQ-5D descriptive system were directly valued. Both aggregate and individual level modeling were conducted to generate values for all 243 health states defined by EQ-5D. Various regression techniques and model specifications were also examined in order to produce the best fit model. Final model selection was based on minimizing the difference between the observed and estimated value for each health state.

Results: The N3 model yielded the best fit for the observed TTO value at the aggregate level. It had a mean absolute error of 0.029 and only 15 predictions out of 101 had errors exceeding 0.05 in absolute magnitude.

Conclusions: The study successfully establishes South Korean population-based preference weights for the EQ-5D. The value set derived here is based on a representative population sample, limiting the interpolation space and possessing better model performance. Thus, this EQ-5D value set should be given preference for use with the South Korean population.

Keywords: EQ-5D, population values, preference-based measures, time trade-off.

Introduction

Economic evaluations of health-care interventions provide important evidence to decision-makers in charge of making efficient resource allocations within their jurisdictions. Quality-adjusted life year (QALY) is one of a number of measurement units in cost-utility analysis for economic evaluation. QALY stands for both quantity and quality of life. To calculate the value of a QALY, a set of value scores needs to be assigned to each of the various health states indicating weights for quality of life, also known as health-related quality of life (HrQoL). It is recommended that these values be calibrated using social preference weights elicited from the general population [1]. In addition, because the preferences for health states can differ across cultures [2], many countries have measured their own population-based preference weights for all possible health states. Several methods to quantify people’s preferences for health status have been developed; these include visual analog scale (VAS), standard gamble, time trade-off (TTO), and person trade-off methods [3].

Together with EQ-5D [4], there are other preference-based health status measures that can be used to classify the health state of individuals and summarize the change of health outcome in a single index score. For example, there are the Health Utilities Index [5], SF-6D [6], and Quality of Well-Being Scale [7]. In Korea, as in many other countries, there is growing interest in EQ-5D due to the increasing need of measuring the change in HrQoL as an outcome of the health care program. The Korean version of EQ-5D has been under development for some time. Its reliability and validity has already been proven [8] and it was included in the Korea National Health and Nutrition Survey, designed to measure population health in 2005.

In order to develop a population-based preference weights for EQ-5D (also known as EQ-5D value set), a valuation study was conducted, in which a subset of health states defined by the EQ-5D descriptive system was directly valued. Based on these observed values, a regression modeling approach is adopted to exploit values for all 243 health states defined by EQ-5D. It must be noted here that there appears to be reported in the literature only one earlier study that attempted to develop the EQ-5D value set for the population in South Korea [9]. However, due to drawbacks in the design of its valuation study and modeling, the sample was not nationally representative and the average of absolute differences between observed and estimated scores was as great as 0.071. To the authors’ knowledge, to this day the demand for a representative and reliable EQ-5D value set for South Korean population is still not met.

The current study establishes the South Korean population-based preference weights for EQ-5D based on the values elicited from a national representative sample using the TTO method. One of the main features of the survey where the preference data were collected is the number of health states involved in the study. Unlike previous valuation studies performed in Korea or in other countries, where either 43 health states defined by EQ-5D or less were directly valued, here the values for a total of 101 EQ-5D health states have been directly observed. Thus, with this unique dataset it is expected that the interpolation spaces in estimating a value set are minimized in comparison to other value sets.
Methods

Study Subjects

The target population for the study is Korean adult citizens, aged 20 and older, estimated at 36.786 million based on the official residential registries on December 31, 2006 [10]. A multistage stratified random sampling was employed aiming at generating a sample representing the age and sex distribution in the target population. Due to limited resources, the target sample size was restricted to 1307. The sampling procedure is explained below.

In the first step, the entire sample was stratified using 15 regions (seven large cities and eight provinces) with the exception of Jeju province, which is an island with a population number equivalent to 1.2 % of the total population. Due to the relatively small number of residents in this province, its exclusion was expected to have only a limited impact on the sampling. The number of subjects was assigned to 15 regions in proportion to the population size of each region. The same process was subsequently repeated within each region using three categorized administrative units: “Dong,” “Eup,” and “Myun” (“Dong” is a town in a district of a city, “Eup” is a main town in a county, and “Myun” is a township in a country; every address can be categorized into one of these units). In the second step, the final field-work locations “Ban” and “Village” (“Ban” is a subdivision of “Dong” or “Eup,” and “Village” is a subdivision of “Eup” or “Myun”) were randomly selected within the strata defined in the first step. In the third step, 8 to 10 households were randomly selected for interview in each “Ban” or “Village.” In those cases where a selected household had more than two persons aged 20 years or more, the interviewers invited the person whose birthday was closest within the next 12 months to the day of interview. Persons residing temporarily at a selected household, such as a lodger, family member in military service, and persons in long-period official trips or overseas duty were excluded.

EQ-5D

EQ-5D is one of the most widely used generic index measures of HrQoL [4]. It consists of two parts, the EQ-5D descriptive system and the EQ-5D VAS. The descriptive system contains five items that measure five dimensions of health including mobility, self-care, usual activities, pain/discomfort, and anxiety/depression. Each dimension is represented by a single item with three levels of responses: no problem, some problems, and extreme problems. A total of 243 health states are defined by this descriptive system. The EQ-5D VAS records the respondent’s self-rate health status on a VAS, where the endpoints are labeled “best imaginable health state” and “worst imaginable health state.” As mentioned earlier, the Korean version of EQ-5D has already been developed, and its validity and reliability has been proven [8].

Health State Selection

The survey included 100 EQ-5D health states together with states “33333” and “11111” for direct valuations. The 100 health states chosen comprise 25 mild, 50 moderate, and 25 severe states. The degree of severity was defined by a standard city-block distance metric in which any movement away from “11111” is simply counted for each dimension and aggregated. For instance, state “11121” and state “21113” are categorized into distance groups 1 and 3, respectively. Mild states are those within the distance groups 1 to 4, where there are no level 3 problems and up to three level 2 problems. Severe states are those within the distance groups 7 to 9, in which there are no level 1 problems and at least two level 3 problems. If a state is neither mild nor severe, then it is classified as a moderate state. For example, although state “21113” is in distance group 3, it would be categorized not as mild but as a moderate state due to having a problem of level 3. The 100 health states selected were distributed into 25 blocks (P. Kind, pers. comm.). To ensure that each block contained health states across different severities, each block had six health states composed of two randomly selected mild states, two severe states, and two moderate states. In the current study, each participant was assigned two blocks; one was picked following the numerical order assigned to 25 blocks (i.e., the 100th respondent evaluated the 25th block) and the other was randomly selected. Thus, each participant evaluated 12 health states from two blocks, in addition to the “11111” and “33333” states. The selection of health states for each of the 25 blocks can be found at: http://www.ispor.org/Publications/value/ ViHSupplementary/ViH12i8_Nam.asp.

Data Collection

The survey instruments and protocol used were similar to those of the Measurement and Valuation of Health (MVH) study in the UK [11]. The details of the survey are as follows:

The survey was based on a face-to-face interview that can be divided into three stages. In the first stage, respondents described their own health at the time of the interview using the validated Korean version of EQ-5D, including answering the five-item descriptive system and self-rated VAS. In the second stage, the respondents were asked to rank the 12 health states from the two blocks assigned plus the states “11111” and “33333” by putting the “best” health state on top and the “worst” at the bottom. It was assumed that each state was experienced for 10 years followed by immediate death. Subsequently, respondents rated the above ranked 14 health states and the state of immediate death using VAS.

Finally, each respondent evaluated the same set of health states but without state “11111” and immediate death using TTO technique with the double-sided time board and a set of health state cards. The method is also known as TTO props method. A thorough description of the method can be found elsewhere [12] and is therefore not repeated in detail here. In short, the respondents were first asked to decide whether a state is better or worse than death. For states regarded as better than death, respondents decided a period of time t in the state “11111,” which they consider as equivalent to 10 years in the target state. The shorter t is, the worse the target state. For the states worse than death, the choice was between dying immediately and spending a length of time (10 – t) in the target state followed by t years in the state “11111.” Consequently, the longer the time chosen to be in the state “11111” to compensate for a shorter time in the target state, the worse the target state is [12]. In TTO valuation scale, the states “11111” and immediate death were treated as anchors and assigned values of 1 and 0, respectively. Respondents were also surveyed on socioeconomic background questions after completing the TTO valuation.

The data were collected between February 6 and April 3, 2007. A total of 61 trained interviewers were recruited for this purpose. On completing the survey, each respondent was rewarded a gift certificate equivalent to about 10 US dollars.

Logical Consistency and Exclusion Criteria

The logical consistency approach was applied to examine the quality of data. Logical consistency is defined as follows: for a given pair of health states, if state A of a pair is better than the state B in at least one dimension and not worse in any other, then
the valuation for the former state (TTOA) must be at least as good as the valuation for the latter state (TTOb) [13,14]. In a situation where this rule is breached, the logical inconsistency is said to occur. For instance, if state “11122” is valued higher than state “11121,” this is logically inconsistent.

Following the MVH study, respondents were excluded if they valued less than three states, valued all states the same, valued all states worse than dead, or if there were four or more logical inconsistencies. These exclusion criteria take into account responses with incomplete or unreliable data. The choice of threshold level (three inconsistencies) is based on Ohinmaa and Sintonen’s study of the Finnish population [15], where it was found that involving valuation data with more than three inconsistencies would significantly affect a derived value set.

Transforming the Data
The observed times \( t \) for each investigated health state in the TTO valuation task were converted into TTO value \( h \). For the states valued as better than death, the TTO value \( h \) is \( t/10 \). For states worse than death, \( h = -t/(10 - t) \). A linear transformation formula was applied to TTO values for states worse than death in order to bound negative values to a maximum of \(-1\), where 0 is equivalent to death. This was done as follows [16]:

\[
b' = h/39
\]  

(1)

The lowest value \( h \) for a state worse than death is \(-39\). This value occurs when a respondent preferred immediate death over the course of 3 months in the target state, followed by 9.75 years in state “11111.” With the above formula, this minimum value of \(-39\) was transformed to \(-1\).

Statistical Analysis

Modeling method. The analyses were conducted at both aggregate and individual levels. In the aggregate level analysis, the mean is used to summarize the score of each health state and to estimate a value set based on the aggregated means. Both ordinary least square (OLS) and weighted least square (WLS) regressions were applied in the aggregate level analysis. The applied weights in WLS regression takes into account the number of respondents who rated a particular state. On the other hand, at the individual level each respondent’s score for a given health state is introduced into the estimated model. OLS regression and either a random or fixed effects model, depending on the Hausman’s test, were employed in the individual level analysis. The individual effects introduced by participants who might have systematically valued health states higher or lower can be eliminated by applying the fixed/random effects model.

Dependent variables. The dependent variable in the regression analysis was computed as 1 minus the transformed TTO value. It represents the measure of disutility by subtracting the value of a given health state from the value of full health. As a result, the predicted value for state “11111” is equal to 1.

Independent variables. In selecting a model, a range of models from earlier applications were reviewed. The simplest model is the main effects model, which comprises 10 dummy variables that indicate the presence of either a level 2 or 3 in a given dimension of the evaluated state. For instance, M2 for mobility level 2, M3 for mobility level 3, SC2 for self-care level 2, SC3 for self-care level 3, UA2 for usual activities level 2, UA3 for usual activities level 3, PD2 for pain or discomfort level 2, PD3 for pain or discomfort level 3, AD2 for anxiety or depression level 2, and AD3 for anxiety or depression level 3. Unlike the model focusing on only main effects, other models such as the N3 and D1 models also take into account the interaction effect. The N3 model includes the N3 term, indicating whether there is any dimension on level 3, along with 10 main effect variables. The D1 model consists of D1, I2, I22, I3, and I32 terms in addition to 10 main effect variables. The D1 term indicates the number of dimensions with problems beyond the first and replaces the constant term. The I2 term represents the number of dimensions at level 2 beyond the first. The I22 term is the square term of I2. The I3 term represents the number of dimensions at level 3 beyond the first. The I32 term is the square term of I3. The following interaction terms were also considered in the modeling process:

N2: whether there is any dimension on level 2,
C2: the number of dimensions on level 2,
C2sq: the square of the number of dimensions on level 2,
C3: the number of dimensions on level 3,
C3sq: the square of the number of dimensions on level 3,
X2: whether there are 2 or more dimensions on levels 2 or 3,
X3: whether there are 3 or more dimensions on levels 2 or 3,
X4: whether there are 4 or more dimensions on levels 2 or 3, and
X5: whether there are 5 dimensions on levels 2 or 3

Model selection. Given the purpose of this study, the best fitting model is the one that minimizes the difference between the observed and the estimated value in each health state. Hence the overall mean absolute error (MAE) was computed for each investigated model. The number of absolute errors greater than 0.05 and 0.10 in each model were also used as criteria. Note that due to the great variability of responses at the individual level, the goodness-of-fit such as adjusted or overall \( R^2 \) at this level is expected to be lower than that at aggregate level analysis.

Various tests were conducted to examine the assumptions made in the model. The normality of residual was investigated using Kolmogorov–Smirnov normality test and the Breusch–Pagan test for examining heteroskedasticity. When heteroskedasticity was found in OLS regression, the standard errors of coefficients in the model were corrected through the estimation of HC3 robust standard errors as described by Long and Ervin [17]. Hausman’s test was used to decide between random effects and fixed effects model. The Ramsey regression equation specification error test (RESET) test for model misspecification was also examined.

To examine the robustness of the chosen model, respondents were randomly split into two half samples. The coefficients were first estimated from one half and then used to generate estimated scores and which were then compared with the observed scores in the other half sample. The final value set is then based on the whole valuation sample.

Comparison of final model with other studies. The mean observed TTO values for EQ-5D health states obtained in the current study were compared using \( t \)-tests to those published by
Jo et al. [9]. For comparability with Jo et al.’s study, the TTO values for states worse than death were based on monotonic transformation. Furthermore, the coefficient estimations from the main effects model in Jo et al.’s study [9] were compared with those from the current data set (also with monotonic transformation) using the same model.

The estimated value set in the current study was compared with the established value set in other countries such as Japan, UK, and USA, as well as with the estimation obtained in the previous Korean study. For this purpose, Spearman’s rank correlation coefficients and mean absolute differences (MADs) between estimated values from the current study and those from others were calculated.

**Results**

**Respondent Sample and Valuation Sample**

Of the 1307 respondents, a total of 233 (17.8%) had logical inconsistencies, of which 39 had four or more inconsistencies and were excluded from the sample. In addition, four other respondents were also excluded: two subjects who gave the same values for all 13 states measured; and another two who valued all states as the state worse than dead. As a result, a total of 1264 respondents formed the valuation sample.

Despite the exclusion of 43 participants, as shown in Table 1 the age and sex distribution of the valuation sample was representative of the Korean population. Table 2 presents the socio-demographic and self-reported health characteristics in the total sample and in the valuation samples. There were no significant differences between the two samples for variables such as education, religion, marital status, experience of chronic condition, and self-reported health problems measured by EQ-5D.

**Modeling Analysis**

Table 3 presents the coefficient estimates and fit statistics results for the aggregate level models using OLS and WLS regression. Only the main effects, N3 and D1 models, are reported here because other models with different interaction terms did not perform better than the above three models. The main effects model based on OLS had an MAE of 0.031 and the number of absolute errors greater than 0.05 and 0.10 were 20 and 1, respectively. All the coefficient estimations were theoretically consistent, having the expected sign and magnitude. The D1 model included significant interaction terms, D1 and I3, only. Despite

| Characteristics     | Total sample (n = 1307) | Valuation sample (n = 1264) | P-value*
|---------------------|-------------------------|----------------------------|--------
| Education, years (%)|                         |                            |        |
| 6 or less           | 10.25                   | 10.05                      | 0.116  |
| 7–12                | 52.11                   | 52.61                      |        |
| 13 or more          | 37.64                   | 37.34                      |        |
| Religion (%)        |                         |                            |        |
| Buddhist            | 25.79                   | 25.71                      | 0.958  |
| Christian or Catholic| 33.82                | 33.86                      |        |
| Others              | 0.38                    | 0.40                       |        |
| Unbeliever          | 39.17                   | 39.16                      |        |
| No answer           | 0.84                    | 0.87                       |        |
| Marital status (%)  |                         |                            |        |
| Married             | 72.61                   | 72.55                      | 0.941  |
| Single              | 23.87                   | 23.89                      |        |
| Widowed             | 2.75                    | 2.77                       |        |
| Divorced/separated  | 0.77                    | 0.79                       |        |
| Experience of chronic condition (%) |            |                            |        |
| Yes                 | 10.41                   | 10.36                      | 0.990  |
| No                  | 89.59                   | 89.64                      |        |
| In EQ-5D, those reporting problems on (%) |     |                            |        |
| Mobility            | 5.89                    | 5.85                       | 1.000  |
| Self-care           | 0.77                    | 0.79                       | 1.000  |
| Usual activities    | 4.05                    | 4.11                       | 0.848  |
| Pain/discomfort     | 21.27                   | 21.04                      | 0.372  |
| Anxiety/depression  | 17.44                   | 17.09                      | 0.100  |

*Comparison between valuation sample and excluded respondents by Chi-square test.

![Image](image_url)
the different interaction terms involved, the N3 and D1 models produced identical results with an MAE of 0.029 and the number of states with absolute errors greater than 0.05 and 0.10 as 15 and 0, respectively. All coefficient estimations in the N3 model had positive signs. In contrast, the two interaction terms in the D1 model, the D1 and I3 terms, had negative signs. The negative sign implies a higher value for a health state with more severe problems. The results of the main effects, N3 and D1 models using WLS regression, were generally worse than those based on OLS regression. Particularly, the number of states with absolute error greater than 0.05 in the main effects, N3 and D1 models using WLS regression, were 18, 22 and 20, respectively. A possible cause could be the outweighed number of values for state “333333” (n = 1264, SD = 0.44) and, as a result, a greater weight assigned to this particular state.

Considering the consistency of the coefficient estimations and minimizing the difference between observed and estimated values, the N3 model based on OLS regression was selected as the best performing model at the aggregate level. It passed the Kolmogorov–Smirnov test for normality of the residuals (D = 0.083, P = 0.497). There was no model or functional form misspecification as suggested by the Ramsey RESET test (F = 0.54, P = 0.466). However, it failed the Breusch–Pagan test for heteroskedasticity (F = 4.92, P = 0.026). Theoretically, the number of absolute errors greater than 0.05 was significantly increased in the results of applying WLS regression as compared to those resulting from the OLS. Therefore, the N3 model based on WLS regression is still preferred. The HC3 procedure was used to correct the biased standard error of coefficient in the OLS.

