VALUE IN HEALTH 17 (2014) 143–156

Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/jval

ISPOR TASK FORCE REPORTS

A Questionnaire to Assess the Relevance and Credibility of Observational Studies to Inform Health Care Decision Making: An ISPOR-AMCP-NPC Good Practice Task Force Report

Marc L. Berger, MD[1], Bradley C. Martin, PharmD, PhD[2,], Don Husereau, BScPharm, MSc[3,4,5], Karen Worley, PhD[6], J. Daniel Allen, PharmD[7], Winnie Yang, PharmD[8], Nicole C. Quon, PhD[9], C. Daniel Mullins, PhD[10], Kristijan H. Kahler, PhD, RPh[11], William Crown, PhD[12]*

[1]*Real World Data and Analytics, Pfizer, New York, NY, USA;* [2]*Division of Pharmaceutical Evaluation and Policy, Little Rock, AR, USA;* [3]*Institute of Health Economics, Edmonton, AB, Canada;* [4]*Department of Epidemiology and Community Medicine, University of Ottawa, Ottawa, ON, Canada;* [5]*University for Health Sciences, Medical Informatics and Technology, Tirol, Austria;* [6]*Comprehensive Health Insights, Humana, Inc., Cincinnati, OH, USA;* [7]*OmedaRx, Portland, OR, USA;* [8]*Blue Shield of California, Woodland Hills, CA, USA;* [9]*Healthcare Quality, Optimer Pharmaceuticals, Inc., Jersey City, NJ, USA;* [10]*University of Maryland, School of Pharmacy, Pharmaceutical Health Services Research, Baltimore, MD, USA;* [11]*HEOR, Novartis Pharmaceuticals Corporation, East Hanover, NJ, USA;* [12]*OptumLabs, Cambridge, MA, USA*

ABSTRACT

Evidence-based health care decisions are best informed by comparisons of all relevant interventions used to treat conditions in specific patient populations. Observational studies are being performed to help fill evidence gaps. Widespread adoption of evidence from observational studies, however, has been limited because of various factors, including the lack of consensus regarding accepted principles for their evaluation and interpretation. Two task forces were formed to develop questionnaires to assist decision makers in evaluating observational studies, with one Task Force addressing retrospective research and the other Task Force addressing prospective research. The intent was to promote a structured approach to reduce the potential for subjective interpretation of evidence and drive consistency in decision making. Separately developed questionnaires were combined into a single questionnaire consisting of 33 items. These were divided into two domains: relevance and credibility. Relevance addresses the extent to which findings, if accurate, apply to the setting of interest to the decision maker. Credibility addresses the extent to which the study findings accurately answer the study question. The questionnaire provides a guide for assessing the degree of confidence that should be placed from observational studies and promotes awareness of the subtleties involved in evaluating those.

Keywords: bias, checklist, comparative effectiveness research, confounding, consensus, credibility, decision making, prospective observational study, quality, questionnaire, relevance, retrospective observational study, validity.

Introduction

Four Good Practices task forces developed consensus-based questionnaires to help decision makers evaluate: 1) prospective and 2) retrospective observational studies, 3) network meta-analysis (indirect treatment comparison), and 4) decision analytic modeling studies with greater uniformity and transparency [1,2]. The primary audiences of these questionnaires are assessors and reviewers of health care research studies for health technology assessment, drug formulary, and health care services decisions who have varying levels of knowledge and expertise. As discussed above, the prospective and retrospective observational task forces, while ultimately coordinating their efforts, worked independently. The work products of these two task forces were quite similar, and based on feedback from reviewers a single questionnaire for observational studies was developed.

Although randomized controlled trials (RCTs) are often sought to inform health system decisions, there is increasing recognition of the limitations of relying on RCTs alone. These studies may not exist because of financial, ethical, or time limitations or if available they may lack sufficient information to guide decision making, which needs input from real-world conditions, diverse populations, or practice settings. Other types of studies, such as observational, modeling, and network meta-analysis, are increasingly sought to fill this gap [3]. There may be barriers, however, to the use of these studies due to the limited number of accepted principles for their

* *Address correspondence to*: Bradley C. Martin, Division of Pharmaceutical Evaluation and Policy, 4301 West Markham Street, Slot 522, Little Rock, AR 72205.

E-mail: bmartin@uams.edu.

1098-3015/$36.00 – see front matter

http://dx.doi.org/10.1016/j.jval.2013.12.011

Background to the Task Force

On May 21, 2011, the Board of Directors approved, in principle, ISPOR's participation with the Academy of Managed Care Pharmacy (AMCP) and the National Pharmaceutical Council (NPC) in the Comparative Effectiveness Research Collaborative Initiative (CER-CI) for advancing appropriate use of outcomes research evidence to improve patient health outcomes. ISPOR's contribution to the CER-CI was to develop articles on how to assess prospective and retrospective observational studies, indirect treatment comparison (network meta-analysis), and decision analytic modeling studies to inform health care decision making. Four Good Practice task forces were created to develop these articles. Task Force Chairs were identified from leaders of ISPOR Good Research Practices task forces. Each Task Force consisted of two members from the AMCP, the NPC, and the ISPOR.Each Task Force met independently via teleconference. In addition, the Task Force Chairs met via teleconferences and face-to-face meetings held on April 20, 2012 (San Francisco, CA, USA), June 3, 2012 (Washington, DC, USA), June 28-29, 2012 (Boston, MA, USA), November 4, 2012 (Berlin, Germany), and May 21, 2013 (New Orleans, LA, USA), to coordinate a common outline and format for these articles. A focus group representing the US formulary decision-making community (22 participants) was convened April 20, 2012, at the AMCP Meeting, San Francisco, CA, USA, for feedback on the draft outline, format, and content of the assessment articles. The content of these reports was presented for comment at the ISPOR Annual International Meetings held June 4, 2012, and May 22, 2013, and the European Congress held November 5 and 6, 2012. Draft prospective observational studies and retrospective observational studies Task Force reports were sent for comment to their respective review group. Comments for each group were considered, and final draft reports were sent to the ISPOR membership for comment on September 5, 2013. Overall, there were 82 written comments for the retrospective Task Force report and 57 written comments for the prospective Task Force report. A number of reviewers commented on the overlap between the two reports. Based on these comments, the prospective and retrospective Task Force reports were combined. All written comments are published on the ISPOR Web site, which can be accessed via the Research menu on ISPOR's home page: http://www.ispor.org. The final report was submitted to *Value in Health*.

evaluation and interpretation. There is a need for transparent and uniform ways to assess their quality [4]. A structured approach reduces the potential for subjectivity to affect the interpretation of evidence and can promote consistency in decision making [5].

Previous tools, including grading systems, scorecards, and checklists, have been developed to facilitate structured approaches to critically appraising clinical research [6,7]. Some are elaborate, requiring software and deliberation among a broad range of experts [8], whereas others are very simple, using scoring systems best suited for randomized clinical trials [9]. With the goal of creating a questionnaire that would promote awareness of the issues related to alternative study designs to a wide audience of users, it was believed that a simple, time-efficient, user-friendly questionnaire incorporating epidemiological principles is needed to give decision makers the means to more appropriately consider results from alternative research designs.

Development of this questionnaire was informed by previous efforts and with several guiding principles derived from a focus group of payers. First, questionnaires would be used by individuals with a broad range of expertise including many without in-depth training in study design and statistics. Second, questionnaires had to be sufficiently comprehensive to promote awareness of the appropriate application of different study designs to decision making; we also sought to produce a questionnaire that would include explanations of core concepts and prompt users to obtain additional education on the underlying methodologies. Last, the use of questionnaires would need to be facilitated by comprehensive educational programs.

Although some might argue that there is no clear distinction between retrospective and prospective observational studies, these were separately considered by two previous ISPOR Good Research Practices task forces [10–13]. The working definitions of these task forces were used in guiding the work in this new initiative.

