Prospective Observational Studies to Assess Comparative Effectiveness:

ISPOR Good Research Practices Task Force Report

INTRODUCTION

Context and Background

In both the U.S. and Europe there has been an increased interest in comparative (or relative) effectiveness of interventions to inform health policy decisions. In the U.S., the American Reinvestment and Recovery Act (ARRA) established a Federal Coordinating Council for Comparative Effectiveness Research (CER). This council defined CER as the “conduct and synthesis of research comparing the benefits and harms of different interventions and strategies to prevent, diagnose, treat and monitor health conditions in 'real world' settings”[1]. It noted that the purpose of this research is to inform patients, providers, and decision-makers, responding to their expressed needs, about which interventions are most effective for which patients under specific circumstances. To provide this information, comparative effectiveness research must assess a comprehensive array of health-related outcomes for diverse patient populations. Interventions may include medications, procedures, medical and assistive devices and technologies, behavioral change strategies, and delivery system interventions. Further, it noted that CER necessitates the development, expansion, and use of a variety of data sources and methods to assess comparative effectiveness.

ARRA provided $1.1 billion in funding to the U.S. Secretary of Health and Human Services, the National Institutes of Health, and the Agency for Healthcare Research and Quality to promote CER. At the request of Congress, the Institute of Medicine developed a list of 100 priority topics for CER, most of which involved processes of care rather than specific therapies. Subsequently, US Health Care Reform legislation – the Patient Protection and Affordability Care Act -- created a new entity, the Patient-Centered Outcomes Research Institute (PCORI) to identify national research priorities for CER, appoint advisory panels on research design, facilitate public comment, and disseminate CER findings, as well as to work to improve the science and methods of CER through developing/updating standards on internal validity, generalizability, and timeliness.

In Europe, the EUnetHTA initiative was initiated in 2006 with a work programme focusing on a pan-European “core model” for HTA in Europe, with initial reports on diagnostics, and medical and surgical interventions. The 2011 EUnetHTA work programme includes research on pharmaceuticals and other technologies, reflecting a recent focus in Europe on the relative effectiveness of pharmaceuticals. The Pharmaceutical Forum was developed in 2005 to bring the European Commission, Member States, representatives of the European Parliament, and a wide range of
stakeholders together to examine challenges relating to providing information to patients on
pharmaceuticals, pricing, and reimbursement policy, and relative effectiveness assessment. In its 2008
report [2], the Forum adopted working definitions of efficacy, relative efficacy, effectiveness and relative
effectiveness. These are shown in Table 1 along with this taskforce’s update of the key features.

<table>
<thead>
<tr>
<th>Term</th>
<th>Efficacy</th>
<th>Relative Efficacy</th>
<th>Effectiveness</th>
<th>Relative Effectiveness</th>
</tr>
</thead>
<tbody>
<tr>
<td>Definition:</td>
<td>Extent to which</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>An intervention does more good than harm under ideal circumstances</td>
<td>An intervention does more good than harm, under ideal circumstances, compared to one or more alternative interventions</td>
<td>An intervention does more good than harm when provided under the usual circumstances of health care practice</td>
<td>An intervention does more good than harm compared to one or more intervention alternatives for achieving the desired results when provided under the usual circumstances of health care practice</td>
<td></td>
</tr>
<tr>
<td>Key Features</td>
<td>Randomization versus placebo</td>
<td>Randomization versus active control or use of indirect comparisons of trials versus placebos</td>
<td>Observational study of several competing interventions; or Randomized Naturalistic Pragmatic Clinical Trial</td>
<td></td>
</tr>
</tbody>
</table>

The report noted that the aim of a relative effectiveness assessment is to compare healthcare interventions in practice in order to classify them according to their practical additional therapeutic value. It acknowledged that differences between the objectives and priorities of different national
healthcare systems may create differences in the way in which healthcare interventions will be evaluated relative to one another and differences in relative effectiveness valued. In a survey of 27 member states in 2007, however, the Forum found that little distinction is currently made in member state assessments between efficacy and effectiveness. Member states mostly relied upon relative efficacy data to inform their health technology assessments (HTA) and felt there was inadequate effectiveness data available.