The results of modeling at the individual level are shown in Table 4. Based on OLS regression, four models were identified, including the main effects, N3, D1, and X5 models. The D1 model here includes I3, I32, and D1 terms. The MAE was 0.031 for the main effects model and 0.030 for the rest. The number of absolute errors greater than 0.05 and 0.10 were 18, 22, 20, and 19, and 1, 0, and 2, respectively. All models generated theoretically consistent coefficient estimations, apart from the D1 model where the I3 and D1 terms had negative signs. The Hausman test rejected the random effects model and in favor of the fixed effects model (F = 116.89, P < 0.001), therefore only results of the fixed effects model are reported. There are two models, main effects and N3, which are based on fixed effects regression and are presented here. The main effects model had an MAE of 0.032 and the number of absolute errors greater than 0.05 and 0.10 were 20 and 2, respectively. The N3 model produced similar results: the MAE was 0.031 and the number of absolute errors greater than 0.05 and 0.10 were 20 and 2, respectively.

The differences between the compared models at the individual level were marginal and therefore it was difficult to select one as the best performing. Also most of the models at the individual level showed significant heteroskedasticity, non-normality of error distribution and model misspecification. For example, the N3 model failed in each of the following tests: Kolmogorov–Smirnov normality, Breusch–Pagan and Ramsey RESET (D = 0.086, P < 0.001; F = 1954.99, P < 0.001; F = 17.90, P < 0.001, respectively).

According to the findings, the N3 model based on OLS regression with aggregate data is the best fitting model, minimizing the difference between the observed and the estimated value in each health state, and is thus chosen as the final model to estimate the value set. When the model robustness was examined by comparing the estimated values from one half sample and the observed values from the other half, both values were highly correlated (r = 0.983) with an MAE of 0.040.

Predicted values are calculated using the final model. For example, we calculated the predicted values of state “32322” as follows:

\[
\text{Predicted value} = \text{full health} - \text{disutility}
\]

Full health = 1.000

Disutility for 32322 state = 0.050 + 0.418 (M3) + 0.046 (SC2) + 0.208 (UA3) + 0.037 (PD2) + 0.043 (AD2) + 0.050 (N3) = 0.852.

Predicted values = 1 – 0.852 = 0.148
### Table 4 Parameter estimates and fit statistics of individual level models using OLS and fixed effect regression

<table>
<thead>
<tr>
<th>Variable</th>
<th>Main effects N3</th>
<th>D1</th>
<th>X5</th>
<th>Main effects N3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Coefficient</td>
<td>SE</td>
<td>Coefficient</td>
<td>SE</td>
</tr>
<tr>
<td>Constant</td>
<td>0.061</td>
<td>0.007</td>
<td>0.060</td>
<td>0.007</td>
</tr>
<tr>
<td>M2</td>
<td>0.087</td>
<td>0.006</td>
<td>0.084</td>
<td>0.006</td>
</tr>
<tr>
<td>M3</td>
<td>0.422</td>
<td>0.008</td>
<td>0.415</td>
<td>0.008</td>
</tr>
<tr>
<td>SC2</td>
<td>0.044</td>
<td>0.006</td>
<td>0.043</td>
<td>0.006</td>
</tr>
<tr>
<td>SC3</td>
<td>0.164</td>
<td>0.007</td>
<td>0.158</td>
<td>0.008</td>
</tr>
<tr>
<td>UA2</td>
<td>0.050</td>
<td>0.006</td>
<td>0.047</td>
<td>0.006</td>
</tr>
<tr>
<td>UA3</td>
<td>0.221</td>
<td>0.007</td>
<td>0.213</td>
<td>0.008</td>
</tr>
<tr>
<td>PD2</td>
<td>0.039</td>
<td>0.006</td>
<td>0.036</td>
<td>0.006</td>
</tr>
<tr>
<td>PD3</td>
<td>0.175</td>
<td>0.008</td>
<td>0.168</td>
<td>0.008</td>
</tr>
<tr>
<td>AD2</td>
<td>0.046</td>
<td>0.006</td>
<td>0.042</td>
<td>0.006</td>
</tr>
<tr>
<td>AD3</td>
<td>0.182</td>
<td>0.007</td>
<td>0.173</td>
<td>0.008</td>
</tr>
<tr>
<td>I3</td>
<td>0.027</td>
<td>0.009</td>
<td>0.027</td>
<td>0.009</td>
</tr>
<tr>
<td>I3</td>
<td>-0.102</td>
<td>0.016</td>
<td>0.102</td>
<td>0.016</td>
</tr>
<tr>
<td>D3</td>
<td>0.012</td>
<td>0.002</td>
<td>0.012</td>
<td>0.002</td>
</tr>
<tr>
<td>X5</td>
<td>-0.048</td>
<td>0.009</td>
<td>0.031</td>
<td>0.010</td>
</tr>
<tr>
<td>MAE</td>
<td>0.033</td>
<td>0.033</td>
<td>0.033</td>
<td>0.033</td>
</tr>
<tr>
<td>No. (of 101) &gt;0.05</td>
<td>18</td>
<td>22</td>
<td>20</td>
<td>19</td>
</tr>
<tr>
<td>No. (of 101) &gt;0.1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>2</td>
</tr>
</tbody>
</table>

0.001 < *P < 0.01; otherwise P < 0.001.
†HC3 robust standard error.
M2, mobility level 2; M3, mobility level 3; SC2, self-care level 2; SC3, self-care level 3; UA2, usual activities level 2; UA3, usual activities level 3; PD2, pain or discomfort level 2; PD3, pain or discomfort level 3; AD2, anxiety or depression level 2; AD3, anxiety or depression level 3; N3, any dimension on level 3; I3, the number of dimensions at level 3 beyond the first; X5, whether there are 5 dimensions on levels 2 or 3; MAE, mean absolute error; OLS, ordinary least square; FE, fixed effect regression; SE, standard error.

The observed and predicted means and the difference between the two values can be found at: http://www.isPOR.org/Publications/value/ViHsupplementary/ViH12i8_Nam.asp.

**Comparison with Previous Korean Study and Other Studies**

Among 23 health states in common, nine health states had significantly lower observed means in the current study than in Jo et al.’s [9], including five severe and two moderate states. Such differences are translated into the coefficient estimates in the model. With the same transformation method for a state worse than death and model specification, all coefficient estimations for level 3 are higher for all dimensions apart from self-care in the current study. The biggest magnitude is observed in the mobility dimension with almost a twofold increase (from 0.310 in Jo et al.’s to 0.606 here).

In comparison with other studies, the value set obtained from our final model is highly correlated with the official value set in Japan (r = 0.969, P < 0.001), USA (r = 0.908, P < 0.001), and UK (r = 0.855, P < 0.001), respectively. The MAD between our Korean study and Japan is 0.056, with USA it is 0.105 and with the UK it is 0.322.

**Discussion**

This study collected TTO values for 101 EQ-5D health states from a South Korean representative sample. Based on these values the population-based preference weights for EQ-5D are developed using the N3 model. This model yields the best fit for the observed TTO value at aggregate level, with an MAE of 0.031 and only 15 (out of a total 101) prediction errors exceeding 0.05 in absolute magnitude.

At the aggregate level, despite the D1 model producing identical results to the N3 model, the negative sign of coefficient estimations for both interaction terms in the D1 model make it less transparent in calculation. It also becomes conceptually difficult to understand why, for instance, health states with more level 3 problems result in an increased value. Thus, the N3 model is preferred.

The empirical comparisons between modeling at aggregate and individual levels in the current study support the use of aggregate level analysis. However, the choice of either aggregate or individual level based analysis is an ongoing debate. Advantages associated with the aggregate level approach include simple modeling, easy interpretation, and being intuitive. On the other hand, the advantages associated with the individual level approach include utilizing the maximum amount of information and treating each respondent’s value on an equal basis. Theoretically, individual level analyses might be expected to produce better results with their capacity to adjust for individual effects. However, in practice it is commonly found that there is too much noise in individual level data that hinders the performance of the estimates. In contrast, aggregate level analysis can alleviate such a problem by regressing at aggregate measures to minimize the unwanted variations.

The choice of central tendency measures, such as the mean or median in the aggregate level analysis, is a debatable issue and the exploration of the impact of the choice of central tendency is beyond the scope of this study.

Compared with other valuation studies, the major contribution of the current study is the number of health states that were directly valued. Unlike other studies, either following the 43 EQ-5D health states in the MVH project or decreasing the number of health states investigated to fewer than 43, this study increases the number of health states. A total of 101 health states were valued, of which 23 overlap with the health states in the MVH set. Therefore, there are at least 2.4 times more health states investigated than in other studies, covering almost 42% of the total health states (101/243) defined by the EQ-5D descriptive system. As a result, this study provides more information regarding how values (observed) are distributed in the valuation system. As a result, this study provides more information regarding how values (observed) are distributed in the valuation system. As a result, this study provides more information regarding how values (observed) are distributed in the valuation system. As a result, this study provides more information regarding how values (observed) are distributed in the valuation system. As a result, this study provides more information regarding how values (observed) are distributed in the valuation system.
South Korean Valuation Study

space defined by EQ-5D, and consequently it limited the interpolation space in the estimations.

There are three possible ways to transfer the TTO value for states worse than death: monotonic, linear, and truncated transformations [16]. The choice of transformation method in our study was purely based on empirical evidence showing that the linear transformation results in the smallest MAE amongst the three methods. There is no theoretical ground for the choice of one method over another. However, there should be awareness of the effect of applying different transformations in the resulting EQ-5D value set and consequently on the cost-effectiveness analysis. For instance, the value set based on linear transformation produces a smaller range of values and therefore the QALY estimation, and possibly QALY gain, will be smaller than those estimated from a value set based on monotonic transformations.

A possible contribution to the discrepancies in observed TTO values and, consequently, coefficient estimation between the current study and Jo et al.’s [9] is the sampling difference. The sampling in our study is from 15 regions representing the whole country (except the Jeju region), while in the latter study it was confined to two adjacent regions only (Seoul and Gyeonggi-do). Our data suggests that the values obtained from the other 13 regions are different from values elicited from Seoul and Gyeonggi-do regions (data not shown). Therefore, the values elicited from these two regions alone cannot be used as a representative preference for the population in South Korea as a whole.

Another possible explanation for differences in coefficient estimations between the two studies could be the number of health states involved. In our study there are 101 health states with directly observed values, whereas Jo et al. [9] use only 42 states. In other words, in our study there is more information available regarding the valuation space defined by EQ-5D, and therefore it minimizes the interpolation spaces in the estimation. Particularly, this study values 26 severe health states directly. In other words, in our study there is more information available regarding the valuation space defined by EQ-5D, and consequently it limited the interpolating space and possessing better value set derived here is based on a population representative sample, limiting the interpolating space and possessing better model performance. Thus, this EQ-5D value set should be used preferentially for the South Korean population.

In conclusion, the study successfully establishes a set of South Korean population-based preference weights for the EQ-5D. The value set derived here is based on a population representative sample, limiting the interpolating space and possessing better model performance. Thus, this EQ-5D value set should be used preferentially for the South Korean population.

Source of financial support: This study was prepared by financial support of Korea Centers for Disease Control and Prevention (KCDC) in the Ministry of Public Health and Social Welfare.

References


International Comparisons in Valuing EQ-5D Health States: A Review and Analysis

Richard Norman, MSc,1 Paula Cronin, MPH,1 Rosalie Viney, PhD,1 Madeleine King, PhD,2 Deborah Street, PhD,1,3,4 Julie Ratcliffe, PhD5

1Centre for Health Economics Research and Evaluation (CHERE), Faculty of Business, University of Technology, Sydney, NSW, Australia; 2Psycho-oncology Co-operative Research Group (PoCoG), University of Sydney, Sydney, NSW, Australia; 3Centre for the Study of Choice (CenSoc), Faculty of Business, University of Technology, Sydney, NSW, Australia; 4Department of Mathematical Sciences, Faculty of Science, University of Technology, Sydney, NSW, Australia; 5Flinders Clinical Effectiveness, Flinders University, Adelaide, Australia

ABSTRACT

Objective: To identify the key methodological issues in the construction of population-level EuroQol 5-dimensions (EQ-5D)/time trade-off (TTO) preference elicitation studies.

Method: This study involved three components. The first was to identify existing population-level EQ-5D TTO studies. The second was to illustrate and discuss the key areas of divergence between studies, including the international comparison of tariffs. The third was to portray the relative merits of each of the approaches and to compare the results of studies across countries.

Results: While most articles report use of the protocol developed in the original UK study, we identified three key areas of divergence in the construction and analysis of surveys. These are the number of health states valued to determine the algorithm for estimating all health states, the approach to valuing states worse than immediate death, and the choice of algorithm. The evidence on international comparisons suggests differences between countries although it is difficult to disentangle differences in cultural attitudes with random error and differences as a result of methodological divergence.

Conclusions: Differences in methods may obscure true differences in values between countries. Nevertheless, population-specific valuation sets for countries engaging in economic evaluation would better reflect cultural differences and are therefore more likely to accurately represent societal attitudes.

Keywords: cost-utility analysis, EQ-5D, health economics methods, health-related quality of life.

Introduction

Cost-utility analysis (CUA), where outcomes are measured in terms of quality-adjusted life-years (QALYs), is the main approach used to measure and value the impacts of treatments. The US Panel on Cost-Effectiveness in Health and Medicine recommends the use of QALYs [1]; the UK National Institute of Health and Clinical Excellence has most commonly used CUA [2,3] and has recently recommended that it should be the preferred outcome measure; and CUA is increasingly used in Australia in the evaluation of pharmaceuticals and medical services. In the recently released PBAC guidelines, a preference is expressed for the use of CUA [4].

While CUA is simple in concept, it presents challenges in practice. QALYs are designed to allow comparisons across interventions with disparate outcomes across different health-care conditions and population groups. Eliciting valuations for all health states that may be relevant to a disease or intervention is time consuming and costly, and comparison of valuations across interventions and diseases requires comparability of methods. Multittribute utility instruments (MAUIs), which comprise a generic descriptive quality of life instrument and a scoring algorithm that covers all health states described by the instrument (e.g., the EQ-5D, the Short Form-6 dimensions (SF-6D), Health Utilities Index, and Assessment of Quality of Life), have facilitated comparability [5,6]. The scoring algorithm for these instruments is usually generated from a stated preference experiment, typically time trade-off (TTO), standard gamble conducted in a population sample. The key advantage of the MAU approach is that it provides community-based valuation of health states for patients who are experiencing the state.

The role of MAUIs in economic evaluation is increasing. For example, the National Institute of Clinical Excellence has recommended the use of the EQ-5D, and the Pharmaceutical Benefits Advisory Committee in Australia has stated a preference for utility weights generated from the use of a MAUI in a clinical trial setting (without specifying a preference for a particular MAUI). Nevertheless, recent reviews have noted that there are significant differences in the performance of different MAUIs [7], which can be attributed to differences in the dimensions covered by the instruments, differences in preference elicitation techniques, and differences in the methods used to derive the scoring algorithm. These differences can have significant impact on valuations of health states and the resulting cost-effectiveness of interventions [8]. There has been relatively little critical appraisal of the methods of development of MAUIs scoring algorithms. In this article, we examine these issues by considering the EQ-5D [9]. We chose the EQ-5D because it is widely used, and there have been a number of different studies undertaken to develop country-specific scoring algorithms. Because the focus of this review is on one MAUI, we do not consider the psychometric aspects of the instrument but, rather, focus on the methods for development of the scoring algorithm. Many of the issues we raise are relevant to other MAUIs.

Overview of the EQ-5D

The EQ-5D is a tool developed by the EuroQol group (Rotterdam, The Netherlands) (www.euroqol.org) and has five dimensions
intended to represent the major areas in which health changes can manifest. These areas are mobility, self-care, usual activities, pain/discomfort, and anxiety/depression. Each dimension contains three levels, classified as “No Problems,” “Some Problems,” and “Extreme Problems.” Details are shown in Table 1. Thus, there are $3^5 = 243$ potential states in the descriptive system. The TTO approach is used to value a selection of these states and then to impute values for the remainder using simple regression. The use of TTO for valuing EQ-5D states is well described in other works [10,11].

Regarding Table 1, it should also be noted that we will treat health states with the same levels as identical throughout this article (e.g., health state 12321 is the same irrespective of language). As of March 2009, the EQ-5D has been translated into 100 different languages (with a further 24 awaiting ratification). As of March 2009, the EQ-5D has been translated into 100 different languages (with a further 24 awaiting ratification). The comparability of versions is a reasonable assumption because all translations are reviewed by EuroQol Group members and ratified by the EuroQol Group Executive Committee. The Euroqol Web site states that translation consists of two forward translations of the EQ-5D English source version, two back translations, lay assessment, and the production of a full report describing each stage of the process (http://www.euroqol.org).

Our analysis of this EQ-5D/TTO approach involves two strands: first, we look at how to elicit societal valuations for EQ-5D states under the York Research Group on the Measurement and Valuation of Health TTO protocol [9]. We begin by identifying some key themes and issues that run across the population valuation studies. Then, we look at international comparisons and discuss whether it is necessary to provide nationality-specific tariffs for the EQ-5D valuation system.

### Methods

The initial target of this study was to identify all large general population valuations studies employing the EQ-5D as the tool for describing health. EMBASE and MEDLINE were searched for such articles. To be considered for inclusion, the analysis had to present primary research in English and be published since 1995. Because it was expected that a proportion of good quality reports may be unavailable in peer-reviewed publications, the reference lists of articles identified in the main search were used to identify further studies. Because all of these identified nonpeer-reviewed publications were available on the EuroQol Web site (http://www.euroqol.org), the list of EuroQol Plenary Meeting Proceedings was scanned for further studies relevant to this work. To be included, a study had to attempt to value all 243 states described by the EQ-5D. Beyond this constraint, we chose to be conservative in our approach to exclusion because we were seeking to identify divergence in approach.

For each included study, details most relevant to the analysis of the methods used were identified. Key areas for discussion were selected. These areas were the precise formulation of the algorithm, the number of states directly valued in the survey to generate weights, the method to value states worse than death, the influence of time preferences of results, and international comparisons in predicted values across EQ-5D space.