- *Prospective observational studies* were defined as those in which participants are not randomized or otherwise assigned to an exposure and for which the consequential outcomes of interest occur *after* study commencement (including creation of a study protocol and analysis plan, and study initiation). They are often longitudinal in nature. Exposure to any of the interventions being studied may or may not have been recorded before the study initiation such as when a prospective observational study uses an existing registry cohort. Exposure may include a pharmaceutical intervention, surgery, medical device, prescription, or decision to treat.
- *Retrospective observational studies* were defined as those that use existing data sources in which both exposure and outcomes have already occurred.

Prospective observational studies have the potential advantage of collecting the specific study measures desired; retrospective studies use existing data sets but have the advantage of generally being less costly and require less time to conduct. Ultimately the principles identified by the two task forces for evaluating prospective and retrospective observational studies were sufficiently similar that a common questionnaire was adopted; however, the distinction between prospective and retrospective perspectives can be important in using this questionnaire and explanations provided with the questionnaire draw on both perspectives. Because the focus of these efforts was specifically on comparative effectiveness research, considerations applying to pharmacovigilance, safety surveillance, and economic analyses were not addressed.

Questionnaire Development

The first issue addressed was whether the questionnaires developed for this joint initiative should be linked to checklists, scorecards, or annotated scorecards. Concerns were raised that a scoring system may be misleading if it did not have adequate measurement properties. Scoring systems have been shown to be problematic in the interpretation of randomized trials [14].

An alternative to a scorecard is a checklist. The Task Force members, however, believed that checklists might also mislead users because a study may satisfy nearly all the elements of a checklist and still harbor "fatal flaws" (defined as design, execution, or analysis elements of the study that by themselves may significantly undermine the validity of the results). Moreover, users might tend to add the number of positive or negative elements and convert it to a score, and then apply the score to their overall assessment of the evidence implicitly (and incorrectly) giving equal

weight to each item. In addition, the acceptability of a study finding may depend on other evidence that addresses the specific issue or the decision being made. A questionnaire without an accompanying score or checklist was felt to be the best way to allow analysts to be aware of the strengths and weaknesses of each piece of evidence and apply their own reasoning.

Questions were developed on the basis of a review of items in previous questionnaires and guidance documents, previous ISPOR Task Force recommendations [10–13], and methods and reporting guidances (including Grading of Recommendations, Assessment, Development and Evaluation, STrengthening the Reporting of OBservational studies in Epidemiology, and the European Network of Centres for Pharmacoepidemiology and Pharmacovigilance) [3,15–35]. The retrospective committee identified all items and themes in these guidance documents and created a list of 174 items. Themes assessing observational study quality not originally in question format were reworded in yes/no question format. Items from previous guidance were categorized and redundant themes were removed. The 174 items were rated by the committee members across five domains: credibility, relevance, feasibility, clarity, and uniqueness. Items that were rated low on these five domains were considered for removal by the committee by consensus of the committee members, resulting in 99 items. The prospective committee followed the same process and created a similar list. After preliminary user testing, items were further reduced and grouped into common conceptual domains for each of the prospective and retrospective questionnaires. At a meeting of the chairs of the four task forces, the domains across all four questionnaires were harmonized as much as possible, and then grouped into two common sections —"Relevance" and "Credibility"—on the basis of key elements essential to evaluating comparative effectiveness evidence.

Four identical questions were developed for the relevance section for this and the other questionnaires developed in this joint initiative. Credibility was further divided into several key domains. For this questionnaire to obtain broad acceptance, and based on early feedback, it was decided to limit the questionnaire length to about 30 items. Whenever possible, efforts were made to avoid the use of jargon and to use similar wording, where appropriate, across all the questionnaires. There was substantial overlap in the design and flow of the questionnaires for prospective and retrospective studies. Following the suggestion by a number of reviewers to combine the two observational study questionnaires, collaboration between the two observational study task forces led to an agreement on a single set of questions and their wording. The logic flow to assess credibility is shown in Figure 1.

How to Use the Questionnaire

Questions that fall under the main categories of relevance and credibility appear in Table 1. Explanations of each question along with specific definitions and issues to consider (formulated as subquestions) are provided in the following section, to facilitate understanding of the appropriate use of the questionnaire.

The questionnaire is an organized series of items with "yes" "no" binary response choices with strengths and weaknesses coded depending on the wording of the item. Users are encouraged to evaluate each item and determine if the study met or did not meet the criteria captured in each item, however, a series of "can't answer" response choices are available to capture instances when the reporting is inadequate or when the user does not have sufficient training to evaluate the item. Upon completion of questions in the relevance section, users are asked to rate whether the study is sufficient or insufficient to inform their decision making. If a study is not considered sufficiently relevant, a user can then opt to truncate the review of its credibility. In the credibility category, the user answers a series of individual yes/no items and then rates each domain as a "strength," a "weakness," or "neutral." On the basis of these evaluations, the user then similarly rates the credibility of the research study as either "sufficient" or "insufficient" to inform decision making. For some questions in the credibility section, a user will be notified that he or she had detected a "fatal flaw." The presence of a fatal flaw suggests a significant opportunity for the findings to be

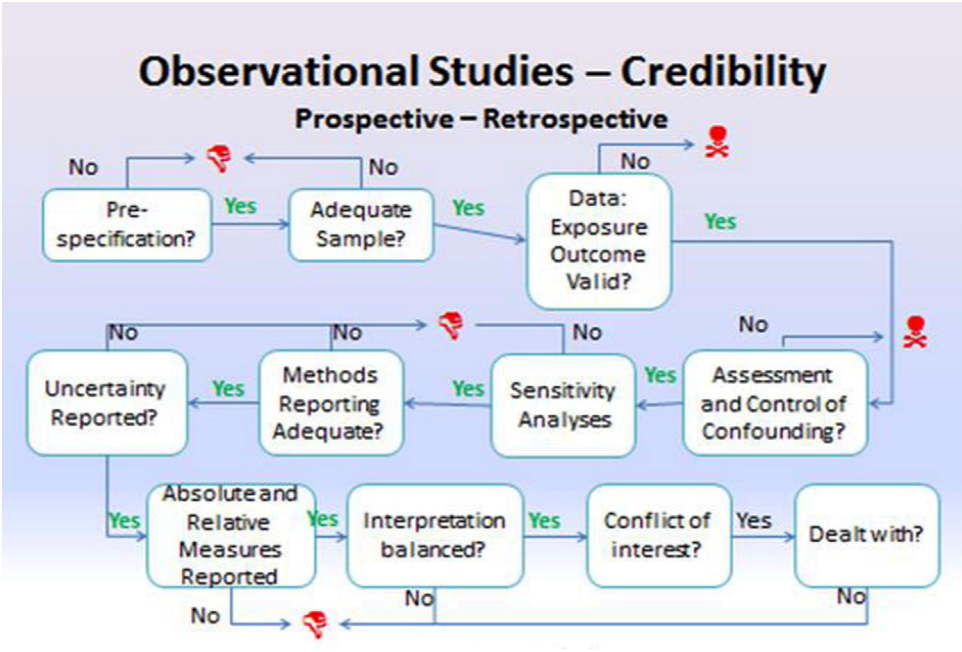

Fig. 1 – Summary flowchart for observational study assessment questionnaire. Red thumbs down icons indicate that a "weakness" had been detected in one of the elements that support credibility. Red skull and cross-bones icons indicate that a potential "fatal flaw" had been detected.

Table 1 – Questionnaire* to assess the relevance and credibility of a prospective observational study.