Generating evidence about new pharmaceuticals, including biological entities, is increasingly being seen as an activity that occurs throughout the entire product life cycle rather than pre-launch for a one-off “at-launch” review. Drug Regulatory Authorities are exploring both early access and provisional access schemes in which some studies about effectiveness and safety are conducted post launch. Similarly, HTA/Pricing & Reimbursement bodies are experimenting with “coverage with evidence development” including risk sharing that involves collection of additional data post listing. At the same time, concerns about safety have led to augmented post launch pharmacovigilance requirements. For most of these efforts, prospective observational studies have been the vehicle for data collection.

Like pharmaceuticals, medical devices demand scrutiny across their total life cycle, albeit a life cycle that is typically much shorter than that of drugs. There is a growing debate about future evidence requirements for medical devices in both the US [3] and Europe. Safety and effectiveness evidence for medical devices, along with novel surgical procedures and diagnostics, has typically involved observational studies.

Definitions

For the purposes of this report we apply the following definitions:

**Observational Study**: a study in which participants are not randomized or otherwise pre-assigned to an exposure. The choice of treatments is up to patients and their physicians (subject to any third party payer constraints).

**Prospective Observational Study**: an observational study in which the consequential outcomes of interest occur after the creation of a study protocol and analysis plan, and study commencement. Patient exposure to any of the interventions being studied may have been recorded prior to the study initiation. Exposure may include a pharmaceutical intervention, surgery, medical device, prescription,
decision made to treat, etc. This contrasts with a retrospective observational study that employs existing secondary data sources in which both exposure and outcomes have already occurred.

It is clear that rising requirements for comparative and relative effectiveness evidence will lead to an increase in the number of prospective observational studies that will be undertaken for consideration by health policy decision makers. Given this, the current ISPOR Task Force is focused on setting out good practices in the design, conduct, analysis and reporting of prospective observational studies that will enhance their value to policy makers. The ISPOR Task Force for Good Research Practices for Retrospective Databases Analysis completed its work in 2009 with a focus on comparative effectiveness research [4-6]. A number of the issues addressed by the 2009 Task Force apply also to the design and conduct of prospective observational studies, and we draw on its recommendations where appropriate. There are, however, additional issues that should be addressed when designing a prospective observational study that is “fit-for-purpose” to prospectively test hypotheses about comparative effectiveness in patients. For example, patients may be asked to provide responses to questionnaires which will require appropriate protocol review by institutional review boards and provisions must be made to protect confidentiality of patient identifiable information.

There have been some prior recommendations on the conduct of and reporting of prospective observational studies. The European Medicines Agency in 2010 drafted a code of conduct to help guarantee high standards, scientific independence, and transparency in post-authorization safety studies conducted under the Europe’s pharmacoepidemiology and pharmacovigilance project (ENCePP) [7]. The ENCePP has also created a checklist of methodological research standards for ENCePP studies [8]. The US Agency for Healthcare Research and Quality (AHRQ) published a “User’s Guide” describing the appropriate use of patient registries in 2007, [9] focusing on some specific observational study designs that support CER. In 2008, the International Society for Pharmacoepidemiology (ISPE) published update guidelines for good pharmacoepidemiology practices [10]. To our knowledge, none of the existing recommendations explain or provide general guidance on how to design a strong prospective observational study to address a comparative effectiveness or relative effectiveness research question.

**Choice of Study Design**

The principle potential study designs to assess comparative effectiveness include: retrospective observational studies, prospective observational studies, randomized clinical trials (RCTs), and
naturalistic ("pragmatic") RCTs which we term PCTs. PCTs are intended to maintain the advantages of randomization while examining outcomes in routine care [11]. However, it is recognized that the definition of what is a pragmatic clinical trial falls on a spectrum [12]. Choice of a CER study design follows from the research question but optimal design must, however, consider issues of value of information, clinical equipoise, timing, feasibility, cost, safety (e.g. risks) versus effectiveness (e.g. benefits), ethics, and legality.

Assuming that one is considering the development of an observational study to assess comparative effectiveness, the first choice to be made is whether the design should be retrospective or prospective. Retrospective studies are typically performed using existing data sets and usually offer advantages in terms of cost and speed of execution. However, the data sets may not contain all of the information desired and therefore definition of exposure and outcomes may not be ideal. Prospective studies offer the opportunity to collect the desired information, but usually are more costly and take longer to complete. Formal or informal value of information analysis is useful in making this choice.