The algorithms were compared by expanding the approach used by Busschbach et al. [12], who compare the directly valued states in the UK, Germany, and Spain. For this, Busschbach et al. used the UK results as the benchmark. The predicted preference scores for the states under the UK algorithm were then ranked in descending order. The preference scores under each of the other algorithms are generated by using the same ordering as in the UK study. We extended this approach by including all identified algorithms. Thus, we can identify any tendency for countries to trade off quantity of life for quality of life, and identify whether countries differ in their relative valuations of the five dimensions.

### Results

10 articles [11,13–21] that met the inclusion criteria were identified, of which eight were published in peer-reviewed journals. These are listed in Table 2. It should be noted that there are, at present, no such results for Canada or Australia, two countries strongly supportive of the use of CUA in health-care decision-making. Two studies utilized the visual analog scale (VAS) as the primary method of valuation [14,15]. Although this technique is widely used in preference elicitation more generally, the age of the two VAS studies in this area suggests that it has been superseded by the TTO although work by Parkin and Devlin suggests that the VAS remains a valuable tool [22].

Three significant methodological differences emerged regarding the survey structure and the development of the algorithm. The first regarded the number of states that need to be directly valued to estimate valuations for the complete EQ-5D space. The second is the approach to valuing states considered to be worse than death. The third is the choice of the algorithm to model those states not directly valued. There were a number of additional issues that might also be considered such as the validity of the TTO method and the assumption of constant proportional trade-off that it is founded on. Nevertheless, it was felt that this had been adequately covered elsewhere [23,24,25].

### The Number of Directly Valued States

Given that the EQ-5D has 243 individual possible states, it is unsurprising that no study has attempted to ask respondents to directly value each of these states. Therefore, the pertinent question becomes how best to form a representative fraction of the entire space that allows a good estimation of the remainder of the EQ-5D states in whichever way that is defined. Two approaches have been adopted to form this representative fraction. The
Table 2: Identified studies

<table>
<thead>
<tr>
<th>Study</th>
<th>Country</th>
<th>Sample size</th>
<th>Age limit/range</th>
<th>States directly valued</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dolan et al. [9]</td>
<td>UK</td>
<td>3395</td>
<td>&gt;18, 13 from 43</td>
<td>Time trade-off</td>
</tr>
<tr>
<td>Badia et al. [1]</td>
<td>Spain</td>
<td>975</td>
<td>Unspecified</td>
<td>Time trade-off</td>
</tr>
<tr>
<td>Bjork and Norinder [14]</td>
<td>Sweden</td>
<td>300</td>
<td>Range between 18 and 75, including at least one of each severity level</td>
<td>Time trade-off</td>
</tr>
<tr>
<td>Greiner et al. [17]</td>
<td>Germany</td>
<td>339</td>
<td>15, 13 from 43</td>
<td>Time trade-off</td>
</tr>
<tr>
<td>Jelsma et al. [17]</td>
<td>Zimbabwe</td>
<td>2488</td>
<td>Range between 15 and 91, including at least one of each severity level</td>
<td>Time trade-off</td>
</tr>
<tr>
<td>Shaw et al. [23]</td>
<td>USA</td>
<td>4048</td>
<td>Range between 18 and 99.3, including at least one of each severity level</td>
<td>Time trade-off</td>
</tr>
<tr>
<td>Devlin et al. [9]</td>
<td>New Zealand</td>
<td>2741</td>
<td>&gt;18, 12 other states</td>
<td>Time trade-off</td>
</tr>
</tbody>
</table>

As expected, the MAE is negatively associated with both the sample size and the number of health states directly valued. Additionally, they contrast these data with the results of Dolan et al. [26], which suggest that not only does the 17-state approach used by Tsuchiya et al. [11] lead to a lower MAE than that of Dolan et al. but also it may lead to a lower MAE than if each respondent valued 17 (or even 22) randomly assigned states from the 42 (although the difference does not appear to be statistically significant). The mean correlation for the predicted and actual values if 22 states from 42 are randomly selected is 0.986 (SD = 0.006), whereas the figures for the 17 states used by Tsuchiya et al. was 0.989 (SD = 0.002).

A related question concerns whether the 17- and 43-state approaches are optimal in terms of study design. For equal precision in each of the effect estimates to be allowed, it is necessary to have equal frequency of appearance for each of the levels for each of the attributes. Because there is a disproportionate number of the better health states, that is, states with attributes at level 1, in the 43 Dolan states [9] or the 17 Tsuchiya states [11], there is greater precision at that healthy end of the scale. The other related issue involves the estimation of interactions. Although only 10 degrees of freedom are required for the estimation of main effects, a further 40 are required to estimate two-factor interactions. Of course, if certain level combinations do not appear together (and perhaps do not make sense together), then estimation of all two-factor interactions becomes impossible.

### Transformation of Values for Worse than Dead States

Although it is plausible that the poorer states in the EQ-5D might be considered worse than immediate death, certain methodological issues arise from generating an algorithm with a subset of states that includes states worse than death. While anchoring death at 0 and full health at 1 gives meaning to states that lie in that range, it is difficult to interpret different values below 0. The lack of a tool that is well suited to this task means that existing articles have taken a variety of approaches to valuing these states, some of which raise further questions.

All articles begin from the same starting point, by asking respondents to choose between immediate death and a period of 10 years of life, some of which is spent in the state worse than death and some in full health. In the majority of articles [11,13,16,17,19,21], if the individual is indifferent between immediate death and x years of the bad state followed by (10 – x) years of full health, the score for the state worse than death is then calculated in the following way:

\[
\text{Preference score (State worse than death)} = (x/10) - 1 \tag{1}
\]
Because $x$ is bounded by 0 and 10, the preference scores for states worse than death are bounded by 0 and $-1$. The one divergence from this orthodoxy is found in Shaw et al. [20], for whom

Preference score (State worse than death) = $x/(10 - x)$  \hspace{1cm} (2)

They allowed the value for $x$ to be between 0.25 and 9.75 years, meaning that the preference score is initially as low as $-39$. This leads to an asymmetry between states better than immediate death and those worse. This is important because it means that the impact of a brief period in the severest health state is of the same magnitude as a much longer period in full health. Although a poor state such as this might be plausible, it could be argued that the uncertainty surrounding interpretation of states worse than death means that the value we place on these states should not have a dominant influence on the final algorithm. Shaw et al. suggested that states worse than death should be bounded by $-1$. Thus, they applied a linear transformation to the raw scores, constraining all scores to be in this range [20]. The major problem with this linear transformation is that the valuations in this range are dependent on the minimum length of time the respondent is allowed to endure in the bad health state. If the minimum period allowable in the poor health state increases to, for example, 1 year, all negative values would be divided by nine. The effect of dividing the different health valuations by different factors (defined by the shortest allowable period in the poor health state) increases to, for example, 1 year, all negative values would be divided by nine.

As health moves away from 0 toward $-1$, the effect of this procedural variable becomes increasingly large and suggests that this divergence from the orthodox position is not justified.

The Construction of the Algorithm

The choice of the algorithm is intrinsically associated with the states directly valued in the TTO. Equal precision around point estimates of main effects depends on equal appearance frequency for each of the levels, which does not occur in the states valued in any of the international articles. Equally, for interactions between levels to be estimated, most of which are plausible, these interactions have to appear in the states directly valued, which certainly are not the case for all pairs of levels. In choosing an algorithm, the benchmark UK study [16] prefers the N3 model, in which the algorithm is a main effects model using dummy variables for levels in each dimension worse than “No Problems,” plus the N3 dummy variable, defined as 1 when any of the dimensions are at level 3 (the worst level). Thus, Valuation $= 1 - \sum$(dummyL,d * co-efficientL,d) + (dummyN3 * co-efficientN3).

Aside from increased predictive value of the model with this interaction term [9], the intuition behind using such a value is not clear. Indeed, the Japanese data showed no improvement in model fit after inclusion of the N3 term. One potential explanation for including the N3 term is that the first dimension moving to level 3 will have significant spillover effects, perhaps not captured by the other dimensions. The need to adapt to a life with a severe impediment has a disutility that is a one-off. Thus, the second dimension to move to level 3 will have a disutility (illustrated by the coefficient associated with the respective dummy variable) but may have a lesser impact than if the move had occurred from a state with no pre-existing level 3 problems. The reverse argument, claiming that the N3 term has no intuitive appeal, might argue that the extra predictive value is a remnant of the correction methods used to adjust states worse than death to constrain them between 0 and $-1$. Because these states are considerably more likely to have level 3 dimensions than the general set of states, it is arguable that applying an erroneous transformation, compressing negative values into too small a range, might be identified through lower coefficients being applied to level 3 parameters beyond the first.
Other than the N3 variable, most studies do not utilize interaction terms in their final models. Nevertheless, the intuitive argument in support of interactions can be illustrated by using a number of examples (e.g., the disutility of not being able to do usual activities may vary, depending on whether the person is mobile because this defines what usual activities consist of). A number investigate alternative model specifications containing interactions [19] but generally (and perhaps surprisingly) find that they do not improve the fit of the model [11,13,16,21].

The final issue regarding the algorithm is the use and interpretation of the constant term. Conventionally, the intercept reflects the value of the function when all explanatory variables are 0 (level 1 in the N3 model). Nevertheless, in this case, this interpretation does not hold because 11111 is axiomatically described as full health and is anchored at 1. In the identified articles, there are two approaches in the discussion of the intercept. In the majority of studies, the intercept is allowed to vary from 0 and is interpreted as the disutility associated with not being at perfect health, independent of the disutility associated with the movement within the dimension per se [10]. This could be justified in the same way as the N3 variable was justified above. An alternative approach is taken in a recent US study [20]. The full algorithm used in this study is given by

\[
\text{Valuation} = 1 - (\Sigma (\text{dummy}_{i,0} \ast \text{coefficient}_{i,0}) + \beta_1 \text{DI} + \beta_2 \text{I2-squared} + \beta_3 \text{I3} + \beta_4 \text{I3-squared}),
\]

where \( \text{DI} \) is the number of dimensions not at level 1 beyond the first, \( \text{I2} \) is the number of dimensions at level 2 beyond the first, and \( \text{I3} \) is the number of dimensions at level 3 beyond the first. The differences between this approach and the more commonly utilized N3 approach are that Shaw et al. [20] do not allow a constant term (because full health is anchored at 1) and that they identified a broader group of statistically significant interaction terms, albeit specified in a different way. One criticism of both approaches is that they are relatively blunt in their approach to interactions. For example, if we consider the interactions concerning dimensions being at level 3, the effect of there being a number of dimensions at level 3 is independent of the specific dimensions at that level.

**International Comparisons**

The final question this article considers is the extent to which the use of these different algorithms affects the preference scores associated with the 243 states in EQ-5D space and, thus, whether the choice of model is likely to alter resource allocation decisions. Our results, comparing the wider range of countries using all states defined by EQ-5D space, are shown in Figures 2 and 3 (note that the UK algorithm is smooth because it has been selected as the base case).

We have compared the algorithms to the benchmark in groups of three. When algorithms from Denmark, Germany, and the Netherlands are compared with the UK study, they generate similar preference scores across the range of health states. Generally, they lie above the UK figures but follow the same trend. This suggests that the various dimensions of the EQ-5D have the same approximate relative importance in these countries, but the absolute disutility attached to worsening in the health state in general is estimated to be lower. Regarding the apparent tendency for the UK algorithm to provide health state valuations that are lower than those for other algorithms, it is worth noting that a modified Research Group on the Measurement and Valuation of Health protocol was used in a repeat experiment in a UK population [27], which produced scores generally higher than those derived from the Dolan et al. algorithm [27].

Divergence from this trend can be seen in the countries shown in Figure 3. The Spanish model does not appear to be systematically different from the UK model but displays more variance from the UK model than the Northern European results, suggesting different emphasis between dimensions. The Japanese results are less than those of all other models for mild health states (as a result of a large constant term in the N3 model) but, for worse states, lie above all other models. Under the Japanese model, there are very few states considered worse than death. Additionally, the Japanese results show considerable variance relative to the UK figures. In comparing the Japanese results with the UK, this seems to be the result of a relatively high importance being associated with mobility and a relatively low importance being associated with pain and discomfort, and anxiety and depression. The US study follows a similar pattern to the Japanese results but displays less variability relative to the UK. This unwillingness to trade off quantity of life for quality of life in Japan and the US means that the spread of HRQoL scores is lower in these countries. As noted by Luo et al. and Noyes et al., this will lead to interventions being less cost-effective in CUA because the quality of life gain is likely to be smaller [28,29].
The uncertain element in interpreting these results is to identify whether the differences in models are a result of genuine differences in national attitudes toward ill health or whether they are the product of different study designs (including any difference caused by translation issues). In support of the former is the fact that Figure 2 suggests convergence between countries in a geographic locality (Northern Europe). Nevertheless, we believe that to firmly identify a trend in models between countries, we would require a greater number of studies than currently exist, preferably using data collected by using the same mechanism and at the same time point. The analysis of subgroups within a population is also of potential interest because it may identify what drives health state valuation patterns, both within a population and potentially between populations. Potential explanatory factors might include wealth, income, religion, or health expenditure.

Conclusions

This article identifies a number of key methodological questions in the construction of population-level EQ-5D/TTO value sets. The number of states that need to be directly valued is considered, and the best solution may depend on whether it is worthwhile to look for interaction terms. We identified study design issues with the sets of states most commonly selected to be directly valued. The decision regarding number of states leads into a number of questions regarding the choice of algorithm. Then, we identified competing approaches for the valuation of states considered to be worse than death and identified that the approach used by Shaw et al. [20] makes valuations heavily dependent on a parameter of model design (specifically the minimum period of the state considered in the TTO) that should have no effect on the valuation.

Whether country-specific algorithms are necessary is a difficult question that we have only partly addressed. There are clear divergences between countries in their valuations, in terms of both their willingness to trade quantity of life for quality and their relative importance of the five dimensions of the EQ-5D. Our findings indicate that a proportion of the divergences in algorithms are likely to be attributable to genuine cultural differences rather than methodological differences between studies, which suggests that country-specific algorithms are of importance. This is particularly true in countries that engage in substantial economic evaluation such as Canada and Australia, which are currently reliant on using algorithms derived from countries asserted to be comparably similar in terms of attitude to health.

Source of financial support: Financial support for this study was provided entirely by NHMRC Project Grant (403303). The funding agreement ensured the authors’ independence in designing the study, interpreting the data, writing, and publishing the report.

References


Too Much Ado about Instrumental Variable Approach: Is the Cure Worse than the Disease?

Onur Baser, MS, PhD
STATinMED Research, LLC, and Department of Surgery, University of Michigan, Ann Arbor, MI, USA

ABSTRACT

Objective: To review the efficacy of instrumental variable (IV) models in addressing a variety of assumption violations to ensure standard ordinary least squares (OLS) estimates are consistent. IV models gained popularity in outcomes research because of their ability to consistently estimate the average causal effects even in the presence of unmeasured confounding. However, in order for this consistent estimation to be achieved, several conditions must hold. In this article, we provide an overview of the IV approach, examine possible tests to check the prerequisite conditions, and illustrate how weak instruments may produce inconsistent and inefficient results.

Methods: We use two IVs and apply Shea’s partial R-square method, the Anderson canonical correlation, and Cragg-Donald tests to check for weak instruments. Hall-Peixe tests are applied to see if any of these instruments are redundant in the analysis.

Results: A total of 14,952 asthma patients from the MarketScan Commercial Claims and Encounters Database were examined in this study. Patient health care was provided under a variety of fee-for-service, fully capitated, and partially capitated health plans, including preferred provider organizations, point of service plans, indemnity plans, and health maintenance organizations. We used controller-reliever copay ratio and physician practice/prescribing patterns as an instrument. We demonstrated that the former was a weak and redundant instrument producing inconsistent and inefficient estimates of the effect of treatment. The results were worse than the results from standard regression analysis.

Conclusion: Despite the obvious benefit of IV models, the method should not be used blindly. Several strong conditions are required for these models to work, and each of them should be tested. Otherwise, bias and precision of the results will be statistically worse than the results achieved by simply using standard OLS.

Keywords: asthma, instrumental variable, propensity score, regression analysis.

Introduction

Causal inference is challenging in all nonexperimental studies because of the possibility of overt and hidden bias [1]. When evaluating certain treatment programs, overt bias can exist because the treatment and control groups are different in terms of certain observable factors, such as age, gender, and comorbidities. Hidden bias may exist as a result of failure to control for unobservable factors, such as doctors’ practice/prescription patterns [2]. Practice patterns might be based on physician rules of thumb, experiences and interactions with patients and colleagues as well as comprehensive empirical evidence. Prescription patterns are influenced by nonclinical factors. For example, health plans provide different financial and nonfinancial incentives to doctors or patients to undergo aggressive treatment [3]. It is not feasible to measure all of these characteristics in observational data.

Propensity score matching and regression analysis are two statistical techniques used to remove overt bias. Although regression analysis is widely used in applied economics literature, propensity score methods are increasingly used in medical literature. A systematic literary search by Stürmer et al. found that the annual number of publications using propensity score methods increased from 8 to 71 from 1998 to 2003 [4]. Last year, the number of propensity score methods used was 171.

Baser describes the conditions outlining which method is optimal for controlling for observable bias [2]. Although effectively controlling for observable bias, neither propensity score matching nor regression adjustment addresses problems because of imbalances in unmeasured factors. For this reason, interest in instrumental variable (IV) approach is growing.

Notions of causality in econometrics and their relationship with IVs and other methods are discussed in Heckman [5]. Econometric literature on notions of causality goes back to early work by Ashenfelter [6] and subsequent work by Ashenfelter and Card [7], Heckman and Robb [8], Lalonde [9], Fraker and Maynard [10], Card and Sullivan [11] and Manski [12,13]. The use of IV technique in outcomes research has increased in recent years because even in the presence of hidden bias, such methods may consistently estimate the average causal effects [14]. We are in the beginning stage of this application on outcomes research studies and believe the surge is yet to start [15–19].

However, like many techniques borrowed from one discipline and applied to another, there is a tendency to apply this method blindly. Researchers, unaware of the shortcomings of this technique, may apply it inappropriately. In this article, we draw attention to the problem of using instruments that explain little of the variation in the endogenous explanatory variables (such as treatment choice variables, etc.). These instruments can lead to large inconsistencies in IV estimates. The magnitude of the bias of IV estimates approaches that of ordinary least squares (OLS) estimates as the R-square between the instruments and the endogenous variable approaches zero. While these results are known in economics literature, their potential implications for empirical work related to outcomes research have not been fully appreciated.