S. no.	Question	Strength	Weakness	Can't answer			
				Not applicable	Not reported	Not enough information	Not enough training
Relevance questions							
1	Is the population relevant?	Yes	No				
2	Are any relevant interventions missing?	No	Yes				
3	Are the outcomes relevant?	No	Yes				
4	Is the context (settings and practice patterns) applicable?	Yes	No				
Comments							
Credibility questions							
Design							
1	Were the study hypotheses or goals prespecified a priori?	Yes	No				
2	If one or more comparison groups were used, were they concurrent comparators or did they justify the use of historical comparison group(s)?	Yes	No				
3	Was there evidence that a formal study protocol including an analysis plan was specified before executing the study?	Yes	No				
4	Were sample size and statistical power to detect differences addressed?	Yes	No				
5	Was a study design used to minimize or account for confounding?	Yes	No				
6	Was the follow-up period of sufficient duration to detect differences addressed?	Yes	No				
7	Were the sources, criteria, and methods for selecting participants appropriate to address the study questions/hypotheses?	Yes	No				
8	Were the study groups selected so that comparison groups would be sufficiently similar to each other (e.g., either by restriction or recruitment based on the same indications for treatment)?	Yes	No				
Comments							
Data							
1	Were the data sources sufficient to support the study?	Yes	No				
2	Was exposure defined and measured in a valid way?	Yes	No				

Table 1 – *continued*

S. no.	Question	Strength	Weakness	Can't answer			
				Not applicable	Not reported	Not enough information	Not enough training
3	Were the primary outcomes defined and measured in a valid way?	Yes	No				
4	Was the follow-up time similar among comparison groups or were the differences in follow-up accounted for in the analyses?	Yes	No				
Comments							
Analyses							
1	Was there a thorough assessment of potential measured and unmeasured confounders?	Yes	No				
2	Were analyses of subgroups or interaction effects reported for comparison groups?	Yes	No				
3	Were sensitivity analyses performed to assess the effect of key assumptions or definitions on outcomes?	Yes	No				
Comments							
Reporting							
1	Was the number of individuals screened or selected at each stage of defining the final sample reported?	Yes	No				
2	Were the descriptive statistics of the study participants adequately reported?	Yes	No				
3	Did the authors describe the key components of their statistical approaches?	Yes	No				
4	Were confounder-adjusted estimates of treatment effects reported?	Yes	No				
5	Did the authors describe the statistical uncertainty of their findings?	Yes	No				
6	Was the extent of missing data reported?						
7	Were absolute and relative measures of treatment effect reported?	Yes	No				
Comments							
Interpretation							
1	Were the results consistent with prior known information or if not was	Yes	No				

Table 1 – *continued*

S. no.	Question	Strength	Weakness	Can't answer			
				Not applicable	Not reported	Not enough information	Not enough training
	an adequate explanation provided?						
2	Are the observed treatment effects considered clinically meaningful?	Yes	No				
3	Are the conclusions supported by the data and analysis presented?	Yes	No				
4	Was the effect of unmeasured confounding discussed?	Yes	No				
Comments							
Conflicts of interest							
1	Were there any potential conflicts of interest?	Yes	No				
2	If there were potential conflicts of interest, were steps taken to address these?	Yes	No				
Comments							

Relevance questions relate to the usefulness of the observational study to inform health care decision making. Each question will be scored with *Yes/No/Can't Answer*. Based on the scoring of the individual questions, the overall relevance of the observational study needs to be judged as *Sufficient or Insufficient*.

If the observational study is considered sufficiently relevant, the credibility is going to be assessed. The **credibility** is captured with questions in the following six domains: *Design, Data, Analysis, Reporting, Interpretation,* and *Conflict of interest*. Each question will be scored with *Yes/No/Can't Answer*. Based on the number of questions scored satisfactory in each domain, an overall judgment of the strength of each domain needs to be provided: *Strength/Neutral/Weakness/Fatal flaw*. If any one of the items is scored as a *no* resulting in a fatal flaw, the overall domain will be scored as a fatal flaw and the study may have serious validity issues. Based on the domain judgments, the overall credibility of the study will be judged as *Sufficient or Insufficient*.

* The questionnaire consists of 33 questions related to the **relevance** and **credibility** of a prospective observational study.

misleading. Consequently, the decision maker must use extreme caution in applying the findings to inform decisions. The occurrence of a fatal flaw, however, does not prevent a user from completing the questionnaire nor does it require the user to judge the evidence as insufficient for use in decision making. The presence of a fatal flaw is intended to raise a strong caution and should be carefully considered when the overall body of evidence is reviewed.

Questionnaire Items

Relevance

Relevance addresses whether the results of the study/apply to the setting of interest to the decision maker. It addresses issues of external validity similar to the population, interventions, comparators, outcomes, and setting framework from evidence-based medicine [36]. There is no correct answer for relevance. Relevance is determined by each decision maker, and the relevance assessment determined by one decision maker will not necessarily apply to other decision makers. Individual studies may be designed with the perspective of particular decision makers in mind (e.g., payer or provider).

Is the population relevant?

This question addresses whether the population analyzed in the study sufficiently matches the population of interest to the decision maker. Population characteristics to consider include demographic characteristics such as age and sex, nationality, and ethnicity; risk factors such as average blood pressure, cholesterol levels, and body mass index; behaviors such as smoking; disease history and onset, stage, and severity of the condition; past and current treatments for the condition; and clinical issues such as comorbidities. For rare diseases or special populations such as pediatrics, the decision to conduct a prospective observational study may be prompted by the general paucity of available information; recruitment will necessarily be related to patient access.

Are any relevant interventions missing?

This question addresses whether the interventions analyzed in the study include ones of interest to the decision maker and whether all relevant comparators have been considered. It is frequently unrealistic that ALL interventions that could be considered to treat a disease or condition be included in the analysis; however, omitting key interventions that represent standards of care introduces the potential for bias and uncertainty. Intervention characteristics should also be specified and defined at a detailed level. For technologies, this includes the device

specification and the technique used (e.g., screening for osteoporosis: whether dual-energy X-ray absorptiometry or some other scanning method was used and whether spine, hip, or wrist measurements were taken).

- *Questions to consider:*
  - *For drugs and biologics, were the doses, durations, and modes of administration specified?*
  - *If usual clinical care was a comparator, were they adequately described to determine whether they resemble the modes of care in the decision setting?*
  - *For surgical interventions, was the skill level of providers, posttreatment monitoring and care, and duration of follow-up reported?*

Are the outcomes relevant?

This question asks what outcomes are assessed in the study and whether the outcomes are meaningful to the patients the decision maker is concerned with. Outcomes such as cardiovascular events (e.g., rates of myocardial infarction or stroke), mortality, patient functioning, health-related quality-of-life or health status measures (e.g., scores from the short-from 36 health survey or the EuroQol five-dimensional questionnaire) may be more relevant to a decision maker than surrogate or intermediate end points (e.g., cholesterol levels).

Is the context (settings and practice patterns) applicable?

The context of the study refers to factors that may affect the generalizability of the study findings to other settings. Factors that should be considered may include the study time frame, the payer or health system setting, provider characteristics, or the geographic area. Some or all of these factors may be different than the population to which the user wants to apply the study results; however, if it is suspected that differences in these factors may affect the treatment response, it should affect the user's judgment of the extent that these findings could be applied to another setting.

Credibility

Credibility addresses the extent to which the study accurately answers the question it is designed or intended to answer and is determined by the design and conduct of the study. Central to credibility assessment in the comparative effectiveness framework is assessing the validity of the causal inferences of the results. It focuses on issues of internal validity, measurement error, and confounding. For example, the observed effect of a new treatment may be due to the manner in which patients were selected for treatment, or the degree to which patients were followed and their outcomes reliably measured, and not due to differences in treatment effectiveness. Appropriate study design and analytic approaches can better separate the contribution of the intervention to observed outcomes versus other factors.

There are a wide range of resources available to assist users in familiarizing themselves with some of the core concepts to assess study quality including many textbooks in fields such as econometrics, statistics, clinical research, epidemiology, and health services research. Some contemporary resources that are freely available that are focused on comparative effectiveness research in the observational framework include Agency for Healthcare Research and Quality's "Developing a Protocol for Observational Comparative Effectiveness Research: A User's Guide"; Patient-Centered Outcomes Research Institute's Methodological Standards; and the STrengthening the Reporting of OBservational studies in Epidemiology initiative, which focuses on study reporting [16,18,37].

Design

Were the study hypotheses or goals prespecified a priori? As stated in a previous report of an ISPOR Task Force, "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 regarding a priori specification of a research hypothesis when conducting observational research, prior specification minimizes the risk of outcome selection bias or "cherry-picking" interesting findings and a related issue of observing spurious findings because of multiple hypothesis testing [10]. For these reasons, we recommend the practice of a priori specification of the research question, study design, and data-analysis plan in a formal study protocol to assure end users that the results were not the product of data exploration (i.e., "fishing" or "dredging" data: searching through the data until you find something interesting) [38]. This is not an indictment of exploratory data analysis; rather that data exploration is more appropriate for hypothesis generation, rather than hypothesis testing. Evidence that prespecified hypotheses were formally stated in a protocol includes registration on a publicly available Web site, such as clinicaltrials.gov, or evidence of a review procedure or process that may include disclosure of an institutional review board (IRB) procedure, or using terms such as "prespecified," "a priori," or "planned analyses" in the report. We note that most prospective observational protocols will require an ethics or IRB review, especially if they entail collection of specific assessments not part of routine medical care. We also recognize that exploratory analyses may be more likely in retrospective observational studies that use existing records.

- *Questions to consider:*
  - *Was the study registered on a publicly available Web site?*
  - *Was IRB review or approval obtained—suggesting that the researchers used a formal protocol?*
  - *Was the study performed as a result of a research grant from a well-established institution—suggesting that researchers had a formal study proposal and analysis plan?*

If one or more comparison groups were used, were they concurrent comparators or did they justify the use of historical comparison group(s)? Concurrent comparators are derived from the same population of subjects and are followed over the same time frame; this approach avoids time-related confounding. Alternatively, historical controls are derived from a population of subjects derived from a time period prior to the treatment that is compared. Concurrent comparators add more strength to research findings, although the use of historical comparators can be justified when a treatment becomes a standard of care and nearly all subjects who could be treated either receive the new treatment, or there are perceived ethical barriers to withhold the new treatment. This suggests that a "no" answer to this question does not automatically invalidate the credibility of study findings. This issue is more commonly encountered in retrospective observational studies and the choice of historical comparison groups should be justified.

Was there evidence that a formal study protocol including an analysis plan was specified before executing the study? Altering study definitions, subject selection criteria, model specifications, and other study procedures can have dramatic effects on study findings and can introduce investigator biases [16–18]. Ideally, study reports would be fully transparent about planned study procedures. They should report findings based on the original plan and be fully transparent regarding post hoc changes to the analysis plan to justify those post hoc alterations. Unfortunately, conventional reporting practices for observational studies are

rarely detailed enough to allow readers to adequately assess what was preplanned and what was not [20]. The best evidence that the study had an analytic plan would be to compare a study report with its registry data; however, only a small fraction of observational studies are registered. Trial registries provide the best documentation that the study procedures followed a predetermined analytical plan. Few observational studies, however, use trial registries, limiting the availability of outside confirmation of an analytical plan for observational studies. Alternatively, users could rely on terms commonly used, such as prespecified or "planned analyses" when describing the methods. The results can also indicate prespecified analyses by declaring results that were preplanned versus those that were post hoc. If the study reports post hoc analyses, the user may cautiously infer that other analyses were preplanned. In addition, when a study reports peer-reviewed grant funding or an IRB review, it suggests that an analysis plan was developed a priori. A prespecified analysis plan cannot be assumed when a study does not offer any indication that there was such a plan developed beforehand.

Were sample size and statistical power to detect difference addressed? An observational study attempts to create a comparison across two samples just as its randomized counterpart and therefore still requires a sample size or power calculation if results are applied to a different study population [10]. Without this, the reader is left with insufficient information as to whether the detectable difference should have occurred on the basis of the expected size of the effect and in advance of the study.

In retrospective studies in which subjects are not prospectively recruited, sample size estimates do not dictate the number of subjects to be included in the study because the investigator will typically include all available subjects recorded in the data source and follow them for as long as subject data are recorded. Including a sample size estimate or power calculation, however, enables readers to interpret findings, particularly null findings.

Was a study design used to minimize or account for confounding? A *confounder* is a factor that distorts the true relationship of the study variables of central interest by virtue of being related to the outcome of interest, but not related to the study question and unequally distributed among the groups being compared. Some study designs can provide stronger methods to deal with potential confounding that may occur because of lack of randomization. These include inception cohorts, new user designs, the use of multiple comparator groups, matching designs, and assessment of outcomes not thought to be affected by the intervention compared (Table 2). Moreover, various considerations factor into the choice of study design including feasibility, cost, and ethical considerations.

The risk of residual, or unobserved, confounding is greater the more dissimilar are the comparison groups. Design and analytic approaches to ensure comparability of treatment and control groups include matching designs, and analytic approaches such as propensity scoring, and instrumental variable techniques. These are described in more detail in previous ISPOR Task Force reports [10–13].

- *Questions to consider:*
  - *Did the study design use strategies to minimize confounding such as inception cohorts, new-user design, or matching?*
  - *Did the study analysis ensure comparability of comparison groups through methods such as propensity scoring?*

Was the follow-up period of sufficient duration to detect differences addressed? Depending on the condition studied and the difference in effect of comparator interventions, the length of time required to detect differences will vary. The more quickly and more frequently outcome events are expected to occur, the shorter the duration of follow-up is required (e.g., asthma exacerbations can be detected with shorter observation periods in comparison with hip fracture). The duration of follow-up is related to the power of the study and its ability to detect differences. There may be feasibility and cost limitations that affect the duration of follow-up in prospective observational studies; however, this should not affect the assessment of the credibility of its findings. It is important to understand that in circumstances in which the follow-up time used to assess outcomes overlaps with the time to assess exposure (e.g., determine survival after someone is diagnosed or is discharged from a hospital and then use that same follow-up period to determine whether he or she receives a prescription) or when imposing a minimum follow-up time as part of the study design, other biases can be introduced (e.g., immortal time bias) in the estimates of treatment effect [10–13,39].

- *Questions to consider:*
  - *Was the duration of follow-up appropriate to the population and condition being studied?*
  - *Could the follow-up time have been influenced by the effects of treatment (e.g., anticoagulation and stroke risk)?*
  - *Was the time frame to assess exposure before the beginning of the follow-up time to record outcomes?*

Were the sources, criteria, and methods for selecting participants appropriate to address the study questions/hypotheses? The sources, criteria, and methods for selecting participants for a study should be similar for the different groups of patients being assessed. Bias (e.g., a consistent measurable effect from systematic rather than random error and not from the intervention) can be introduced if the comparator group or the data source or methods for assessing or selecting patient groups vary [40]. Also, the data source should provide some level of assurance that key measures are reliably recorded. For retrospective studies, assurance that relevant health information was captured will vary. Administrative data sources should be checked to ensure that subjects had continuous eligibility for health benefits and were eligible to receive all relevant sources of care. Not all persons eligible for insurance in the United States will be eligible for pharmacy or mental benefits, for example. Studies that rely on provider-supplied data such as electronic medical records (EMRs)

Table 2 – Study designs used to minimize the effect of confounding variables.