The discussion in this article does not provide detailed or rigorous treatment of the theory that premises the IV approach. In recent years, several books on IV methods, with various levels of sophistication, have been published. Wooldridge’s book is an excellent source for researchers with an elementary level of

Address correspondence to: Onur Baser, President, STATinMED Research, 218 N. Fourth Avenue, Suite 205, Ann Arbor, MI 48104, USA. E-mail: obaser@statinmed.com

10.1111/j.1524-733X.2009.05567.x

© 2009, International Society for Pharmacoeconomics and Outcomes Research (ISPOR) 1098-3015/09/1201 1201–1209

VALUE IN HEALTH
statistical knowledge [20]. His more advanced book provides detailed information on IV for readers with an advanced mathematical background [21]. The work of Bowden and Turkington is geared toward mathematical outcomes researchers [22]. Curious readers are encouraged to consult these books for a more detailed analysis.

**Overview of IV Estimation**

Suppose we want to estimate the effect of treatment (T) on outcome (Y), i.e., estimate β₁ in:

\[ Y = \beta_0 + T\beta_1 + \epsilon \]

For simplicity, we assume dichotomous treatment variable (T), homogenous treatment effect, linear regression, and no covariates. ε is unobservable. Least squares estimate of the equation yields the following estimator:

\[ \hat{\beta}_{OLS} = \frac{\sum T Y - \sum T \sum Y}{\sum T^2} \]

which is the difference in mean outcomes. In order to reach a consistent estimator, the key assumption is that treatment (T) is not correlated with the unobserved determinants of the outcome (ε).

\[ E(T'\varepsilon) = 0 \Rightarrow T'(Y - T\hat{\beta}_{OLS}) = 0 \Rightarrow \hat{\beta}_{OLS} = (T'T)^{-1}T'Y \]

The OLS assumption is unlikely to hold, because treatment is related to omitted factors influencing outcome. For example, patients who are more severely ill in ways known to their physicians but not to the analyst might not get the treatment, or vice versa.

In order to obtain a consistent estimator of β₀ and β₁ when treatment and omitted factors are related, we need additional information.

The information comes by way of a new variable—an IV—(Z) that satisfies the following properties:

1. \( \text{Cor}(Z, \epsilon) = 0 \) Z should have no partial effect on the outcomes variable and should not be correlated with other factors that affect the outcomes variable.
2. \( \text{Cor}(Z, T) \neq 0 \) Z must be related, either positively or negatively, to the treatment indicator.

If these two conditions are satisfied, then the IV estimator is:

\[ E(Z'\varepsilon) = 0 \Rightarrow Z'(Y - Z\hat{\beta}_{IV}) = 0 \Rightarrow \hat{\beta}_{IV} = (Z'Z)^{-1}Z'Y \]

Note that all IV results apply asymptotically. Small sample estimation properties of IV are more complex and, as discussed in the next section, not generally understood. Variants on this approach include two-stage least squares [21,23,24], limited information maximum likelihood estimator [21,25], general method of moments [21,26], and sample selection corrections ("Heckit") [21,27].

The coin toss in the context of randomized controlled trials (RCTs) is a perfect example of IV. The coin toss does not affect the outcome of interest directly (assumption (i)) but it determines treatment assignment (assumption (ii)). Following are some examples of IVs that have been used in applied research:

1. **Geography** (distance, rivers, small area variation) [28–32].
2. **Legal/political institutions** (laws, election dynamics) [31,33,34].
3. **Administrative rules** (wage/staffing rules, reimbursement rules, eligibility rules) [35–37].
4. **Naturally occurring randomization** (blood type of recipients, draft, birth timing, lottery, roommate assignment) [36,38–40].

**Why Not Always Use IV?**

The immediate question that arises is if the IV method is superior to risk adjustment methods such as propensity score matching or multivariate regression in the sense that these methods both cover observable and unobservable factors, why not always use the IV method?

First, it is hard to find variables that meet the definition of valid instruments. Conceptually, most variables that have an effect on treatment variables may also have a direct effect on the outcomes variable.

Second, the standard errors of IV estimates are likely to be larger than those of OLS estimates, creating publication bias. It is important to understand that publication bias may exist even without the authors of individual studies being aware of it. The potential problem simply arises because of the desire to report significant findings. Although lack of treatment is useful to report, evidence against the null hypothesis—that is, favorable to the finding of a treatment effect—is more likely to be reported. Because IV estimates have to be larger in order to be significant, published results tend to create publication bias [41].

Third, the desirable properties of the IV estimator hold in large sample sizes. With simulation, Grootendorst showed that in smallsample sizes, the estimates can be highly inaccurate [17]. Also, refer to Kinal for related issues with small properties, where the IV estimator may have no expected value [42].

Fourth, the interpretation of IV is difficult, especially when the treatment effect is heterogeneous. Randomized clinical trials estimate the treatment effect in well-defined populations. Therefore, there are always issues of external validity. Analogous issues arise with IV estimates. They estimate treatment effect for the “marginal” patients whose treatment is affected by the instruments. Therefore, they often do not estimate treatment effect in the general population [43].

The last sets of problems are related to weak instruments, which are the focus of this article.

**Weak Instruments**

There is a very important difference between the two requirements for an IV. Because assumption (i) is a covariance between the IVs and the unobservable error u, it can never be directly checked or even tested. Rather, we must maintain this assumption by appealing to clinical behavior (in the presence of multiple instruments, indirect tests can be conducted. See Wooldridge for details [21]).

By contrast, assumption (ii) that IV is correlated with treatment choice can be tested, given a random sample from the population. If this correlation is weak, this may lead to large inconsistencies in IV estimates with the bias in the same direction as that of OLS estimates [44]. Because IV estimates also have larger standard errors than those of OLS estimates, as pointed out by Bound et al., “...the cure can be worse than the disease” [45].

**Testing for Weak Instruments**

Staiger and Stock formalized the definition of weak instruments and most researchers appear to have concluded (incorrectly) from that work that if F-statistics on coefficients of exogenous
variables on the endogenous treatment indicator is greater than 10, one need to worry no further about weak instruments [43].

Another statistic commonly used, as recommended by Bound et al., is the $R^2$ of the regression with instruments partialed out [45]. However, Shea showed, in general, the distribution of this $F$-statistic is nonstandard [46]. Also, for models with multiple endogenous variables, these indicators may not be sufficiently informative [46].

To grasp the pitfalls facing empirical researchers here, consider the following simple example. We have a model with two endogenous covariates (treatment and insurance choice) and two instruments (distance to nearest specialized hospital and small area variation). Distance to nearest specialized hospital increases the likelihood of being admitted in a specialized hospital. Therefore, patients near a specialized hospital are more likely to be treated by specialized medical staff, in a special care unit, and with other dimensions of higher intensity. Distance to nearest specialized hospital might also affect the insurance choice. The differences between each managed care plan lie mainly in the degree of compensation one receives for medical treatment outside the managed care network. Patients who live close to a specialized hospital are more likely to choose a low premium insurance plan.

Small area variations in hospital surgical volumes might affect the treatment quality. Because volume is positively correlated with surgical quality, patients who live in high volume areas might get better treatment choices than patients who live in low volume areas. Therefore, both distance to nearest specialized hospital and small area variation in hospital volumes are valid instruments because they have a direct effect on treatment or insurance choice but indirectly related to the outcome.

Suppose the distance to the nearest specialized hospital is highly correlated with treatment and insurance choice, but the small area variation is just a noise. In this case, because we have one instrument for two endogenous variables, this model is underidentified. Bound et al.'s $F$-statistics [45] and partial $R^2$-squared measures from regression with instruments will not reveal this weakness. Indeed, the $F$-statistics are statistically significant and without investigation, but we may not realize the model cannot be estimated in this form. The statistics proposed by Bound et al. [45] diagnose instruments relevant only in the presence of one endogenous covariate.

When multiple endogenous variables are used, other statistics can be used. Shea provided such statistics [46]. Shea's $R^2$-squares take the intercorrelations among instruments into account. As a rule of thumb, if an estimated equation yields a large value of the standard partial $R^2$-squares and a small value of the Shea measure, we should conclude that the instruments lack sufficient relevance to explain all the endogenous regressors.

A more general approach to weak instruments was proposed by Anderson [47] and discussed in Hall and Peixe [48]. Anderson's approach considers the canonical correlations of the excluded and included instruments. This test shows whether some instruments are redundant. Stock and Yogo go into more details and provide useful rules of thumb regarding the weakness of instruments based on a statistic from Cragg and Donald [49,50].

Data Sources and Construction of Variables

We illustrate the implications of a weak IV using MarketScan data to examine the effect of controller medication on health-care expenditures for asthma patients. Briefly, the MarketScan Commercial Claims and Encounters Database contains detailed descriptions of inpatient, outpatient, medical, and outpatient prescription drug services for approximately 13 million persons in 2005 who were covered by corporate-sponsored health-care plans. Details of the patient selection criteria are provided in Crown et al. [51] and summarized as follows:

1. Patients with evidence of asthma were selected from the intersection of the medical claims and encounter records, enrollment files, and pharmaceutical data files.
2. Individuals meeting at least one of the following criteria were deemed to show evidence of asthma:
   - At least two outpatient claims with primary or secondary diagnoses of asthma.
   - At least one emergency room (ER) claim with primary diagnosis of asthma, and a drug transaction for an asthma medication 90 days before or 7 days after the ER claim.
   - At least one inpatient claim with a primary diagnosis of asthma.
   - A secondary diagnosis of asthma and a primary diagnosis of respiratory infection in an outpatient or inpatient claim.
   - At least one drug transaction for an anti-inflammatory agent, oral antileukotrienes, long-acting bronchodilator, or inhaled or oral short-acting beta-agonists.
3. Patients with a diagnosis of chronic obstructive pulmonary disease and having one or more diagnoses or procedure codes indicating pregnancy or delivery, or who were not continuously enrolled for 24 months, were excluded from our study group.

The sociodemographic characteristics include age of the household, percentage of the patients who were female, and geographic region (northeast, north-central, south, west, and “other” region). Charlson comorbidity index scores are generated to capture the level and burden of comorbidity. Point-of-service plans and other plan types, including health maintenance organizations and preferred provider organizations, are included. The analytic file contains patients with fee-for-service (FFS) health plans and those with partially or fully capitated plans. Data on costs are not available for the capitated plans however. Therefore, the value of patients’ service utilization under the capitated plans is priced and imputed using average payments from the MarketScan FFS inpatient and outpatient services by region, year, and procedure.

The outcomes variable is total health-care costs. The MarketScan database contains information on all payments processed with regard to reimbursement for particular services, including secondary payers and patient out-of-pocket costs. For services in which these MarketScan employers are a secondary payer (i.e., the patient has other primary insurance), the amounts paid by other insurers is also documented and included in the cost. In cases where services delivered are completely covered by another primary insurance, these claims are not included in the database. Data on costs were not available for the capitated plans. Therefore, the value of patients’ service utilization under the capitated plan was priced and imputed using average payments from the MarketScan FFS inpatient and outpatient services by region, year, and procedure.

The endogenous variable is treatment choice ($c = 1$ if controller, $c = 0$ if reliever). Asthma drugs can be reliever medications (used to relieve symptoms in an acute asthma exacerbation or asthma attack) or primarily controller medications (used to control pulmonary inflammation and prevent an attack).

The IVs are controller/reliever copay ratio and physician/practice prescribing pattern. Copayments for outpatient pharma-
drugs by therapeutic class for each plan to calculate the ratio of mean controller payments to mean reliever copayments. These plan-level ratios are attached to each patient’s record for a given plan. Our second IV involves calculation of the proportion of patients obtaining controller medication for each physician/provider tax identification number. In many cases, this tax identification number includes a multiphysician medical practice, but in some cases is unique to one physician.

**Results**

The objective of this study was to estimate the cost of illness for asthma patients treated with controller and reliever medications using the IV approach.

Table 1 reports the demographic characteristics of the sample, stratified by treatment choice. Patients using controller medication had a mean age of 40 years (compared with 30 years for patients using reliever medication) and were more likely to be female. The racial distribution in counties was similar between the two groups. Patients treated with controller medication were more likely to receive their health-care coverage as an employee compared with those with reliever medication. Significant differences in mean income between the two groups were evident from county-level US census data compared with the claims data. Patients treated with controller medication had higher numbers of major diagnostic categories, higher Charlson comorbidity scores, and higher rates of asthma specific comorbidities. The descriptive table shows that the ratio of mean controller copayments to mean reliever copayments was lower for the treated controller group relative to the reliever group. Physicians were more likely to prescribe controller medication to the patient group treated with controller medication. The unadjusted total health-care costs were significantly higher for patients treated with controller medication relative to the ones who were treated with reliever medication.

Because the Hausman test showed the treatment choice is endogenous, the IV method has been applied (P < 0.000). We have two possible candidates for instrument: controller/reliever copay ratio and physician/practice prescribing pattern. The first key assumption for IV is that it does not independently affect the outcome, so it is not associated with measured and unmeasured health status. Table 2 shows a different division of the sample from Table 1, namely division according to quintile of IVs. The first assumption, that copay ratios and prescribing patterns affect health-care costs only through its effect on the

### Table 1 Summary of asthma patient characteristics

<table>
<thead>
<tr>
<th>Variables</th>
<th>Controller medication (n = 3,903)</th>
<th>Reliever medication (n = 11,049)</th>
<th>P-values</th>
<th>STD difference</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Explanatory Variables</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>39.74 ± 16.52</td>
<td>30.30 ± 18.07</td>
<td>0.0000</td>
<td>53.38</td>
</tr>
<tr>
<td>(%) Female</td>
<td>0.63 ± 0.48</td>
<td>0.57 ± 0.49</td>
<td>0.0000</td>
<td>11.87</td>
</tr>
<tr>
<td>Race</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>White</td>
<td>0.84 ± 0.14</td>
<td>0.85 ± 0.13</td>
<td>0.2737</td>
<td>5.49</td>
</tr>
<tr>
<td>Black</td>
<td>0.09 ± 0.11</td>
<td>0.09 ± 0.10</td>
<td>0.3819</td>
<td>4.41</td>
</tr>
<tr>
<td>Other</td>
<td>0.02 ± 0.02</td>
<td>0.02 ± 0.02</td>
<td>0.7958</td>
<td>3.47</td>
</tr>
<tr>
<td>Geographic Regions</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Northeast</td>
<td>0.02 ± 0.15</td>
<td>0.02 ± 0.13</td>
<td>0.1442</td>
<td>2.66</td>
</tr>
<tr>
<td>North-Central</td>
<td>0.81 ± 0.39</td>
<td>0.85 ± 0.36</td>
<td>0.0000</td>
<td>10.86</td>
</tr>
<tr>
<td>South</td>
<td>0.02 ± 0.15</td>
<td>0.02 ± 0.13</td>
<td>0.1442</td>
<td>2.66</td>
</tr>
<tr>
<td>West</td>
<td>0.04 ± 0.20</td>
<td>0.04 ± 0.19</td>
<td>0.4918</td>
<td>1.27</td>
</tr>
<tr>
<td>Year of Patient Identification</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1996</td>
<td>0.40 ± 0.49</td>
<td>0.39 ± 0.49</td>
<td>0.3035</td>
<td>1.91</td>
</tr>
<tr>
<td>1997</td>
<td>0.26 ± 0.44</td>
<td>0.28 ± 0.45</td>
<td>0.0042</td>
<td>4.59</td>
</tr>
<tr>
<td>1998</td>
<td>0.31 ± 0.46</td>
<td>0.31 ± 0.46</td>
<td>0.5362</td>
<td>1.15</td>
</tr>
<tr>
<td>1999</td>
<td>0.02 ± 0.15</td>
<td>0.02 ± 0.13</td>
<td>0.0209</td>
<td>4.15</td>
</tr>
<tr>
<td>Member Type</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Employee</td>
<td>0.49 ± 0.50</td>
<td>0.35 ± 0.48</td>
<td>0.0000</td>
<td>29.54</td>
</tr>
<tr>
<td>Spouse</td>
<td>0.27 ± 0.45</td>
<td>0.20 ± 0.40</td>
<td>0.0000</td>
<td>18.72</td>
</tr>
<tr>
<td>Dependents</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4-11 Years</td>
<td>0.08 ± 0.27</td>
<td>0.19 ± 0.39</td>
<td>0.0000</td>
<td>32.31</td>
</tr>
<tr>
<td>12-18 Years</td>
<td>0.11 ± 0.32</td>
<td>0.21 ± 0.41</td>
<td>0.0000</td>
<td>25.59</td>
</tr>
<tr>
<td>Others</td>
<td>0.04 ± 0.20</td>
<td>0.06 ± 0.24</td>
<td>0.0000</td>
<td>9.99</td>
</tr>
<tr>
<td>County Mean Household Income</td>
<td>$25,829 ± $6,686</td>
<td>$24,997 ± $6,141</td>
<td>0.0000</td>
<td>12.95</td>
</tr>
<tr>
<td>Number of Major Diagnosis Categories</td>
<td>6.80 ± 2.39</td>
<td>6.06 ± 2.15</td>
<td>0.0000</td>
<td>32.58</td>
</tr>
<tr>
<td>Charlson Comorbidity Index</td>
<td>0.92 ± 1.18</td>
<td>0.49 ± 0.92</td>
<td>0.0000</td>
<td>40.73</td>
</tr>
<tr>
<td>Asthma-Specific Comorbidities</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Allergic Rhinitis</td>
<td>0.28 ± 0.45</td>
<td>0.18 ± 0.39</td>
<td>0.0000</td>
<td>23.94</td>
</tr>
<tr>
<td>Migraine</td>
<td>0.07 ± 0.26</td>
<td>0.05 ± 0.22</td>
<td>0.0000</td>
<td>8.43</td>
</tr>
<tr>
<td>Depression</td>
<td>0.10 ± 0.30</td>
<td>0.10 ± 0.30</td>
<td>0.3604</td>
<td>1.69</td>
</tr>
<tr>
<td>Gastrointestinal Disorder</td>
<td>0.28 ± 0.45</td>
<td>0.21 ± 0.41</td>
<td>0.0000</td>
<td>16.44</td>
</tr>
<tr>
<td>Sinusitis</td>
<td>0.27 ± 0.44</td>
<td>0.23 ± 0.42</td>
<td>0.0000</td>
<td>7.81</td>
</tr>
<tr>
<td>Anxiety</td>
<td>0.03 ± 0.16</td>
<td>0.02 ± 0.15</td>
<td>0.4556</td>
<td>1.38</td>
</tr>
<tr>
<td>Instrumental Variables</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Controller/Reliever Copayment</td>
<td>1.32 ± 0.30</td>
<td>1.29 ± 0.26</td>
<td>0.0000</td>
<td>14.08</td>
</tr>
<tr>
<td>Tax Provider ID Controller %</td>
<td>0.61 ± 0.05</td>
<td>0.60 ± 0.04</td>
<td>0.3144</td>
<td>20.73</td>
</tr>
<tr>
<td>Tax Provider ID Reliever %</td>
<td>0.39 ± 0.05</td>
<td>0.40 ± 0.04</td>
<td>0.3144</td>
<td>20.73</td>
</tr>
<tr>
<td>Outcomes Variable</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total Cost</td>
<td>$4,321 ± $7,011</td>
<td>$2,792 ± $6,151</td>
<td>0.0000</td>
<td>23.18</td>
</tr>
</tbody>
</table>

ID, identification number; STD, standard deviation.
likelihood of treatment choice, cannot be tested directly. We can, however, indirectly test this assumption and see how realistic it is a priori. This assumption would be satisfied if a person’s copay ratio or physician practice/prescribing pattern was not associated with the clinical severity of the asthma, the primary unobserved variable that will determine the treatment. If this is true, IVs should also be independent of observed variables such as age, gender, and comorbidities, associated with health status and hence the likelihood of treatment choice. The data in Table 2 show that observable factors are independent across the quintiles.