Inception cohorts are designated groups of persons assembled at a common time early in the development of a specific clinical disorder (e.g., at first exposure to the putative cause or at initial diagnosis), who are followed thereafter.
New-user design begins by identifying all the patients in a defined population (in terms of both people and time) who start a course of treatment with the study medication. Study follow-up for end points begins at precisely the same time as initiation of therapy (t0). The study is further restricted to patients with a minimum period of nonuse (washout) before t0.
Matching designs include a deliberate process to make two comparable groups, a study group and a comparison group, matched on factors extraneous to the main purpose for the investigation but which might interfere with the interpretation of the study's findings.
Assessment of outcomes thought not to be affected by the interventions compared may permit an assessment of residual confounding [10]. These may be described as falsification tests.

should attempt to ensure that persons are receiving care in the provider network and not from other providers.

- *Questions to consider:*
  - *Was there an adequate rationale provided for key inclusion and exclusion criteria?*
  - *Was there an assurance that subject encounters or data were adequately recorded over the entire study time frame for each subject?*

Were the study groups selected so that comparison groups would be sufficiently similar to each other (e.g., either by restriction or recruitment based on the same indications for treatment)? One of the most common starting points to enhance the comparability of treatment groups is to restrict subjects to a common set of conditions or patient characteristics. For example, if one were to compare beta blockers and diuretics as antihypertensive therapy, it would be important to restrict both treated groups to those without any evidence of previous cardiovascular disease including angina because beta blockers have a Food and Drug Administration indication for angina (a strong risk factor for subsequent myocardial infarction) whereas diuretics do not. A study that does not include a strategy to ensure similarity between groups could suffer from confounding by indication.

- *Question to consider:*
  - *Did the study design ensure comparability of comparison groups through methods such as restriction?*

Data

Were the data sources sufficient to support the study? The data sources should contain valid measures of treatment, outcome, and covariates including confounding variables, possess a large enough sample, and be of sufficient duration to detect differences. Often a single data source will not contain all the information necessary to conduct the study, and linkages to other data sources may be necessary [22]. Prospective studies may design study-specific data collection forms; they may also rely on existing mechanisms to collect data. The quality of the study data is a broad concept, and it can be affected by multiple factors.

- *Questions to consider:*
  - *Were all outcomes, exposures, predictors, potential confounders, and effect modifiers clearly defined?*
  - *Were the sources of data and details of methods of assessment for each variable of interest adequately described?*
  - *Were the assessment methods and/or definitions the same across treatment groups?*
  - *Were the reliability and validity of the data described, including any data quality checks and data cleaning procedures?*
  - *Have the data sources been used previously for research?*
  - *Were reliable and valid data available on important confounders or effect modifiers?*

Was exposure defined and measured in a valid way? Exposure to treatment is ideally documented by evidence that patients actually took the medication or received the treatment, though this is rarely done [23]. Exposure may be documented by evidence of a prescription being written, being filled, a claim being filed, and measures of medication possession. Exposure is typically defined by whether a subject received (or did not receive) a treatment; however, exposure may also be defined in terms of the intensity of exposure, including the dose and/or duration of treatment(s).

- *Questions to consider:*
  - *Does the study describe the data source(s) used in the study of the ascertainment of exposure (e.g., pharmacy dispensing, general practice prescribing, claims data, EMR data, chart review, self-report, face-to-face interview)?*
  - *Does the study describe how exposure is defined and measured (e.g., operational details for defining and categorizing exposure)?*
  - *Does the study discuss the validity of exposure measurement (e.g., precision, accuracy, prospective ascertainment, and exposure information recorded before outcome occurred)?*

Were the primary outcomes defined and measured in a valid way? Selection of primary outcomes is perhaps the most critical part of study design [10,17]. Outcomes more directly observable, such as laboratory measures, hospitalization, and death, may require less sophisticated validation approaches than do outcomes that are more difficult to observe or rely on investigator classification and subjectivity, such as time to referral or medication adherence. Primary outcomes can be documented in a patient chart (paper, EMR), inferred from claims (e.g., through administrative codes suggesting hospitalization for myocardial infarction), documented in a diary, or reported in a face-to-face interview. The validity of outcome measures will vary on the basis of the outcome measure, the context of the study, and the data sources used and ultimately some degree of judgement will be required.

In the retrospective framework, the validity of the outcome measure may require more careful consideration, particularly when the data source was not designed for the specific research purpose, which is commonly encountered in studies that use medical charts, EMRs, or administrative claims. Ideally, some evidence of the validity (sensitivity, specificity, positive and negative predictive values) of the outcome measure contrasted with a measure where direct observation of the outcome could be verified is reported. This can be accomplished by conducting a substudy to verify the accuracy of the outcome definitions using a more accurate or detailed data source (e.g., direct report by subject or provider). Weaker evidence of the validity can be inferred when previous studies that report the validity of the outcome measures and definitions using a similar but different data source (e.g., validity of outcome definition obtained from one administrative data source that is structured similarly to the study's data source) are cited. Outcome measure definitions that have been used in previous analyses permit comparison between studies but do not ensure validity.

Was the follow-up time similar among comparison groups or were the differences in follow-up accounted for in the analyses? Patients may drop out of a study or discontinue medications for many reasons including lack of effectiveness, adverse effects, or routine life events such as moving or changing insurers. Differential follow-up between treatment groups can introduce a bias in observed treatment effects. A credible study will explain as best as possible the reasons for these differences and use appropriate statistical techniques (reporting rates and utilizing time to event analytic approaches that account for censoring) to minimize the effect of variable follow-up time. Even when these analytic approaches are used, immortal time bias cannot be ruled out and is more likely when the duration of follow-up time is affected by the treatment group selections or if being selected into a treatment group is dependent on person time.

Analyses

Was there a thorough assessment of potential measured and unmeasured confounders? The choice and effectiveness of

treatments may be affected by practice setting, the health care environment and the experience of health care providers, as well as the medical history of patients. Treatment inferences from observational studies are all potentially biased from imbalances across treatment groups on confounding variables whether the variables are observed or not [10,17]. Confounding can be controlled statistically using a wide arrange of multivariate approaches. If a statistical model excludes a confounding variable, however, the estimates of treatment effects suffer from omitted variable bias in just about all analyses except when a technique such as instrumental variables or a regression discontinuation approach that uses a "natural" randomizer is undertaken [26,41]. Ideally, one should look to see that the authors have considered all potential confounding factors and conducted a literature review to identify variables that are known to affect the outcome variable. Unfortunately, this is not typically reported in articles [13]. A table of potential confounders with citations to previous research describing these associations is sometimes available in an article and would be suggestive of an explicit search to identify known confounding variables. In addition to a literature search, credible research should use clinical judgment or consensus techniques to identify confounders. Often, the data will not contain information for some confounding variables (e.g., race, income level, and exercise level) and these variables will be omitted from the analysis [26]. When the analysis does not include key confounders, a thorough research report will discuss the potential effect of these, including the direction and magnitude of the potential bias.

- *Questions to consider:*
  - *Was there evidence that a literature search was performed to identify all potential measured and unmeasured confounders?*
  - *Were known influential confounders unrecorded or not included in the adjusted analyses?*

Were analyses of subgroups or interaction effects reported for comparison groups? Exploring and identifying heterogeneous treatment effects, or effect modification, are some of the potential advantages of large observational studies. Interaction occurs when the association of one exposure differs in the presence of another exposure or factor. The most common and basic approaches for identifying heterogeneous treatment effects are to conduct subgroup analyses or to incorporate interaction terms within the analysis [10,31]. Interactions may be explored using additive or multiplicative approaches in which the differences in effect depart from either the addition of effects of two factors or exposures, or the multiplicative effect of those factors. Caution is warranted when the main treatment effect is not significantly associated with the outcome, but significant subgroup results are reported (which could have been the result of a series of post hoc subgroup analyses) [42].

Were sensitivity analyses performed to assess the effect of key assumptions or definitions of exposure or outcome measures? Various decisions must be made in designing a study. This includes how populations, interventions, and outcomes are defined; how missing data are dealt with; how outliers are dealt with; which analytic approaches were taken; and to what extent unmeasured confounders may affect the results. A credible study will indicate which of these decisions had an important effect on the results of the analyses and will report the effect of using a reasonable range of alternative choices on the results. Key issues to consider include whether sensitivity analyses were reported using different statistical approaches, and according to key definitions; whether the analysis accounted for outliers and examined their effect in a sensitivity analysis; and whether the authors discussed or conducted a sensitivity analysis to estimate the effect of unmeasured confounders on the observed difference in effect.

- *Questions to consider:*
  - *Were sensitivity analyses reported using different statistical approaches?*
  - *Were sensitivity analyses reported to test the effect of key definitions?*
  - *Did the analysis account for outliers and examine their effect in sensitivity analysis?*
  - *Did the authors discuss or conduct a sensitivity analysis to estimate the effect of unmeasured confounders on the observed difference in effect?*

Reporting

When methods are described in sufficient detail, it permits others to replicate the analysis on similar data sets or to reproduce the results if given access to the data set from the study. An adequate description delineates all key assumptions, describes the key study measure, and why a methodological approach was selected over alternatives. Increasingly, published reports rely on online appendices to more fully describe methods and results as these should be checked in addition to any errata and letters to the editor to identify corrections or alternative viewpoints in interpreting the data. Although there are several checklists that have been developed to address adequacy of reporting including minimum reporting standards for medical journals, a straightforward consideration is whether the reader can understand precisely how study authors arrived at their particular findings [16].

Was the number of individuals screened or selected at each stage of defining the final sample reported? Reporting the number of individuals screened at each stage of the selection process is important to assess the potential selection bias of participants and can provide cursory evidence that the study procedures were implemented correctly. The final analyzable sample can most easily be interpreted when a text description or a flow diagram is provided that describes the initial pool of potential subjects and the sample after each inclusion and exclusion is applied. This allows the reader to more easily assess the extent to which the analyzable sample may differ from the target population and which criteria materially affected the final sample.

Were the descriptive statistics of the study participants adequately reported? A basic summary of the observable characteristics of the study population should be provided including descriptive statistics on the mean value (and distribution where appropriate) for demographic variables, prevalence of comorbidities, and other potential confounders reported by treatment groups to enable the reader to assess the comparability of treatment groups and the potential for confounding. Vast differences on key confounding measures may suggest a higher likelihood of residual confounding even after adjusting for the observable characteristics.

Did the authors describe and report the key components of their statistical approaches? The authors should fully describe their statistical approach and provide citations for any specific variation of econometric or statistical methods that they used. One question to consider in assessing the adequacy of the statistical reporting include whether the authors used statistical techniques to examine the effect of multiple variables simultaneously (i.e., multivariate analysis) and whether they discussed how well the

models predicted what they are intended to predict. Authors may conduct a statistical test, such as r-squared, pseudo r-squared, c-statistics, and c-indices, to demonstrate the predictive capacity of the statistical model used. Other key items of statistical reporting relate to statistical techniques used to adjust for multiple analyses of the same data, reporting of unadjusted estimates of treatment effects, and reporting of the full regression model in either the publication or the appendix. Techniques commonly used in observational studies include propensity score methods and instrument variable methods. Some of these techniques may require more extensive reporting of their development, use, and evaluation. A guiding principle for statistical reporting is to "Describe statistical methods with enough detail to enable a knowledgeable reader with access to the original data to verify the reported results" [43].

- *Questions to consider:*
  - *Did the authors describe and report the key components of their statistical approaches?*
  - *Were any modeling assumptions addressed?*
  - *If the authors use multivariate statistical techniques, do they discuss how well the models predict what they are intended to predict?*
  - *Were unadjusted estimates of outcomes reported?*
  - *Was the full regression model (not just the adjusted treatment effects) available in the publication or at least an appendix?*
  - *If propensity score methods were used, were the methods appropriately described? Method of developing the propensity score, use of the propensity score (matching, weighting, regression, etc.), evaluation of the propensity score (e.g., standardized differences before and after matching).*
  - *If instrumental variable methods were used, were the methods appropriately described (rational for the instrumental variable, evaluation of the strength of the instrument).*

Were confounder-adjusted estimates of treatment effects reported? Confounder-adjusted estimates of treatment effects can be obtained in various ways. Most commonly, treatment effects are estimated from a coefficient of independent variable(s) in a multivariate regression, or systems of equations that include a set of control covariates that represent potential confounders. Treatment effects may also be estimated by taking differences from propensity-matched treated and comparison subjects. Any nonrandomized study must report confounder-adjusted estimates if it is attempting to make any inference regarding the effects from treatment. Unadjusted estimates should also be reported to allow for comparison with the adjusted results.

Did the authors describe the statistical uncertainty of their findings? There will always be uncertainty when evaluating outcomes in observational research because estimates must be based on population samples. Uncertainty from sampling should be presented in the form of either a Bayesian credibility interval (range of values that is likely to contain the true size of the effect given the data) or a CI (range of values that is likely to contain the true estimate, e.g., 95%). *P* values can provide some sense of uncertainty but are not sufficient for reanalysis. Because they are a product of both uncertainty and the magnitude of the effect observed, they can be misleading when either sample populations or effect sizes are large.

Was the extent of missing data reported? There is considerable opportunity to introduce bias into estimates of treatment effects if data are missing [28,44]. Missing data can occur in prospective observational studies because many studies rely on secondary data sources that rely on routine data entry. Because it is possible that the reason for the missing data is related to the reason for observed treatment effects, the extent of missing data should be reported. The potential for bias from missing data can be further explored in sensitivity analyses or through analyses that attempt to correct for missing data by making assumptions. Credibility i s enhanced if the authors indicate that an a priori analytic strategy to deal with missing data (including any data imputation method) was explicitly created as part of their data analysis plan.

Were absolute and relative measures of treatment effect reported? Reporting the effect of treatment(s) in both absolute and relative terms provides the decision maker the greatest understanding of the magnitude of the effect [45]. Absolute measures of effect included differences in proportions, means, rates, number needed to harm, and number needed to treat and should be reported for a meaningful time period. Relative measures of effect are rate ratios, proportions, or other measures and include odds ratios, incidence rate ratios, relative risks, and hazard ratios.

Interpretation

Were the results consistent with previous known information or if not was an adequate explanation provided? To aid interpretation of research, study authors should undertake a thorough review of the literature to compare their findings to all known previous findings exploring the same or similar objectives. Research authors should provide plausible explanations for disparate findings and identify methodological differences or advance a theoretical or biologic rationale for the differences. Authors should provide plausible explanations that have led to findings that are different in direction or magnitude.

Are the observed treatment effects considered clinically meaningful? In analyses of large observational studies, sometimes relatively minor differences between treatment groups can attain levels of statistical significance because of the large sample sizes. The results should be interpreted not only in terms of their statistical association but also by the magnitude of effect in terms of clinical importance. Some authors may identify previously developed minimally important clinical differences to support their assertions. In addition, the larger the treatment effect that is observed, the smaller the chances that residual confounding can change a significant finding to a null finding.

Are the conclusions supported by the data and analysis presented? Overstating the implications of the study results is commonly encountered in the literature [46]. The study should be fully transparent describing the study limitations and importantly how study limitations could affect the direction and magnitude of the findings and ultimately the study conclusions. Users of a study should consider whether conclusions are cautious and appropriate given the objectives, limitations, multiplicity of analyses, results from similar studies, and other relevant evidence. Authors should discuss the limitations of the study, including the potential direction and magnitude of any potential bias, to help users of the study understand the degree to which these limitations may reduce the strength of the casual inferences.