We also tested whether these instruments satisfied the second assumption: are they highly correlated with treatment? By looking at the quintiles in Table 2, we can see the correlation between treatment choice and the IVs across the quintiles. Among the physicians who are most likely to give prescriptions for controllers, the number of patients who are getting controller medication is almost two times higher. However, the likelihood of being prescribed controller medications was similar across the quintiles of copay ratios.

First, we used only controller/reliever copay ratio as an IV (see Table 3). Shea’s partial $R^2$ was very small for this equation. The Cragg–Donald statistic failed to reject its null hypothesis of underidentification. The Anderson canonical correlation failed to reject its null hypothesis at the 10% level, suggesting that the instrument may be inadequate to identify the equation. Second, we used only physician/practice prescribing patterns as an IV. Shea’s partial $R^2$ was 0.58 for this equation. The Cragg–Donald statistic rejected its null hypothesis of underidentification. The Anderson canonical correlation rejected its null hypothesis, suggesting that the instrument was adequate to identify the equation. We also attempted both instruments at the same time, but the Hall–Peixe test showed that ratio of copays as an instrument was redundant. The Hausman–Taylor

Table 2 Descriptive statistics on asthma patients by quintiles of instrumental variables

<table>
<thead>
<tr>
<th>Variables</th>
<th>Q1</th>
<th>Q2</th>
<th>Q3</th>
<th>Q4</th>
<th>Q5</th>
<th>Q1</th>
<th>Q2</th>
<th>Q3</th>
<th>Q4</th>
<th>Q5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Explanatory Variables</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age (years)</td>
<td>31.62</td>
<td>31.81</td>
<td>30.52</td>
<td>30.93</td>
<td>40.22</td>
<td>31.20</td>
<td>31.49</td>
<td>26.59</td>
<td>31.55</td>
<td>41.61</td>
</tr>
<tr>
<td>(%) Female</td>
<td>0.59</td>
<td>0.61</td>
<td>0.56</td>
<td>0.56</td>
<td>0.62</td>
<td>0.57</td>
<td>0.59</td>
<td>0.52</td>
<td>0.59</td>
<td>0.62</td>
</tr>
<tr>
<td>Race</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>White</td>
<td>0.88</td>
<td>0.87</td>
<td>0.86</td>
<td>0.79</td>
<td>0.78</td>
<td>0.87</td>
<td>0.86</td>
<td>0.75</td>
<td>0.85</td>
<td>0.79</td>
</tr>
<tr>
<td>Black</td>
<td>0.08</td>
<td>0.07</td>
<td>0.08</td>
<td>0.08</td>
<td>0.13</td>
<td>0.08</td>
<td>0.08</td>
<td>0.11</td>
<td>0.07</td>
<td>0.12</td>
</tr>
<tr>
<td>Other</td>
<td>0.02</td>
<td>0.02</td>
<td>0.02</td>
<td>0.03</td>
<td>0.03</td>
<td>0.02</td>
<td>0.02</td>
<td>0.04</td>
<td>0.02</td>
<td>0.03</td>
</tr>
<tr>
<td>Geographic Regions</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Northeast</td>
<td>0.00</td>
<td>0.00</td>
<td>0.02</td>
<td>0.02</td>
<td>0.06</td>
<td>0.00</td>
<td>0.00</td>
<td>0.21</td>
<td>0.00</td>
<td>0.04</td>
</tr>
<tr>
<td>North-Central</td>
<td>1.00</td>
<td>0.95</td>
<td>0.92</td>
<td>0.61</td>
<td>0.54</td>
<td>0.99</td>
<td>0.91</td>
<td>0.28</td>
<td>0.86</td>
<td>0.59</td>
</tr>
<tr>
<td>South</td>
<td>0.00</td>
<td>0.00</td>
<td>0.02</td>
<td>0.02</td>
<td>0.06</td>
<td>0.00</td>
<td>0.00</td>
<td>0.21</td>
<td>0.00</td>
<td>0.04</td>
</tr>
<tr>
<td>West</td>
<td>0.00</td>
<td>0.02</td>
<td>0.02</td>
<td>0.17</td>
<td>0.07</td>
<td>0.00</td>
<td>0.00</td>
<td>0.17</td>
<td>0.07</td>
<td>0.09</td>
</tr>
<tr>
<td>Year of Patient Identification</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1996</td>
<td>0.08</td>
<td>0.00</td>
<td>0.87</td>
<td>0.39</td>
<td>0.37</td>
<td>0.93</td>
<td>0.09</td>
<td>0.02</td>
<td>0.00</td>
<td>0.29</td>
</tr>
<tr>
<td>1997</td>
<td>0.85</td>
<td>0.00</td>
<td>0.04</td>
<td>0.04</td>
<td>0.38</td>
<td>0.07</td>
<td>0.86</td>
<td>0.87</td>
<td>0.00</td>
<td>0.17</td>
</tr>
<tr>
<td>1998</td>
<td>0.07</td>
<td>0.93</td>
<td>0.09</td>
<td>0.56</td>
<td>0.24</td>
<td>0.00</td>
<td>0.06</td>
<td>0.10</td>
<td>1.00</td>
<td>0.44</td>
</tr>
<tr>
<td>1999</td>
<td>0.00</td>
<td>0.07</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.00</td>
<td>0.00</td>
<td>0.01</td>
<td>0.00</td>
<td>0.10</td>
</tr>
<tr>
<td>Member Type</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Employee</td>
<td>0.40</td>
<td>0.40</td>
<td>0.37</td>
<td>0.36</td>
<td>0.40</td>
<td>0.35</td>
<td>0.40</td>
<td>0.33</td>
<td>0.40</td>
<td>0.43</td>
</tr>
<tr>
<td>Spouse</td>
<td>0.19</td>
<td>0.20</td>
<td>0.18</td>
<td>0.24</td>
<td>0.31</td>
<td>0.20</td>
<td>0.19</td>
<td>0.18</td>
<td>0.20</td>
<td>0.31</td>
</tr>
<tr>
<td>Dependents</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4-11 Years</td>
<td>0.16</td>
<td>0.14</td>
<td>0.17</td>
<td>0.21</td>
<td>0.14</td>
<td>0.16</td>
<td>0.17</td>
<td>0.27</td>
<td>0.16</td>
<td>0.12</td>
</tr>
<tr>
<td>12-18 Years</td>
<td>0.19</td>
<td>0.19</td>
<td>0.22</td>
<td>0.16</td>
<td>0.12</td>
<td>0.22</td>
<td>0.18</td>
<td>0.19</td>
<td>0.17</td>
<td>0.12</td>
</tr>
<tr>
<td>Others</td>
<td>0.07</td>
<td>0.07</td>
<td>0.06</td>
<td>0.03</td>
<td>0.03</td>
<td>0.07</td>
<td>0.06</td>
<td>0.02</td>
<td>0.06</td>
<td>0.02</td>
</tr>
<tr>
<td>County Mean Household Income</td>
<td>$24,691</td>
<td>$26,040</td>
<td>$24,584</td>
<td>$27,610</td>
<td>$24,897</td>
<td>$24,127</td>
<td>$24,640</td>
<td>$26,216</td>
<td>$26,631</td>
<td>$26,069</td>
</tr>
<tr>
<td>Number of Major Diagnosis Categories</td>
<td>6.27</td>
<td>6.31</td>
<td>6.30</td>
<td>5.72</td>
<td>6.34</td>
<td>6.30</td>
<td>6.27</td>
<td>5.64</td>
<td>6.18</td>
<td>6.38</td>
</tr>
<tr>
<td>Charlson Comorbidity Index</td>
<td>0.54</td>
<td>0.57</td>
<td>0.57</td>
<td>0.55</td>
<td>0.80</td>
<td>0.57</td>
<td>0.56</td>
<td>0.56</td>
<td>0.56</td>
<td>0.80</td>
</tr>
<tr>
<td>Asthma-Specific Comorbidities</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Allergic Rhinitis</td>
<td>0.20</td>
<td>0.21</td>
<td>0.21</td>
<td>0.26</td>
<td>0.19</td>
<td>0.21</td>
<td>0.21</td>
<td>0.21</td>
<td>0.22</td>
<td>0.19</td>
</tr>
<tr>
<td>Migraine</td>
<td>0.06</td>
<td>0.06</td>
<td>0.05</td>
<td>0.06</td>
<td>0.05</td>
<td>0.06</td>
<td>0.05</td>
<td>0.06</td>
<td>0.05</td>
<td>0.05</td>
</tr>
<tr>
<td>Depression</td>
<td>0.11</td>
<td>0.11</td>
<td>0.10</td>
<td>0.10</td>
<td>0.07</td>
<td>0.11</td>
<td>0.11</td>
<td>0.08</td>
<td>0.11</td>
<td>0.07</td>
</tr>
<tr>
<td>Gastrointestinal Disorder</td>
<td>0.22</td>
<td>0.23</td>
<td>0.22</td>
<td>0.21</td>
<td>0.28</td>
<td>0.22</td>
<td>0.22</td>
<td>0.23</td>
<td>0.23</td>
<td>0.28</td>
</tr>
<tr>
<td>Sinusitis</td>
<td>0.24</td>
<td>0.24</td>
<td>0.25</td>
<td>0.23</td>
<td>0.23</td>
<td>0.25</td>
<td>0.24</td>
<td>0.24</td>
<td>0.24</td>
<td>0.22</td>
</tr>
<tr>
<td>Anxiety</td>
<td>0.02</td>
<td>0.03</td>
<td>0.03</td>
<td>0.02</td>
<td>0.02</td>
<td>0.03</td>
<td>0.02</td>
<td>0.03</td>
<td>0.02</td>
<td>0.02</td>
</tr>
<tr>
<td>Treatment</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Controller Medication</td>
<td>0.22</td>
<td>0.24</td>
<td>0.25</td>
<td>0.26</td>
<td>0.34</td>
<td>0.25</td>
<td>0.23</td>
<td>0.26</td>
<td>0.25</td>
<td>0.34</td>
</tr>
<tr>
<td>Outcomes Variable</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total Cost</td>
<td>$3,043</td>
<td>$3,223</td>
<td>$2,805</td>
<td>$2,806</td>
<td>$4,134</td>
<td>$2,940</td>
<td>$3,064</td>
<td>$3,150</td>
<td>$3,150</td>
<td>$3,110</td>
</tr>
<tr>
<td>IV-1 controller/reliever ratio</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IV-2 controller/reliever ratio</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3 Testing the strength of the instruments

<table>
<thead>
<tr>
<th>Instruments</th>
<th>Shea’s $R^2$</th>
<th>Partial $R^2$</th>
<th>F-stat</th>
<th>P-value</th>
<th>Cragg–Donald (P-value)</th>
<th>Anderson’s test (P-value)</th>
</tr>
</thead>
<tbody>
<tr>
<td>IV-1</td>
<td>0.0023</td>
<td>0.0023</td>
<td>1.78</td>
<td>0.1563</td>
<td>2.79 (0.145)</td>
<td>2.57 (0.165)</td>
</tr>
<tr>
<td>IV-2</td>
<td>0.58</td>
<td>0.58</td>
<td>18.65</td>
<td>0</td>
<td>28.54 (0.000)</td>
<td>28.57 (0.000)</td>
</tr>
</tbody>
</table>

IV, instrumental variable.
test showed the rejection of controller–reliever copay ratio as an adequate instrument [53].

Our estimation method has two stages. We first estimated the likelihood of prescribing controller medication as a function of the exogenous covariates and IVs via conventional probit analysis. The results are presented in Table 4. For the second stage outcome (health-care expenditures) regression, following the Park test as described by Manning and Mullahy, we chose generalized linear models (GLMs) with log links and gamma family. We regressed the total expenditure on treatment choice, following the Park test as described by Manning and Mullahy, we chose generalized linear models (GLMs) with log links and gamma family [54]. We regressed the total expenditure on treatment choice, following the Park test as described by Manning and Mullahy, we chose generalized linear models (GLMs) with log links and gamma family [54]. We regressed the total expenditure on treatment choice, following the Park test as described by Manning and Mullahy, we chose generalized linear models (GLMs) with log links and gamma family [54]. We regressed the total expenditure on treatment choice, following the Park test as described by Manning and Mullahy, we chose generalized linear models (GLMs) with log links and gamma family [54]. We regressed the total expenditure on treatment choice, following the Park test as described by Manning and Mullahy, we chose generalized linear models (GLMs) with log links and gamma family [54]. We regressed the total expenditure on treatment choice, following the Park test as described by Manning and Mullahy, we chose generalized linear models (GLMs) with log links and gamma family [54]. We regressed the total expenditure on treatment choice, following the Park test as described by Manning and Mullahy, we chose generalized linear models (GLMs) with log links and gamma family [54].

In the outcomes tables (Table 6), we compared raw outcomes between the patient groups that used controller medication only and reliever medication only, then we used the standard regression technique to adjust the raw outcome differences for observable differences in demographic and comorbid diseases characteristics between these two groups. These estimates were then compared with three IV estimators: one with weak IV, one with strong IV, and one with using both of them as an IV (one being a redundant IV).

The predicted cost differences were similar between standard regression and IV regression with weak instruments ($260 vs. $270). However, standard errors of the IV estimator increased almost 10-fold. Therefore, the differences were insignificant according to IV estimation with a weak instrument. However, using the right IV, with a strong relationship between treatment variables, there was a significant relationship between the health-care cost of the controller-only user group and reliever-only user group ($894, \( P = 0.000 \)). Because the coefficient on treatment choice in standard GLM regression is positive, downward bias can explain the positive relationship between the unobserved severity level and reliever medication use.

### Discussion

A widely recognized problem in observational research is that because of unobservable differences between individuals, it is unclear to what extent differences in outcomes reflect differences in treatment choices, even if we follow standard risk adjustment models, such as regression analysis or propensity score matching. IV approach is a novel method to control for both observed and unobserved differences between individuals. However, this method is based on two strong assumptions and ignoring those assumptions can result in severe bias and inefficiency of the estimators.

A valid IV, which helps determine whether an individual is treated but does not determine other factors that affect outcome of interest, can overcome using the method of OLS. Current literature clarifies how to interpret estimated treatment effects using IVs. Because it is not possible to estimate the treatment effect for each individual, researchers rely on average treatment effect (ATE), which is the average of the individual treatment effects across the whole population of interest. When the treatment being evaluated has the same effect for everyone, any assumptions can result in severe bias and inefficiency of the estimators.

A widely recognized problem in observational research is that because of unobservable differences between individuals, it is unclear to what extent differences in outcomes reflect differences in treatment choices, even if we follow standard risk adjustment models, such as regression analysis or propensity score matching. IV approach is a novel method to control for both observed and unobserved differences between individuals. However, this method is based on two strong assumptions and ignoring those assumptions can result in severe bias and inefficiency of the estimators.

A valid IV, which helps determine whether an individual is treated but does not determine other factors that affect outcome of interest, can overcome using the method of OLS. Current literature clarifies how to interpret estimated treatment effects using IVs. Because it is not possible to estimate the treatment effect for each individual, researchers rely on average treatment effect (ATE), which is the average of the individual treatment effects across the whole population of interest. When the treatment being evaluated has the same effect for everyone, any assumptions can result in severe bias and inefficiency of the estimators.

A valid IV, which helps determine whether an individual is treated but does not determine other factors that affect outcome of interest, can overcome using the method of OLS. Current literature clarifies how to interpret estimated treatment effects using IVs. Because it is not possible to estimate the treatment effect for each individual, researchers rely on average treatment effect (ATE), which is the average of the individual treatment effects across the whole population of interest. When the treatment being evaluated has the same effect for everyone, any assumptions can result in severe bias and inefficiency of the estimators.

A valid IV, which helps determine whether an individual is treated but does not determine other factors that affect outcome of interest, can overcome using the method of OLS. Current literature clarifies how to interpret estimated treatment effects using IVs. Because it is not possible to estimate the treatment effect for each individual, researchers rely on average treatment effect (ATE), which is the average of the individual treatment effects across the whole population of interest. When the treatment being evaluated has the same effect for everyone, any assumptions can result in severe bias and inefficiency of the estimators.
To be sure the IV method estimates is the ATE among those who alter their treatment status because they react to the instrument. This is called local average treatment effect (LATE). When patients do not make decisions to reach the instrument based on factors that also determine treatment gains, the LATE equals the ATE among those exposed to the treatment.

In our application, the consequence of instruments with little explanatory power is increasing bias in the estimated IV coefficients and worsening the large sample approximations to the finite sample distributions. With weak instruments, the large sample bias of the IV estimator is the same as that of the OLS estimator, and IV becomes inconsistent and nothing is gained from instrumenting. One recommendation when faced with a weak instrument is to be parsimonious in the choice of instruments because if we use redundant instruments, even for the cases where the identification is not a problem, final estimates are inefficient.