Was the effect of unmeasured confounding discussed? Unmeasured confounding is always a potential issue in any observational research framework. Unmeasured confounding is more likely to bias the results when patients might be channeled to one treatment over another with studies that investigate known or likely

outcomes. Outcome events that are lesser known or unsuspected are less likely to be considered in the clinician's or patient's decision process, and, in turn, are less likely to suffer from confounding. A good discussion will identify factors thought to be confounders that were not recorded or were unmeasurable, and identifies the potential direction of the bias. A credible study would ideally assess residual confounding through simulations to explore how strongly a confounder would have to be correlated with treatment and outcome to move the results to a null finding.

Conflicts of interest

Two questions were used in the conflicts of interest domain in all the questionnaires.

Were there any potential conflicts of interest? Conflicts of interest may be stated by authors; however, reviewers may also seek information from public sources including Web-based curricula vitae or faculty pages. In some cases, conflicts are not stated in a research report simply because of editorial policy rather than a lack of their existence. Although some journals adhere strictly to uniform standards for disclosing conflicts, such as those promoted by the International Committee of Medical Journal Editors, others may not. Readers should not misinterpret absence of a stated conflict as evidence of absence of a conflict.

If there were potential conflicts of interest, were steps taken to address these? Potential conflicts of interest may include financial interests in the study results, desire for professional recognition, or other nonmonetary incentives. Steps to address potential conflicts of interest include disclosure of any potential conflicts of interest; involving third parties in the design, conduct, and analysis of studies; and agreements that provide independence of researchers (including freedom to publicly disseminate results) from funding entities [30].

Summary of Questionnaire Results

The Web-based version of the questionnaire provides a summary of the assessments as follows:

In evaluating this study, I made the following judgments:

- *I found the study (relevant, not relevant) for decision making because I considered that the population/interventions/outcomes/setting (applied, did not apply) to the decision I am informing.*
- *I found the study (credible, not credible) for decision making because:*
  - *There (were, were not any) fatal flaws—that is, critical elements that call into question the validity of the findings.*
    - *The presence of a fatal flaw suggests significant opportunities for the findings to be misleading and misinterpreted; extreme caution should be used in applying the findings to inform decisions.*
    - *The following domains contained fatal flaws:*
  - *There are strengths and weakness in the study:*
    - *The following domains were evaluated as strengths:*
    - *The following domains were evaluated as weaknesses:*

Discussion

User Testing

Following preliminary testing by members of the four task forces, the wording of some of the questions was modified. The revised questionnaire was made available to volunteers from the payer community (members of the AMCP) as well as the pharmaceutical industry. Ninety-three volunteers were solicited to participate. Each volunteer was asked to test one questionnaire using three studies and rate them accordingly. Studies were provided that were previously rated by the Task Force as either "good quality," "medium quality," or "poor quality." Sixty-five volunteers participated, of which 25 were assigned to the prospective observational study questionnaire; the response rate from this group was 72%. Twenty were assigned to the retrospective observational study questionnaire; the response rate from this group was 70%. Although there were not enough users to perform a formal psychometric evaluation, the good-quality studies were generally rated as sufficient with respect to credibility, while the poor-quality studies were generally rated not sufficient. Based on the answer to the question: "Is the study 'Sufficiently' or 'Insufficiently' credible to include in the body of evidence?" there was 59% and 54% multirater agreement among ratings provided for the prospective and retrospective questionnaires, respectively. Ratings were not provided 15% to 36% of the time. Multirater agreement exceeded 80% for 16 of the 28 original items in the retrospective credibility domains. In the original format, there were supplementary questions and few test users completed those and hence these were not included as specific items in this revised questionnaire. These results were used to modify the questionnaire to its current format and enhanced explanations were developed to increase interrater agreement.

Educational Needs

Internationally, the resources and expertise available to inform health care decision makers vary widely. Although there is broad experience in evaluating evidence from RCTs, there is less experience and greater skepticism regarding the value of other types of evidence [12,18]. The volume and variety of real-world evidence, however, is increasing rapidly with the ongoing adoption of electronic EMRs along with the linkage of claims data with laboratory, imaging, and EMR data. Volume, variety, and velocity (the speed at which data are generated) are three of the hallmarks of the era of "big data" in health care. The amount of information from these sources could easily eclipse that from RCTs in coming years. This implies that it is an ideal time for health care decision makers and those who support evidence evaluation to enhance their ability to evaluate this information.

Although there is skepticism about the value of evidence from observational, network meta-analysis/indirect treatment comparison, and modeling studies, they continue to fill important gaps in knowledge for payers, providers, and patients. ISPOR has provided Good Research Practice recommendations on what rigorous design, conduct, and analysis looks like for these sources of evidence [10–13,47,48]. These questionnaires, including the one discussed in this report, are an extension of those recommendations and serve as a platform to assist the decision maker in understanding what a comprehensive evaluation of this research requires. By using this questionnaire, our intent is to make observational evidence more accessible and raise the level of sophistication by decision makers and the bodies that support them through the use and interpretation of evidence.

To that end, we anticipate additional educational efforts and promotion of these questionnaires and that they will be developed and made available to an increasing number of health care decision makers. In addition, an interactive Web-based tool has been developed at https://healthstudyassessment.org/ to facilitate uptake and support the educational goal of the questionnaire.

Acknowledgments

We express our gratitude for the tireless efforts of Rebecca Corey and Randa Eldessouki to organize our committee meetings, coordinate and collate user feedback, and provide editorial support and prodding us at just the right time to keep this effort moving forward. We also thank Anand Shewale who assisted with the analysis of user feedback and editing. We are also thankful to the volunteers who provided helpful comments on previous versions of this work.

Source of financial support: A portion of Dr. Martin's effort was supported by the UAMS Translational Research Institute (grant no. 1UL1RR029884).

REFERENCES

[1] Jansen J, Trikalinos TA, Cappelleri JP, et al., Indirect treatment comparison/network meta-analysis study questionnaire to assess study relevance and credibility to inform healthcare decision-making: an ISPOR-AMCP-NPC Good Practice Task Force report. Value Health 2014;17:157–73.
[2] Caro JJ, Eddy DM, Kan H, et al., A modeling study questionnaire to assess study relevance and credibility to inform health care decision making: an ISPOR-AMCP-NPC Good Practice Task Force report. Value Health 2014;17:174–82.
[3] Garrison LP Jr, Neumann PJ, Erickson P, et al., Using real-world data for coverage and payment decisions: the ISPOR Real-World Data Task Force report. Value Health 2007;10:326–35.
[4] Brixner DI, Holtorf AP, Neumann PJ, et al., Standardizing quality assessment of observational studies for decision making in health care. J Manag Care Pharm 2009;15(3):275–83.
[5] Balshem H, Helfand M, Schunemann HJ, et al., GRADE guidelines, 3: rating the quality of evidence. J Clin Epidemiol 2011;64:401–6.
[6] Atkins D, Eccles M, Flottorp S, et al., Systems for grading the quality of evidence and the strength of recommendations, I: critical appraisal of existing approaches. The GRADE Working Group. BMC Health Serv Res 2004;4(1):38.
[7] Glasziou P, Vandenbroucke JP, Chalmers I. Assessing the quality of research. BMJ 2004;328:39–41.
[8] Guyatt GH, Oxman AD, Vist GE, et al., GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ 2008;336:924–6.
[9] Moher D, Jadad AR, Tugwell P. Assessing the quality of randomized controlled trials: current issues and future directions. Int J Technol Assess Health Care 1996;12:195–208.
[10] Berger ML, Dreyer N, Anderson F, et al., Prospective observational studies to assess comparative effectiveness: the ISPOR Good Research Practices Task Force report. Value Health 2012;15:217–30.
[11] Berger ML, Mamdani M, Atkins D, Johnson ML. Good research practices for comparative effectiveness research: defining, reporting and interpreting nonrandomized studies of treatment effects using secondary data sources: the ISPOR Good Research Practices for Retrospective Database Analysis Task Force Report—part I. Value Health 2009;12:1044–52.
[12] Cox E, Martin BC, Van Staa T, et al., Good research practices for comparative effectiveness research: approaches to mitigate bias and confounding in the design of nonrandomized studies of treatment effects using secondary data sources: the International Society for Pharmacoeconomics and Outcomes Research Good Research Practices for Retrospective Database Analysis Task Force Report–part II. Value Health 2009;12:1053–61.
[13] Johnson ML, Crown W, Martin BC, et al., Good research practices for comparative effectiveness research: analytic methods to improve causal inference from nonrandomized studies of treatment effects using secondary data sources: the ISPOR Good Research Practices for Retrospective Database Analysis Task Force Report—part III. Value Health 2009;12:1062–73.
[14] Juni P, Witschi A, Bloch R, Egger M. The hazards of scoring the quality of clinical trials for meta-analysis. JAMA 1999;282:1054–60.
[15] GRACE Initiative. GRACE principles. Available from: http://www.graceprinciples.org. [Accessed December 4, 2013].
[16] STROBE checklist. 2013. Available from: http://www.strobe-statement.org/fileadmin/Strobe/uploads/checklists/STROBE_checklist_v4_cohort.pdfstatement.org/fileadmin/Strobe/uploads/checklists/STROBE_checklist_v4_cohort.pdf. [Accessed December 4, 2013].
[17] The European Network of Centres for Pharmacoepidemiology and Pharmacovigilance. Guide on methodological standards in pharmacoepidemiology (Revision 1). 2012. Available from: http://www.encepp.eu/standards_and_guidances/documents/ENCePPGuideofMethStandardsinPE.pdf. [Accessed December 4, 2013].
[18] Agency for Healthcare Research and Quality. AHRQ user guide for developing a protocol for observation comparative effectiveness research. 2013. Available from: http://www.ncbi.nlm.nih.gov/books/NBK126190/. [Accessed December 4, 2013].
[19] Slutsky J, Atkins D, Chang S. AHRQ series paper 1: comparing medical interventions: AHRQ and the effective health-care program. J Clin Epidemiol 2010;63:481–3.
[20] Helfand M, Balshem H. AHRQ series paper 2: principles for developing guidance: AHRQ and the effective health-care program. J Clin Epidemiol 2010;63:484–90.
[21] Chou R, Aronson N, Atkins D, et al., AHRQ series paper 4: assessing harms when comparing medical interventions: AHRQ and the effective health-care program. J Clin Epidemiol 2010;63:502–12.
[22] Owens DK, Lohr KN, Atkins D, et al., AHRQ series paper 5: grading the strength of a body of evidence when comparing medical interventions–Agency for Healthcare Research and Quality and the effective health-care program. J Clin Epidemiol 2010;63:513–23.
[23] Rangel SJ, Kelsey J, Colby CE, et al., Development of a quality assessment scale for retrospective clinical studies in pediatric surgery. J Ped Surg 2003;38:390–6, discussion 390–6.
[24] Stroup DF, Berlin JA, Morton SC, et al., Meta-analysis of observational studies in epidemiology: a proposal for reporting. Meta-analysis Of Observational Studies in Epidemiology (MOOSE) group. JAMA 2000;283:2008–12.
[25] Fairman KA, Curtiss FR. Rethinking the "whodunit" approach to assessing the quality of health care research–a call to focus on the evidence in evidence-based practice. J Manag Care Pharm 2008;14(7):661–74.
[26] Motheral B, Brooks J, Clark MA, et al., A checklist for retrospective database studies–report of the ISPOR Task Force on Retrospective Databases. Value Health 2003;6:90–7.
[27] Tooth L, Ware R, Bain C, et al., Quality of reporting of observational longitudinal research. Am J Epidemiol 2005;161:280–8.
[28] Atkins D, Best D, Briss PA, et al., Grading quality of evidence and strength of recommendations. BMJ 2004;328:1490.
[29] Theobald K, Capan M, Herbold M, et al., Quality assurance in non-interventional studies. Ger Med Sci 2009:7, Doc 29.
[30] Vandenbroucke JP, von Elm E, Altman DG, et al., Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): explanation and elaboration. Ann Intern Med 2007;147:W163–94.
[31] Chavers S, Fife D, Wacholtz M, et al., Registration of observational studies: perspectives from an industry-based epidemiology group. Pharmacoepidemiol Drug Saf 2011;20:1009–13.
[32] Dreyer NA. Making observational studies count: shaping the future of comparative effectiveness research. Epidemiology 2011;22(3):295–7.
[33] Dreyer NA, Tunis SR, Berger M, et al., Why observational studies should be among the tools used in comparative effectiveness research. Health Aff (Millwood) 2010;29(10):1818–25.
[34] National Academy of Sciences. Initial national priorities for comparative effectiveness research. Available from: http://www.iom.edu/~/media/Files/Report%20Files/2009/ComparativeEffectivenessResearchPriorities/CER%20report%20brief%2008-13-09.ashx. [Accessed December 4, 2013].
[35] Papathanasiou AA, Zintzaras E. Assessing the quality of reporting of observational studies in cancer. Ann Epidemiol 2010;20:67–73.
[36] Sackett D, Strauss S, Richardson W, et al. Evidence-Based Medicine: How to Practice and Teach EBM. London, UK: Churchill Livingstone, 2000.
[37] PCORI methodology standards. Available from: http://www.pcori.org/assets/PCORI-Methodology-Standards.pdf. [Accessed December 4, 2013].
[38] Smith GD, Ebrahim S. Data dredging, bias, or confounding. BMJ 2002;325:1437–8.
[39] Levesque LE, Hanley JA, Kezouh A, Suissa S. Problem of immortal time bias in cohort studies: example using statins for preventing progression of diabetes. BMJ 2010:340.
[40] Ellenberg JH. Selection bias in observational and experimental studies. Stat Med 1994;13:557–67.
[41] Linden A, Adams JL, Roberts N. Evaluating disease management programme effectiveness: an introduction to the regression discontinuity design. J Eval Clin Prac 2006;12:124–31.
[42] Austin PC, Mamdani MM, Juurlink DN, Hux JE. Testing multiple statistical hypotheses resulted in spurious associations: a study of astrological signs and health. J Clin Epidemiol 2006;59:964–9.
[43] International Committee of Medical Journal Editors. Uniform requirements for manuscripts submitted to biomedical journals: manuscript preparation and submission: preparing a

manuscript for submission to a biomedical journal. Available from: http://www.icmje.org/manuscript_1prepare.html. [Accessed December 4, 2013].
[44] Dwan K, Gamble C, Williamson PR, et al., Systematic review of the empirical evidence of study publication bias and outcome reporting bias - an updated review. PloS One 2013;8:e66844.
[45] Forrow L, Taylor WC, Arnold RM. Absolutely relative: how research results are summarized can affect treatment decisions. Am J Med 1992;92:121–4.
[46] Gotzsche PC. Readers as research detectives. Trials 2009;10:2.
[47] Jansen JP, Fleurence R, Devine B, et al., Interpreting indirect treatment comparisons and network meta-analysis for health-care decision making: report of the ISPOR Task Force on Indirect Treatment Comparisons Good Research Practices: part 1. Value Health 2011;14:417–28.
[48] Caro JJ, Briggs AH, Siebert U, et al., Modeling good research practices-overview: a report of the ISPOR-SMDM Modeling Good Research Practices Task Force-1. Value Health 2012;15:796–803.