In the past, ingenious instruments have been proposed and methods produced closer to “true” estimates than standard risk adjustment models. In a recent article, Stukel compared four analytic methods to remove the effects of selection bias in observational studies: multivariable model risk adjustment, propensity score risk adjustment, propensity-based matching, and IV analysis [56]. She concluded that the IV method produced closer results to the results from RCTs, which balances both measurable and immeasurable factors.

If we go back to our original question and answer: Do we have a method to control for both observed and unobserved bias? The answer is “theoretically YES” but practical application is very limited because of the difficulty in finding the right instrument. Researchers should test whether their instruments satisfy the two key assumptions. Application of the instrument without prior tests may produce inconsistent and inefficient results, which is worse than applying simple OLS or propensity score matching. Therefore, the challenge for outcomes researchers remains to find and adopt the right instruments for outcomes research studies. Otherwise, the cure can be worse than the disease.

The empirical work for this article was completed when the author was an employee of Thomson-Reuters and presented at the ISPOR conference.

Source of financial support: None.

---

Table 5  Second stage estimation using generalized linear model with log link and gamma family

<table>
<thead>
<tr>
<th>Variables</th>
<th>No IV</th>
<th>With IV-1</th>
<th>With IV-2</th>
<th>With IV-1 and IV-2</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Explanatory Variables</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>0.02</td>
<td>0.02</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>(%) Female</td>
<td>0.03</td>
<td>0.06</td>
<td>0.03</td>
<td>0.02</td>
</tr>
<tr>
<td>Race</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>White</td>
<td>-0.41</td>
<td>-0.40</td>
<td>-0.19</td>
<td>-0.18</td>
</tr>
<tr>
<td>Black</td>
<td>-0.45</td>
<td>-0.46</td>
<td>-0.72</td>
<td>-0.72</td>
</tr>
<tr>
<td>Other</td>
<td>-0.72</td>
<td>-0.75</td>
<td>-0.15</td>
<td>-0.16</td>
</tr>
<tr>
<td>Geographic Regions</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Northeast</td>
<td>0.10</td>
<td>0.10</td>
<td>0.19</td>
<td>0.19</td>
</tr>
<tr>
<td>North-Central</td>
<td>-0.12</td>
<td>-0.14</td>
<td>-0.25</td>
<td>-0.25</td>
</tr>
<tr>
<td>West</td>
<td>-0.19</td>
<td>-0.20</td>
<td>-0.01</td>
<td>-0.02</td>
</tr>
<tr>
<td>Year of Patient Identification</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1996</td>
<td>-0.05</td>
<td>-0.05</td>
<td>-0.14</td>
<td>-0.14</td>
</tr>
<tr>
<td>1998</td>
<td>0.02</td>
<td>0.02</td>
<td>-0.09</td>
<td>-0.09</td>
</tr>
<tr>
<td>1999</td>
<td>0.05</td>
<td>0.06</td>
<td>-0.23</td>
<td>-0.23</td>
</tr>
<tr>
<td>Member Type</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Employee</td>
<td>0.10</td>
<td>0.10</td>
<td>0.30</td>
<td>0.30</td>
</tr>
<tr>
<td>Spouse</td>
<td>0.14</td>
<td>0.14</td>
<td>0.43</td>
<td>0.43</td>
</tr>
<tr>
<td>Dependents</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12–18 years</td>
<td>0.21</td>
<td>0.21</td>
<td>0.22</td>
<td>0.22</td>
</tr>
<tr>
<td>Others</td>
<td>0.37</td>
<td>0.37</td>
<td>0.96</td>
<td>0.96</td>
</tr>
<tr>
<td>County Mean Household Income per $10,000</td>
<td>-0.01</td>
<td>-0.01</td>
<td>0.06</td>
<td>0.06</td>
</tr>
<tr>
<td>Number of Major Diagnosis Categories</td>
<td>0.22</td>
<td>0.22</td>
<td>0.18</td>
<td>0.18</td>
</tr>
<tr>
<td>Charlson Comorbidity Index</td>
<td>0.12</td>
<td>0.13</td>
<td>0.12</td>
<td>0.12</td>
</tr>
<tr>
<td>Asthma-Specific Comorbidities</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Allergic Rhinitis</td>
<td>0.04</td>
<td>0.07</td>
<td>-0.17</td>
<td>-0.17</td>
</tr>
<tr>
<td>Migraine</td>
<td>0.08</td>
<td>0.08</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>Depression</td>
<td>0.34</td>
<td>0.34</td>
<td>0.12</td>
<td>0.12</td>
</tr>
<tr>
<td>Gastrointestinal Disorder</td>
<td>0.17</td>
<td>0.17</td>
<td>0.19</td>
<td>0.19</td>
</tr>
<tr>
<td>Sinusitis</td>
<td>0.10</td>
<td>0.10</td>
<td>0.11</td>
<td>0.11</td>
</tr>
<tr>
<td>Anxiety</td>
<td>0.24</td>
<td>0.24</td>
<td>0.07</td>
<td>0.07</td>
</tr>
<tr>
<td>Treatment Indicator</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Controller</td>
<td>0.11</td>
<td>0.12</td>
<td>0.38</td>
<td>0.25</td>
</tr>
<tr>
<td>First Stage Residual</td>
<td>N/A</td>
<td>N/A</td>
<td>-0.18</td>
<td>-0.12</td>
</tr>
<tr>
<td>Residual</td>
<td></td>
<td></td>
<td>0.12</td>
<td>0.12</td>
</tr>
</tbody>
</table>

Table 6  Comparison of differences in outcome measures between the controller-only users and reliever-only users (standard errors are in parentheses)

<table>
<thead>
<tr>
<th>Cost</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unadjusted</td>
<td>$1,471 ($114)</td>
</tr>
<tr>
<td>OLS Estimate</td>
<td>$260 ($75)</td>
</tr>
<tr>
<td>IV Estimate with Weak IV</td>
<td>$270 ($613)</td>
</tr>
<tr>
<td>IV Estimate with Strong IV</td>
<td>$894 ($611)</td>
</tr>
<tr>
<td>IV Estimate with Redundant IV</td>
<td>$601 ($618)</td>
</tr>
</tbody>
</table>

IV, instrumental variable; OLS, ordinary least squares.
References

Probabilistic Sensitivity Analysis: Be a Bayesian

Hendriek C. Boshuizen, PhD, Pieter H. M. van Baal, PhD

National Institute of Public Health and the Environment (RIVM), Bilthoven, The Netherlands

ABSTRACT

Objective: To give guidance in defining probability distributions for model inputs in probabilistic sensitivity analysis (PSA) from a full Bayesian perspective.

Methods: A common approach to defining probability distributions for model inputs in PSA on the basis of input-related data is to use the likelihood of the data on an appropriate scale as the foundation for the distribution around the inputs. We will look at this approach from a Bayesian perspective, derive the implicit prior distributions in two examples (proportions and relative risks), and compare these to alternative prior distributions.

Results: In cases where data are sparse (in which case sensitivity analysis is crucial), commonly used approaches can lead to unexpected results. We show that this is because of the prior distributions that are implicitly assumed, namely that these are not as “uninformative” or “vague” as believed. We propose priors that we believe are more sensible for two examples and which are just as easy to apply.

Conclusions: Input probability distributions should not be based on the likelihood of the data, but on the Bayesian posterior distribution calculated from this likelihood and an explicitly stated prior distribution.

Keywords: Bayesian methods, maximum likelihood estimation, prior probability distribution, probabilistic sensitivity analysis.

Background

In economic evaluation employing modeling techniques, the model typically contains several unknown parameters [1]. The outcome of a study will depend on the values that are postulated for these parameters. These parameters are seldom based on hard facts; in most cases, there is uncertainty about their magnitude.

Probabilistic sensitivity analysis (PSA) has become the state-of-the-art method for determining the uncertainty in the outcomes of cost-effectiveness calculations for health-care interventions because of the uncertainty in input parameters. For instance, in the UK, the National Institute for Clinical Excellence recommendations state that PSA should be employed in order to yield unbiased estimates of expected net monetary benefits, and more importantly, to characterize decision uncertainty [2–4].

In a PSA [5], the uncertainty in each parameter is quantified in terms of a probability distribution of this parameter. One then carries out a Monte Carlo simulation, in which one randomly draws one value for each parameter from its probability distribution and then calculates the outcome corresponding to the set of parameters drawn. This process is repeated M times, yielding M outcome values that represent the distribution of the outcome values (for a given choice of the distributions of the input parameters of the model). PSA is a conceptually simple and intuitive method, and as such has considerable appeal. It can be seen as an implementation of Bayesian statistics, as the view that parameters have a probability distribution is a hallmark of the Bayesian outlook. Moreover, the decision context in which economics evaluations are carried out is essentially Bayesian [6,7].

Parameter values usually come from data that are collected in a single study, studies that combine data from multiple studies (meta-analysis), expert opinion, or applying complex methods of Bayesian evidence synthesis [8]. An important step in performing a PSA is defining the probability distribution to quantify the uncertainty in the input parameters. One guide in this field (Briggs et al. [9] hereafter called BCS) describes methods to fit distributions of parameter values “directly to the data.” In our experiences, this book, which grew out of a popular course on health economic modeling, is a popular guide for those carrying out cost-effectiveness analyses, and its methods are followed widely. Although it clearly states (and advocates) the Bayesian context of decision modeling, and describes the underlying theory, this guide’s final recommendations with respect to the choice of input probability distributions are not discussed from the viewpoint of the underlying Bayesian prior distributions. We are aware that a Bayesian perspective with respect to the choice of input probability distributions may scare some applied modelers.

However, as we will argue below, if seen from a Bayesian perspective, fitting parameter values “directly to the data” implies choices for prior distributions that need justification, as, in our opinion, more suitable alternative choices are possible. More importantly, we argue that in the case of some parameters, it is just as easy to estimate input probability distributions by assuming a more sensible alternative prior distribution. We will elaborate on two important types of parameters, namely a probability (or proportion), and a ratio (e.g., a relative risk [RR]). Although the prior distribution plays only a minor role whenever data are abundant, this is not always the case, especially given the current trend toward modeling of many specific subgroups [10]. Also, as our proposals are just as simple to use as those proposed in BCS, there are no practical reasons for not using them.

Our article is confined to the situation where the uncertainty of different parameters in the model is assumed to be mutually independent, for instance in cases where they are based on different sources. When multiple parameters are correlated, for instance because they are based on the same data source, the correlation between the uncertainty should also be taken into account. If not, the outcome of the PSA might be severely biased [11–13]. Also, in the case that input parameters are based on Bayesian evidence synthesis of trial data, there will be correlation between the estimates of individual parameters. In all these cases,
the PSA should sample from the joint posterior distribution of these parameters. This topic, although important in many situations, is not dealt with here.

The structure of the article is the following. First, we will discuss the implicit prior distributions for ratio and proportion parameters obtained by fitting distributions “directly to the data” (as described by BCS) from a Bayesian perspective. Next, we derive alternative prior distributions and show that their use leads to more intuitive input distributions for PSA.

**Being a Bayesian**

From a Bayesian point of view, the distributions that enter as input into the PSA are themselves “posterior distributions” based both on a “prior distribution” and on the data according to the following central formula in Bayesian statistics:

\[
p(\theta|x) \propto p(x|\theta) \times p(\theta)
\]

This formula states that the posterior distribution of the input parameter \(p(\theta|x)\) is proportional to the product of the likelihood of the data \(p(x|\theta)\) and the prior distribution of the parameter \(\theta\). The posterior distribution \(p(\theta|x)\) is the distribution we want to use in our PSA, as this gives the probability distribution of the parameter after we take the data into account.

Fitting input distributions directly to the data has considerable appeal as it seems to avoid the potentially messy business of having to choose a prior. However, this is deceiving: the methods as described in BCS implicitly assume a particular prior distribution. In their chapter on choosing distributions for input parameters, BCS give guidelines for the choice of distributions for (among others) proportions and ratios. In the case of a proportion, other possible priors are discussed by BCS in their technical appendix to that chapter, and distributions based on other priors are also applied in another article from Briggs et al. [14]. However, if we consider the underlying prior distributions, their final recommendations would not be our preference. In the case of a ratio, we propose an alternative prior that, as far as we know, has not yet been discussed in the health economic literature. Although we did not come up with this prior for that particular purpose, this alternative prior also obviates the shortcoming of the expected value of the input distribution not being equal to the point estimate as computed from the data. As the binomial proportion lends itself well to explaining the Bayesian method, we will discuss this first.

**Binomial Proportion**

For the binomial proportion, BCS advise using a beta distribution characterized by two parameters \(a\) and \(b\), and propose to use the number of positive outcomes observed in the data for \(a\), and the number of negative outcomes for \(b\). This approach implies a so-called Haldane prior distribution, which is proportional to \(p^{-1}(1-p)^{-1}\) (where \(p\) is the proportion). Alternative distributions to use in PSA are beta\((a,\ b)\) (assuming a uniform prior), or beta\((a+0.5,\ b+0.5)\) (assuming a Jeffrey’s prior, proportional to \(p^{-1/2}(1-p)^{-1/2}\) [15,16]. Figure 1 displays the probability density functions for these different prior distributions. To illustrate the differences between using alternative priors, we take the following simple example: In a trial, there are two arms, each with 100 patients. The object of study is (among other outcomes) the overall mortality, which in our example is rare: there is only one death in arm A, while there are 0 deaths in arm B. Figure 2 displays the posterior probabilities based on these data using the three different priors.

![Figure 1 Possible prior probability for a binomial proportion: Haldane prior (dashed line), Jeffrey’s prior (solid line), and uniform prior (dotted line). The Haldane prior has an infinitely large value at \(P = 0\) and \(P = 1\), and infinitely small values at \(P\)-values in between. Although all values in between 0 and 1 are infinitely small, they are not all equal, as shown by our plotted function.](image1)

![Figure 2 Resulting posterior probability density distributions for the data: 0 death in 100 participants (upper plot) and 1 death in 100 participants (lower plot) using a Haldane prior (dashed line), a Jeffrey’s prior (solid line), or a uniform prior (dotted lines).](image2)
The advantage of using the Haldane prior is that the posterior mean is equal to the empirical rate. This resembles the advantage of the maximum likelihood estimator (MLE) for a proportion in statistics. An MLE estimate gives the parameter value at which the likelihood of the data is at its maximum and also yields a result equal to the empirical rate. The argument therefore might be made that using a Haldane prior conforms more to common methods of statistical analysis. We therefore will briefly review the arguments behind the MLE.

First, the MLE can be seen as the maximum of the posterior distribution given a uniform prior. This posterior is given by the dotted line in Figure 2. The likelihood (apart from a normalizing constant) also looks like the dotted line in Figure 2. The MLE is, in statistical terms, thus equal to the posterior mode of a posterior distribution based on a uniform prior distribution, rather than to the posterior mean of this posterior distribution.

The philosophy leading to using the mode rather than the mean of the likelihood in MLE estimation has (among others) to do with what is called asymptotic unbiasedness and efficiency: the first means that if one repeats the experiment an infinite number of times, the MLE will give the true answer on average. For instance, if the real mortality is 1 in 200, then repeating the experiment with N = 100 many times will give 60.6% trials with 0 cases, 30.4% trials with one case, 7.6% trials with two cases, and 1.4% trials with three or more cases. This implies that 60.6% of the MLEs are equal to 0, 30.4% are equal to 0.01, 7.6% are equal to 0.02, and 1.4% are equal to 0.03 or more. On average, the MLE than is 0.005.

Efficiency means that the average amount that the estimate is “off target” (defined as the root mean square error) is as small as possible. This is related to the fact that if the size of the experiment increases (say you have 1000 patients with 0 or 10 deaths), the MLE stays in place (while the posterior mean will not). These characteristics, however, are not of much use in PSA: one does not have data from repeating the experiment many times: the knowledge that a particular estimate is “on average” unbiased and precise does not tell you whether it is right or precise in this particular case. Common sense will tell that an observation of 0 or 1 deaths in a group of 100 persons in most cases will not mean that the true mortality rate is exactly 0 or 0.01, but rather that it is compatible with many other “true” values beside 0.0 or 0.01. Also, although an observation of 0 deaths in 100 participants gives the same empirical rate as 0 in 10,000 participants, the true mortality rates that are compatible with these data will be lower in the last case.

In many cases, one is pretty sure that a particular event can occur in some patients and not in others, although its probability might be very low or very high. If you are sure that (based on evidence external to the collected data) one single case of the event has ever occurred, a possible event rate of zero is no longer possible. In this case, we would advocate a uniform prior (that assigns only a very small prior probability to the chance that the rate is zero or 1), and in this case the distribution to be used in PSA is beta(a + 1, b + 1). Briggs et al. [14] also noted the problem of using beta(a,b) for data with zero rates, and proposed using distributions based on such priors here.

However, there might be situations where one seriously questions whether the event could occur at all. For example, one might model the probability of hair color changing to pink as a result of taking a particular red-colored drug, which it is rumored to do on the Internet. The outcome is included in the model, but we do not think this effect possible from a biological point of view. Still, the Haldane prior does not seem prudent in this case either as this would make a zero rate already a certainty based on a study with only a few subjects. A better choice in this case is the Jeffrey’s prior (beta(0.5, 0.5)), an intermediate prior between the Haldane prior and uniform prior.

### Ratio Measures

Ratio measures, like odds ratios (ORs) and (hazard) rate ratios (HRs), are generally modeled on the logarithmic (log) scale, using mostly logistic regression (modeling the log OR) for count data or proportional hazards regression (modeling the log HR) for survival data. These models yield parameter estimates and parameter (co)variances on the log scale. When such ratio measures are used as input in a health economic model, BCS advise that their uncertainty is modeled by a normal distribution on the log scale, using the parameter estimate (log OR or log HR) as mean, and the standard error of the log(OR) or log(HR) as standard deviation. From a Bayesian point of view, this normal distribution is the posterior distribution of the log(OR) or log(HR) given the data used in the logistic or proportional hazards model and assuming a uniform prior distribution on log(OR) or log(HR), that is, assuming that all values of log(OR) or log(HR) have equal probability.

In modeling effects of interventions, however, one uses the ratio itself, not the log of the ratio. For instance, event rates are modeled by multiplying baseline event rates with the RR itself, not with its logarithm. So, it is the ratio itself, and not the log(ratio) that is usually proportional to the outcome in a health economic setting.

The implicit uniform prior on log(OR) or log(HR) (uniform prior on the log scale) at first sight seems reasonable: all values of log(OR) or log (HR) between minus infinity and plus infinity are equally likely and the average of this prior on the log scale is 0 (as each positive value of log(OR) will cancel out a corresponding negative value), corresponding to an OR or HR of 1. However, an informative prior one scale can be informative on another scale. On the linear scale, this prior implies a prior that is proportional to 1/OR or 1/HR. Despite the fact that this prior assigns probabilities that decline with increasing OR or HR values, this prior has an average of infinity. Thus, with a limited amount of empirical information, the average of the posterior will partly reflect the infinitely high average of the implicit prior.

To illustrate this, we will use the example of using an RR on larynx cancer in current smokers compared to never smokers, as taken from a meta-analysis [17] (RR 6.76 with a 95% confidence interval of [2.86, 16.0]). The statistical methods used in meta-analysis yield an estimate of log(RR) and a standard error of the estimated log(RR). Figure 3 shows three different prior distributions that could be considered in this case and which are discussed below. Figure 4 displays the corresponding posterior distributions (the potential input distributions for PSA) for this RR on larynx cancer that are based on these three different prior distributions. Two of the plotted prior distributions in Figure 3 are improper priors, that is, their integral is infinite. This means that they cannot be rescaled to a probability density, which must integrate to 1. In other terms, their scaling constant is infinitely small and we can therefore not plot these prior distributions. Here, we therefore plotted a function that is proportional to these priors, choosing an arbitrary scaling constant. Note, however, that an improper prior still can have a finite mean.

We start with discussing the prior that is implicit in the meta-analytic model, as recommended by BCS (the solid lines in Figs. 3 and 4). The implicit prior in this case is a uniform prior on the scale on which the analysis was carried out, in this case on the log(RR). The posterior probability of the RR in this case follows the lognormal distribution given by the solid curve in Figure 4.
This characteristic and that is also mathematically convenient is a lognormal distribution, characterized by its parameters $\mu_{prior}$ and $\sigma_{prior}$.

Such a lognormal prior on RR yields a posterior for RR (dashed line in Fig. 4) that is a lognormal distribution with

$$\mu_{post} = \mu_{lik} - (\mu_{lik} - \mu_{prior}) \frac{\sigma^2}{\sigma_{lik} + \sigma^2_{prior}}$$

and

$$\sigma_{post}^2 = \frac{\sigma_{prior}^2 \sigma_{lik}^2}{\sigma_{lik}^2 + \sigma_{prior}^2}$$

where $\mu_{prior}$ is the posterior mean of log(RR), $\mu_{lik}$ the point estimate for log(RR) from the meta-analysis, and $\sigma_{lik}$ the standard error of log(RR) from the meta-analysis. By choosing a sufficiently large value for $\sigma_{prior}$, one could make sure that the prior is sufficiently vague, so it would not play an important role. A conservative choice for $\mu_{prior}$ would set the average prior RR to 1, implying

$$\mu_{prior} = -\frac{1}{2} \sigma_{prior}^2$$

Thus,

$$\mu_{post} = \mu_{lik} - \left(\mu_{lik} - \sigma^2_{prior}\right) \frac{\sigma_{lik}}{\sigma_{lik}^2 + \sigma^2_{prior}} = \mu_{lik} - \sigma^2_{post} \frac{\mu_{lik}}{\sigma_{lik} + \sigma^2_{prior}} = \mu_{lik} - \frac{1}{2} \sigma^2_{post}$$

if $\sigma^2_{prior} \gg \sigma^2_{lik}$

$$\sigma^2_{post} = \sigma^2_{lik}$$

$$\mu_{post} = \mu_{lik} - \frac{1}{2} \sigma^2_{lik}$$

This prior is given by the dashed line in Figure 3. These are exactly the values for $\mu_{prior}$ and $\sigma_{prior}$ that BCS advise to use if one wants the expectation of the probabilistic distribution to correspond to the point estimate from a log link generalized linear model. It can be seen that this advice can be justified from using the vague prior described earlier.

In our example, the RR was larger than 1. If it would have been smaller than 1 (for instance, in case of a treatment effect), then using a uniform prior on the log scale would yield a posterior with a mean closer to 1 than the original point estimate (representing a smaller treatment effect). The alternative prior again would yield a mean equal to the original point estimate.

**Discussion**

The importance of a Bayesian outlook in health economic modeling has been stated before [6,18]. We feel that this should be extended to the topic of defining input distributions as used in sensitivity analysis. The reasoning that such input distributions should be founded on data only is tempting. Such reasoning has brought us likelihood-based statistics, but, unfortunately, for purposes other than summarizing evidence from data, the reasoning is deceptive. When working with the uncertainty of an estimated parameter, one uses the Bayesian concept of posterior probability, which is always based on an assumed prior. The choice of such a prior therefore should be discussed.

One important option of the Bayesian outlook is that one can specify “informative” prior distributions for the parameters of interest, using any relevant information available and without being constrained to a prespecified form for the posterior distributions. The resulting posterior distributions can then be used in
the economic analysis, which then is based on all existing relevant information. This is surely the optimal approach from a scientific point of view, but in economic evaluations, where financial stakes are often high, one prefers to use information based on hard facts as much as possible, in order to avoid the semblance of having manipulated the data.

In health economical modeling expectancies rather than modes or medians are the summarizing entities of interest. If input parameters have skewed distributions, it generally is the mean of this distribution rather than its mode or median that is related to the expected value of the outcome. It is therefore intuitive to “doctor” the distribution of an input parameter in a way that its mean is equal to our “best estimate” for that parameter (mostly an MLE estimator). We show that in case of a ratio measure, the justification of this method is that one implicitly chooses a prior distribution on the RR that has an average of 1 (which is conservative, but sensible), instead of the standard prior which has an average of zero on the log ratio scale, but an average of infinity on the ratio scale. The latter, in our opinion, does not make a sensible prior, as infinity is an unrealistic value.

For proportions, however, we do not advocate the approach of using a beta distribution with parameters equal to the number of positives \( a \) and negatives \( b \) in the data, despite its advantage of similarly forcing the distribution into having an expected value equal to the MLE. The reason is that it prescribes complete certainty in cases where zero positives or negatives are observed. Especially when events are rare and/or data sets are split up in many subpopulations, such situations are not uncommon.

Instead, we propose using a uniform prior distribution in those cases where one is sure that values of 0% and 100% are extremely unlikely, implying the use of a beta distribution with parameters \( a+1 \) and \( b+1 \). If values of 0% or 100% are likely, Jeffrey’s prior is a good choice, implying a beta-distribution with parameters \( a+0.5 \), \( b+0.5 \).

Although these recommendations are in the spirit of BCS, they nevertheless differ subtly from the final recommendations made in BCS. Its relevance to the outcome, however, in many cases will be minimal. First, in all cases where data are ample, the influence of the type of prior that has been chosen will vanish. Second, the uncertainty of modeling is only partly due to parameter uncertainty [20], and parameter uncertainty in turn is only partly due to the type of uncertainty on which we have focused here. The estimates of uncertainty given by a statistical model only reflect the uncertainty for populations that are completely similar to the study population in which the data were observed. In reality, there often is considerable uncertainty on whether study results can be generalized to the population of interest in the health economic problem. Ignoring this uncertainty influences results of PSA more than the subtle changes because of following the recommendations given here. Nevertheless, the recommendations given here are just as simple and just as easy to implement, and so there is no real reason not to use them.

We restricted the topic of this article to two simple, but frequently occurring types of input parameters. We did not discuss topics as including uncertainty on the distributional form of the input parameters [18,19] or using more complex methods of Bayesian evidence synthesis [7], mostly using MCMC. In the latter case, the posterior joint distribution from these methods can directly be used as input for PSA.

Summarizing, considering the (implicit) priors used in constructing input distributions for PSA, we recommend using a lognormal distribution for ratios, with median \( \exp(\log(\text{RR}) - 0.5 \times \text{se}^2) \) and standard error equal to \( \text{se} \), and a beta distribution with parameters \( a+1 \) and \( b+1 \) for proportions. Only in cases where a real proportion of 0 or 1 is anticipated, a beta distribution with parameters \( a+0.5 \) and \( b+0.5 \) might be preferable.

Source of financial support: The work was carried out as part of the regular work of the institute (no external funding).

References

16 Gelman A, Carlin JB, Stern HS, Rubin DB. Bayesian Data Analy-}
Value in Health’s Editorial Staff would like to thank the following for reviewing manuscripts for Value in Health, from the period June 2008–June 2009:

© 2009, International Society for Pharmacoeconomics and Outcomes Research (ISPOR) 1098-3015/09/1215 1215–1222
Reviewer Acknowledgment

Brennan, Victoria, MSc, RTI, Manchester, United Kingdom
Bridges, John, EP, B. Ec. (Hons) M. Ec. (Hons) PhD, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA
Brookes, Andrew, University of Glasgow, Glasgow, United Kingdom
Brixner, Diana, L, RPh, PhD, University of Utah, Salt Lake City, Utah, USA
Brodersen, John, PhD, University of Copenhagen, Institute of Public Health, Copenhagen, Denmark
Brooks, John, Matthew, PhD, University of Iowa, Iowa City, Iowa, USA
Brooks, Richard, Gibraltar, United Kingdom
Brown, Gary, Center for Value-Based Medicine, Flourtown, PA, USA
Brown, Jeffrey, PhD, Harvard Pilgrim Health Care/Harvard Medical School, Boston, MA, USA
Brüggenjürgen, Bernd, Charité University Medical Center, Berlin, Germany
Busch, Katharina, Abbott GmbH & Co KG, Ludwigshafen, Germany
Bullinger, Monika, PhD, Centre for Psychosocial Medicine, University Hospital Hamburg Eppendorf, Hamburg, Germany
Bushmakin, Andrew, Pfizer Inc, Groton, CA, USA
Busschbach, Jan, van, Erasmusmc, Rotterdam, The Netherlands
Caldwell, Deborah, M, University of Bristol, Bristol, United Kingdom
Campbell, Jonathan, University of Washington, Seattle, WA, USA
Campolina, Alessandro, Gonçalves, MD, MSc, Unifesp, Sao Paulo, Brazil
Cantrrell, Christopher, Ron, PhD, Hurdle Mills, NC, USA
Cappelli, Carlo, University of Brescia, Brescia, Italy
Carbone, Villemine, Dr., University of Pavia, PIRAEUS, Attiki, Greece
Carlson, Josh, John, PhD, University of Washington, Seattle, WA, USA
Carlton, Jill, University of Sheffield, Sheffield, United Kingdom
Caro, J, Jaime, UBC, Concord, MA, USA
Carr, Louise, University of YORK, York, United Kingdom
Castelnuovo, Emanuela, University of York, York, United Kingdom
Cella, David, PhD, Northwestern University, Evanston, IL, USA
Cerrito, Patricia, PhD, University of Louisville, Louisville, KY, USA
Charot, Isabelle, PhD, Pfizer Canada, Kirkland, Canada
Chapman, Ken, University Health Network, Toronto, Canada
Chassany, Olivier, AP-HP, Paris, France
Chen, Tianhui, Dr., German Cancer Research Center, Heidelberg, Germany
Chen, Yen-Fu, University of Birmingham, Birmingham, United Kingdom
Chhabra, Amit, Pfizer Inc, New York, NY, USA
Chhe, Wei Chu, MD, PhD, National Taiwan University, Taipei, Taiwan
Choi, Sang-Eun, Seoul National University, Seoul, Republic of Korea
Chomistek, Andrea, Harvard School of Public Health, Boston, MA, USA
Cipriano, Lauren, BSc, BA, Stanford University, Mountain View, CA, USA
Cleary, Susan, Health Economics Unit, Cape Town, South Africa
Conner-Spady, Barbara, PhD, University of Calgary, Calgary, Canada
Cook, Greg, A, PhD, Bristol Myers Squibb Australia Pty Ltd, Noble Park North, Victoria, Australia
Cook, Joseph, PhD, University of Washington, Seattle, WA, USA
Cooper, K, University of Southampton, Southampton, United Kingdom
Copley-Merriman, Kati, RTI-HS, Ann Arbor, MI, USA
Craig, Benjamin, Matthew, PhD, Moffitt Cancer Center, Tampa, FL, USA
Cramer, Joyce, Yale Univ School of Medicine, Orange, CT, USA
Crown, William, H, PhD, i3Innovus, Waltham, MA, USA
Davis, Peter, University of Auckland, Auckland, New Zealand
de Lissovoy, Gregory, MPH, PhD, United BioSource Corporation, Bethesda, MA, USA
de Portu, Simona, Faculty of Pharmacy, University of Naples, Federico II., Naples, Italy
de Vries, Rob, MSc, University of Groningen, Groningen, The Netherlands
Dennett, Susan, PhD, RPh, Strategic Health Outcomes, Carmel, IN, USA
Dere, Willard, Amgen, Thousand Oaks, CA, USA
Dias, Sofia, PhD, University of Bristol, Bristol, United Kingdom
Diaz, Mireya, PhD, Case Western Reserve University, Cleveland, OH, USA
Dijkeman, Lea, MSc, Academic Medical Center, Amsterdam, The Netherlands
Dodke, Richard, Philips-University, Marburg, Germany
Donnan, Peter, T, BA, BSc, MSc, PhD, University of Dundee, Dundee, United Kingdom
Doran, Christopher, UNSW, Sydney, New South Wales, Australia
Dorsey, Ray, CTCC, University of Rochester, Rochester, NY, USA
Dragomir, Alice, University of Montreal, Montreal, Canada
Drummond, Mike, University of York, York, United Kingdom
Eckermann, Simon, Douglas, PhD, Flinders University, Adelaide, South Australia, Australia
Edwards, Natalie, Health Services Consulting Corporation, Boxborough, MA, USA
Elbashir, Elamin, Merck & Co., Inc., North Wales, USA
Engstrom, Andreas, MSc, Pharmaceutical Benefits Board, Solna, Sweden
Reviewer Acknowledgment

Epstein, David, Mark, MSC, University of York, York, United Kingdom
Erder, Moshe, Haim, PhD, Forest Laboratories, Jersey City, NJ, USA
Eremenco, Sonya, United BioSource Corporation, Bethesda, USA
Erhart, Michael, PhD, University Clinic Hamburg-Eppendorf, Hamburg, Germany
Erickson, Steven, R., PharmD, University of Michigan, Ann Arbor, MI, USA
Faddy, Malcolm, Queensland University of Technology, Brisbane, Queensland, Australia
Faulli Munkerud, Siri, The Norwegian Medical Departement, Oslo, Norway
Fenwick, Elisabeth, University of Glasgow, Glasgow, United Kingdom
Ferrero, Montse, IMIM-Hospital del Mar, Barcelona, Spain
Fischer, Alastair, NICE, London, United Kingdom
Fliege, Herbert, Charite—University Medicine Berlin, Germany
French, Simon, Manchester Business School, Manchester, United Kingdom
Frick, Kevin, David, PhD, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA
Frost, Michelle, June, PhD, Schering-Plough Pty Limited Australia, Macquarie Park, New South Wales, Australia
Fu, Alex, Z., PhD, Cleveland Clinic, Cleveland, OH, USA
Fullerton, Pete, PhD, RPh, University of Washington, Seattle, WA, USA
Galani, Carmen, MD, MSc, HealthEcon, Basel, Switzerland
Gallego, Giselle, University of Technology Sydney, Broadway, New South Wales, Australia
Gandjour, Afshin, University of Cologne, Köln, Germany
Gao, Kun (Sue), MPH, PhD, Thousand Oaks, CA, USA
Gao, Xin, PhD, Alt Associates, Bethesda, MD, USA
Garcia-Altes, Anna, Barcelona, Spain
Garrison, Louis, P., PhD, University of Washington, Seattle, WA, USA
Geisler, Benjamin, Peter, MD, MPH, Beth Israel Deaconess Medical Center/Massachusetts General Hospital, Somerville, MA, USA
Getisio, Denis, Caro Research, Halifax, Nova Scotia, Canada
Globe, Denise, PhD, Amgen, Thousand Oaks, CA, USA
Goeree, Ron, McMaster University, Hamilton, Canada
Gold, Heather, Weill-Cornell Medical College, New York, NY, USA
Goring, Sarah, St. Paul's Hospital, Vancouver, Canada
Graf von der Schulenburg, J., Matthias, Gottfried Wilhelm Leibniz University Hanover, Hanover, Germany
Grau, Santiago, Hospital del Mar, Barcelona, Spain
Green, Colin, PhD, Institute for Health Service Research, Peninsula Medical School
Greenberg, Dan, PhD, Ben-Gurion University, Beer-Sheva, Israel
Griffin, Susan, York, North Yorkshire, United Kingdom
Griffiths, Ulla, LSHTM, London, United Kingdom
Grimison, Peter, NHMRC Clinical Trials Centre, University of Sydney, CAMPERDOWN, New South Wales, Australia
Groleau, Danielle, Novo Nordisk, Mississauga, Canada
Grutters, Jannke, PC, MSc, MAASTRO clinic, Maastricht, The Netherlands
Gu, Ning Yan, University of Southern California, Los Angeles, CA, USA
 Günther, Oliver, H., University of Leipzig, Leipzig, Germany
Gutierrez, Ben, AstraZeneca, Wilmington, PA, USA
Gwadry-Sridhar, Femida, PhD, Parkwood Hospital, London, Canada
Gwaltney, Chad, Brown University, Providence, RI, USA
Gyrd-Hansen, Dorte, University of Southern Denmark, Odense, Denmark
Hankins, M., Brighton & Sussex Medical School, Falmer, East Sussex, United Kingdom
Hannen, Janel, PhD, UW-Madison, Madison, WI, USA
Harrison, Jeffrey, University of Auckland, Auckland, New Zealand
Harrison, Mark, J., PhD, The University of Manchester, Manchester, United Kingdom
Hassan, Cesare, Rome, Italy
Hauera, A., Brett, PhD, RTI Health Solutions, RTP, NC, USA
Hay, Joel, PhD, USC, Los Angeles, CA, USA
Hayes, Richard, University of Melbourne, Melbourne, Victoria, Australia
Heeg, Bart, Phamrerit, Rotterdam, The Netherlands
Henk, Henry, J., PhD, Eden Prairie, MN, USA
Henriksen, Martin, Linköping University, Linköping, Sweden
Hess, Lisa, M., PhD, Indiana University Purdue University Indianapolis, Indianapolis, IN, USA
Hiligsmann, Mickael, MPH, MSc, University of Liege, Liege, Belgium
Hirano, Keisuke, University of Arizona, Tucson, AZ, USA
Hirsch, Jan, USSD, La Jolla, CA, USA
Hoch, Jeffrey, Cancer Care Ontario, Toronto, Canada
Holland, Margaret, L., MS, University of Rochester, Rochester, NY, USA
Hollenbeak, Christopher, S, Penn State College of Medicine, Hershey, PA, USA
Hoogendoorn, Martine, MSc, Erasmus MC, Rotterdam, The Netherlands
Hoque, Monjurul, KwaZulu-Natal Department of Health, Durban, South Africa
Hsiao, Fei-Yuan, University of Maryland School of Pharmacy, Baltimore, MD, USA
Reviewer Acknowledgment

Lichtenberg, Frank, Columbia University, New York, NY, USA
Lilja, Bengt, PhD, AstraZeneca, Molndal, Sweden
Lin, Wanchuan, Peking University, Beijing, China
Lloyd, Andrew, Oxford Outcomes Ltd, Oxford, United Kingdom
Lubeck, Deborah, PhD, Half Moon Bay, CA, USA
Luo, Xuemei, PhD, Global Outcomes Research, New London, CN, USA
MacKillop, James, UGA, Athens, GA, USA
Madan, Alok, PhD, MPH, Medical University of South Carolina, Charleston, SC, USA
Madan, Jason, University of Sheffield, Sheffield, United Kingdom
Mahoney, Elizabeth, Mid America Heart Institute of Saint Luke’s Hospital, Kansas City, MO, USA
Malone, Dan, University of Arizona, Tucson, AZ, USA
Manns, Braden, University of Calgary, Calgary, Canada
Marchetti, Monia, Ospedale C. Massaia, Asti, Italy
Marcinjak, Martin, Dybel, PhD, Eli Lilly and Company, Indianapolis, IN, USA
Marra, Carlo, A., PharmD, PhD, University of British Columbia, Vancouver, Canada
Martin, Bradley, C., PharmD, PhD, UAMS, Little Rock, AR, USA
Marx, Steven, RPh, MS, PharmD, BSc, Abbott, Abbott Park, IL, USA
Mathias, Susan, MPH, Health Outcomes Solution, Winter Park, FL, USA
Mattias, Neyt, Belgian Health Care Knowledge Centre (KCE), Brussels, Belgium
Matza, Louis, UBC, Bethesda, MD, USA
Mauldin, Patrick, Medical University of South Carolina, Charleston, SC, USA
Mauskopp, Josephine, PhD, Latrometrics, Durham, NC, USA
McBride, Doreen, Charité University Medical Centre, Berlin, Germany
McCullam, Patrick, BS, PharmD, Eli Lilly, Indianapolis, IN, USA
McCombs, Jeff, University of Southern California, School of Pharmacy, Los Angeles, CA, USA
McEntegart, Damen, MSC, Nottingham, United Kingdom
McGhan, William, E, PharmD, PhD, University of Sciences in Philadelphia, Philadelphia, PA, USA
McKenna, Stephen, Patrick, Galen Research, Manchester, United Kingdom
McKinnon, Peggy, PharmD, BCPS, FCCP, Ballwin, MO, USA
McLaughlin, P.J., Penn State University College of Medicine, Hershey, PA, USA
McLeod, Lori, RTI Health Solutions, Research Triangle Institute, NC, USA
McVeigh, Tina, NYC Dept. of Health and Mental Hygiene, New York, NY, USA
Mealing, Stuart, Oxford Outcomes Consultancy, Oxford, United Kingdom
Mebrye, Rick, J., Pfizer, Moorhead, MN, USA
Membe, Stephen, B(H)Sc, MDE, CADTH, Ottawa, Canada
Metcalf, Hugh, Newcastle University, Newcastle, United Kingdom
Migliaccio-Walle, Kristen, B.S., Caro Research, Concord, MA, USA
Mihaylova, Boy, University of Oxford, Oxford, United Kingdom
Miksd, Rebecca, Beth Israel Deaconess Medical Center, Boston, MA, USA
Miller, Jeffrey, D., Abt Bio-Pharma Solutions, Lexington, MA, USA
Minshall, Michael, E., IMS-Health, Noblesville, IN, USA
Mitchell, Matthew, D., PhD, Diagnostic Research Design & Reporting, Glenside, PA, USA
Mordini, Margaret, RTI Health Solutions, Ann Arbor, MI, USA
Morris, Louis, PhD, Louis A. Morris & Associates, Dix Hills, NY, USA
Moss, Sue, Institute of Cancer Research, Surrey, United Kingdom
Mueller, Dirk, DMX, Cologne, Germany
Müller-Nordhorn, Jacqueline, MD, DPH, Charité University Medical Center, Berlin, Germany
Mullings, Anthony, DM, UW, Kingston, Jamaica
Mullins, C. Daniel, PhD, University of Maryland, Baltimore, MD, USA
Murawski, Matthew M., RPh, PhD, Purdue University, West Lafayette, IN, USA
Neil, Nancy, PhD, ICON Clinical Research, San Francisco, CA, USA
Nelso, Linda, Merck & Co., Inc., North Wales, PA, USA
Neumann, Peter, J., ScD, Tufts Medical Center, Boston, MA, USA
Nichol, Michael, B., University of Southern California, Los Angeles, CA, USA
Nixon, Annabel, Oxford Outcomes, Oxford, United Kingdom
Noyes, Katia, University of Rochester, Rochester, NY, USA
Nuijten, Mark, J.C., PhD, Imta, Erasmus University and Ars Accessus Medica, Rotterdam, The Netherlands
O’Donnell, John, C., PhD, AstraZeneca LP, Wilmington, DE, USA
O’Morain, Colm, Trinity College Dublin, The University of Dublin, Dublin, Ireland
Oderka, Gary, University of Utah, Salt Lake City, UT, USA
Oishi, Takeko, PhD, Boston, MA, USA
Onwudiwe, Nneka, C., PharmD, Silver Spring, MD, USA
Osterhaus, Jane, T., PhD, Wasatch Health Outcomes, Park City, UT, USA
Ouwens, Mario, Mapi Values, Utrecht, The Netherlands
Ozanne, Elissa, MGH, Boston, MA, USA
Reviewer Acknowledgment

OWMINKOWSKI, Ron, Thomson Healthcare, Ann Arbor, MI, USA
Palmer, Steve, University of York, York, United Kingdom
Pattar, Mari, University of Wisconsin, WI, USA
PASHOS, Chris, L., PhD, HERQuLES Health Economic Research, Lexington, MA, USA
PASQUALE, Margaret, K., PhD, Procter & Gamble Pharmaceuticals, Cincinnati, OH, USA
Paul, John, E., PhD, Univ. of North Carolina at Chapel Hill, Chapel Hill, NC, USA
Pawar, Vivek, PhD, i3 Innovus, Medford, MA, USA
Payne, Katherine, PhD, The University of Manchester, Manchester, Lancashire, United Kingdom
Penson, David, University of Southern California/ Noris Cancer Center, Los Angeles, CA, USA
Petrienco, Michele, PhD student, University of Turin, Turin, Italy
Phillip, Roy, K., Mid-Western Regional & Maternity Hospitals, Limerick, Munster, Ireland
Pickard, A. Simon, PhD, University of Illinois at Chicago, Chicago, IL, USA
Priest, Ginny, Griffith University, Brisbane, Queensland, Australia
Prosser, Lisa, A., PhD, Harvard Pilgrim Health Care, Boston, MA, USA
Qu, Yongming, Eli Lily and Company, Indianapolis, IN, USA
Quon, Jeffrey, St. Paul’s Hospital, Vancouver, Canada
Raish, Dennis, W., VA Cooperative Studies Program Clinical Research Pharmacy, Albuquerque, NM, USA
Ramsey, Scott, D., MD, PhD, Fred Hutchinson Cancer Research Center, Seattle, WA, USA
RAPP, Thomas, PhD, University of Maryland School of Pharmacy, Baltimore, MD, USA
Rattinger, Gail, University of Maryland, Baltimore, MD, USA
Ravens-Sieberer, Ulrike, University of Bielefeld, Bielefeld, Germany
Ray, Joshua, A, MSc, F.Hoffmann-La Roche, Ltd., Basel, Basel-Stadt, Switzerland
Redekop, Ken, Erasmus MC, Rotterdam, The Netherlands
Reed, Shelby, D., PhD, Duke University, Durham, NC, USA
Reeder, Clairborne, Eugene, PhD, Xcenda, Columbia, FL, USA
Reilly, MC, Reilly Associates, New York, NY, USA
Richardson, Jennifer, London, United Kingdom
Rive, R., Lundbeck SAS, Paris, France
Roccchi, Angela, Axia Research, Hamilton, Canada
Rodriguez-Blazquez, Carmen, Carlos III Institute of Health, Madrid, Spain
Roser, Hubertus, AiM Assessment in Medicine e.K., Schopheim, Germany
Rupnow, Marcia, ET, Ethicon, Inc., Somerville, NJ, USA
Rutheemann, Jens, MD, MBA, Duisburg, Germany
Rutten, Maureen, Erasmus University/Erasmus Medical Centre, Rotterdam, The Netherlands
Sabes-Figuera, Ramon, Institute of Psychiatry, King’s College London, London, United Kingdom
Saez, Marc, University of Girona, Girona, Spain
Salas, Maribel, MD, DSc, MSc, AstraZeneca, Wilmington, PA, USA
Sankaranarayanan, Jayashri, MPharm, PhD, University of Nebraska Medical Center, Omaha, NE, USA
Schafer, Delia, Kendall, St Leonards, New South Wales, Australia
Schulman, Kevin, A., Duke University, Durham, NC, USA
Schwenkglenks, Matthias, M.A., University of Basel, Basel, Switzerland
Scotland, Graham, Stewart, MSc, University of Aberdeen, Aberdeen, Aberdeenhire, United Kingdom
Scuffham, Paul, A., PhD, Griffith University, Brisbane, Queensland, Australia
Sculpher, Mark, PhD, University of York, York, United Kingdom
Shah, Bijal, M., BPharm, PhD, Touro University—CA, Vallejo, CA, USA
Shah, Nilay, PhD, Mayo Clinic, Rochester, MA, USA
Shaya, Fadwa, University of Maryland Baltimore, Baltimore, MD, USA
Shi, Jinhai, MD, PhD, Amgen Inc., Oak Park, CA, USA
Shi, Liheng, PhD, MScPharm, Tidane University, NO, USA
Shih, Ya-Chen, Tina, PhD, University of Texas M.D. Anderson Cancer Center, Houston, TX, USA
Shinogle, Judith, A., PhD, MSc, University of Maryland, College Park, MD, USA
Sidhu, Manpreet, K., PhD BA (Hons), Astellas Pharma Europe Ltd, Staines, Middx, United Kingdom
Silva, Jose, UFMG, Belo Horizonte, Brazil
Simons, Steven, Katholieke Universiteit Leuven, Leuven, Belgium
Sintonen, Harri, PhD, University of Helsinki, Helsinki, Finland
Skredgel, Chris, MDE, Dalhousie University, Halifax, Canada
Sloan, Frank, Duke University, Durham, NC, USA
Slovacek, Ladislav, University of Defence, Hradec Kralove, Czech Republic
Smoalen, Harry, J., BSEE, MSIE, Medical Decision Modeling Inc., Indianapolis, IN, USA
 Snyder, Claire, F, PhD, Johns Hopkins School of Medicine, Baltimore, MD, USA
Snyder, Steve, Lehigh University, Bethlehem, PA, USA
Soares, Marta, University of York, York, United Kingdom
Sogaard, Rikke, MSc, MPH, PhD, University of Southern Denmark, Odense, Denmark
Sorbero, Melony, PhD, MS, MPH, RAND, Pittsburgh, PA, USA
Spackman, Eldon, University of Washington, Seattle, WA, USA
Reviewer Acknowledgment

Stahl, James, MGH-Inst. for Tech. Assessment, Boston, MA, USA
Stipiner, John, University of Colorado Health Sciences Center, Denver, CO, USA
Stern, Craig, Stephen, PharmD, MBA, Pro Pharma, Northridge, CA, USA
Streten, Lotte, Brunel University, Uxbridge, United Kingdom
Stevanovic, Dejan, Health Care Center Odzaci, Odzaci, Serbia
Stevens, John, University of Sheffield, Sheffield, United Kingdom
Stevenson, Matt, University of Sheffield, Sheffield, United Kingdom
Stollenwerk, Bjoern, Helmholtz Zentrum Muench, Neuherberg, Germany
Strassels, Scott, PharmD, PhD, BCPS, University of Texas, Austin, TX, USA
Stull, Donald, PhD, United BioSource Corporation, Bethesda, MD, USA
Sullivan, Patrick, University of Colorado—Denver School of Pharmacy, Aurora, CO, USA
SUMMERS, KENT, H., PhD, Purdue University School of Pharmacy, West Lafayette, IN, USA
Sun, Shawn, Takeda Global Research & Development, Deerfield, IL, USA
Sundaram, Murali, MS, PhD, QualityMetric Incorporated, Lincoln, MA, USA
Sung, Vivian, Alpert Medical School at Brown University/Women and Infants Hospital, Providence, RI, USA
Symmons, Deborah, University of Manchester, Manchester, United Kingdom
Symonds, Tara, Pfizer Ltd, Sandwich, Kent, United Kingdom
Szabo, Sheilagh, Oxford Outcomes Ltd., Vancouver, Canada
Tappenden, Paul, BA, MSc, University of Sheffield, Sheffield, South Yorkshire, United Kingdom
Tarride, Jean-Eric, McMaster University, Hamilton, Canada
Taylor, Thomas, Research Triangle Institute, Ann Arbro, MI, USA
Thomas, Simu, Novartis Pharmaceuticals, East Hanover, NJ, USA
Thompson-Coon, Jo, Universities of Plymouth and Exeter, Exeter, United Kingdom
Toscani, Michael, PharmD, New Hope, PA, USA
Treeak, Charoen, Faculty of Pharmacy, Srinakharinwirot University, Nakhonmayok, Ongkarak, Thailand
Trueman, Paul, University of York, York, United Kingdom
Turk, Florian, Novartis Pharma AG, Basel, Switzerland
van Asselt, Antoinette, Dorothea Isabelle, MSc, University Hospital Maastricht, Maastricht, The Netherlands
van der Pol, Marjon, University of Aberdeen, Aberdeen, United Kingdom
van Gestel, A., MSc, Academic hospital Maastricht, Maastricht, The Netherlands
van Nispen, RM, EMGO Institute, VU University Medical Center Amsterdam, Amsterdam, The Netherlands
van Schayck, Onno, CAPHRI, Maastricht, The Netherlands
Varni, James, W., PhD, Texas AM & University, College Station, TX, USA
Verdier, Angela, Mazarat-Aurouze, France
Vickrey, Barbara, MD, UCLA, Los Angeles, CA, USA
Viney, Rosalie, University of Technology, Sydney, Sydney, New South Wales, Australia
Vogenberg, Randy, PhD, RPh, EPSR, Sharon, RI, USA
Voller, Heinz, Klinik am See, Rüdersdorf, Germany
von Feber, Liseotte, University of Dusseldorf, Dusseldorf, Germany
Wagner, Todd, H., VA Palo Alto, Menlo Park, CA, USA
Wailoo, Allan, University of Sheffield, Sheffield, United Kingdom
Walker, David, Ray, PhD, Baxter Healthcare, McGaw Park, IL, USA
Walker, Simon, University of York, York, United Kingdom
Walter, Otto, Westfälische Wilhelms-Universität, Münster, Germany
Wang, Jingshu, PhD, Baltimore, MD, USA
Waters, Heidi, Christina, MS, MBA, Centocor, Inc, Horsham, PA, USA
Watkins, John, Premera Blue Cross, Montlake Terrace, WA, USA
Weatherly, Helen, University of York, York, United Kingdom
Weis, Kathleen, Ann, DrPh, PhD, Pfizer Inc., New York, NY, USA
Welte, Robert, PhD, MPH, MSc, GlaxoSmithKline, Munich, Germany
Welton, Nicky, University of Bristol, Bristol, United Kingdom
Wertheimer, Albert, Temple University, Philadelphia, PA, USA
Whalley, Diane, University of Manchester, Manchester, United Kingdom
Wild, Diane, Oxford Outcomes Ltd, Oxford, United Kingdom
Williams, Valerie, RTI Health Solutions, Research Triangle Park, NC, USA
Willis, Michael, IHE, Lund, Sweden
Wilke, Richard, J., PhD, Pfizer, Inc, Bridgewater, NJ, USA
Wilson, Leslie, University of California, San Francisco, San Francisco, CA, USA
Wittenberg, Eve, PhD, Schneider Institutes for Health Policy, Waltham, MA, USA
Wolowicz, Sobrel, PhD, RTI Health Solutions, Manchester, United Kingdom
Wonder, Michael, John, B Sc (Hons), B Pharm, Novartis Pharmaceuticals Australia Pty Ltd, North Ryde, New South Wales, Australia
Wu, Eric, Q., Analysis Group, Inc, Boston, MA, USA
Wu, Joanne, University of Southern California, Los Angeles, CA, USA
Xie, Feng, McMaster University, Hamilton, Canada
Reviewer Acknowledgment

XIE, YANG, University of Iowa, Iowa City, IA, USA
YANG, MIN, MD, PhD, QualityMetric, Inc., Lincoln, RI, USA
YAO, GUIQING, LILY, University of Birmingham, Birmingham, United Kingdom
YOO, JEAN, U.S. Department of Veterans Affairs, Menlo Park, CA, USA
YOUNG, NICHOLAS, J., University of Colorado, Denver, CO, USA
YOUNG, TRACEY, University of Sheffield, Sheffield, United Kingdom
YU, ANDREW, PENG, Analysis Group, Boston, MA, USA
ZAMORA, JAVIER, PhD, Hospital Ramón y Cajal, Madrid, Saint Helena, Spain
ZANG, WENBIN, Southwestern Univ. of Finance and Economics, Chengdu, Sichuan, China
ZARATE, VICTOR, MD, MSc, Pontificia Universidad Católica de Chile, Santiago, Chile
ZAWADZKI, REZI, Palo Alto, CA, USA
ZHANG, JAMES, X., Virginia Commonwealth University, Richmond, VA, USA
ZHAO, ZHONGYUN, Johnson & Johnson, Raritan, NJ, USA
ZUBEK, VALENTINA, BAYER, PhD, Aureon Laboratories, Yonkers, NY, USA
ZUPANCIC, JOHN, Harvard University, Cambridge, MA, USA