While one is developing an observational study design, it is important to consider whether there is clinical equipoise for the treatments of interest, and if not, whether the proposed study design and analysis elements will be sufficient to address issues of confounding and bias. Clinical equipoise has been defined by Freedman as existence of ‘genuine uncertainty within the expert medical community - not necessarily on the part of the individual investigator - about the preferred treatment [13]. Equipoise is defined at both the individual physician/patient and population-levels [14]. When true equipoise exists, an observational study will provide good information to inform comparative effectiveness decisions. Most situations fall somewhere between true equipoise and clear preferences regarding which treatments should be used in specific patients (albeit in the absence of good comparative effectiveness evidence). In the presence of very strong treatment preferences, observational designs and analytic approaches may not be adequate to address for confounding and bias. In these situations, a randomized study design may be preferable and one might explore a naturalistic randomized naturalistic pragmatic clinical trial (PCT) study design.

When the clinical situation does not involve very strong treatment preferences, either a prospective observational study or PCT can be of value and provide useful information. PCTs offer randomization which may be considered quite important when the principle study question focuses on the relative contributions of different treatments to the observed outcomes. Randomization does provide an additional tool to mitigate concerns about potential confounding; however, in naturalistic
PCTs, randomization is not enforced and a significant amount of treatment switching may occur [15, 16]. Accounting for switching using various statistical approaches may enable assessment of the relative contribution of the originally assigned therapies to the observed relative effectiveness. On the other hand, when the principal question of interest is what are the relative outcomes associated with strategies that begin with various treatments, or what are the relative outcomes associated with a device verses a drug, then observational studies may be preferred. Prospective observational studies are generally less costly than PCTs and may pose fewer feasibility issues (e.g. acceptability of randomization) which factor into any calculation of the projected value of information.

When prevailing clinical opinion is very strong, neither a prospective observational study nor a naturalistic pragmatic clinical trial may be able to provide useful information due to confounding. Here, the only option is to perform a standard RCT of comparative efficacy where treatment assignment is enforced, adherence to treatment is monitored, and switching is not permitted; such situations most frequently arise when the treatment in question addresses a fatal condition (FIGURE 1). This was the situation in which the efficacy of high-dose chemotherapy with autologous bone marrow transplantation (HDC-ABMT) as therapy for breast cancer was studied. The therapy made sense and there was strong professional interest in offering this option to patients. Indeed, more than 41,000 patients underwent HDC-ABMT for breast cancer and most health plans provided coverage [17]. In April 2000, the results of a major randomized controlled trial of HDC-ABMT for the treatment of metastatic breast cancer showed that there was no survival advantage to HDC-ABMT relative to standard-dose chemotherapy [18]. Following these results and similar findings from several other RCTs, HDC-ABMT fell out of practice.

There is a relative paucity of head-to-head RCTs that answer questions of interest to policy makers. On occasion, large multi-center RCTs are funded in the United States by the National Institutes of Health or by the Veterans Administration – such as ALLHAT,[19] and these provide important insights into the relative outcomes obtained with various treatments. However, these studies are very expensive to conduct, take a long time to complete and sometimes do not answer questions pertinent to health policy decision makers when the results become available. There is little doubt that such trials will provide important information in the future, however it is likely that the number of such studies that will be conducted will be limited.

Choice of study design may also be dictated by the technology being studied. For devices, particularly implants, and for some biologics interventions, the interventionist’s skill and point on the
learning curve have important impacts on the safety and effectiveness of the product. RCTs are limited in addressing issues of learning as only elite interventionists are usually recruited. The impact of the learning curve can be better evaluated in prospective cohort studies using comprehensive registries with long term follow up [20]. Moreover it is often not possible to conduct double or sometimes even single blinding in implantable device studies and some surgery studies, particularly when the intervention/device is compared to non-intervention. Finally, device development is characterized by frequent modifications of existing products rendering RCTs less useful because by the time the RCT has been reported it is likely that a new generation of the product is in use in routine clinical practice. For these reasons, a prospective observational study is a valuable design option.

Sometimes the study goal may involve questions not answerable by randomized trials – either standard RCTs or naturalistic randomized pragmatic trials. The typical example that comes to mind is that involving the effect of harmful exposures, such as smoking, where it is clear that randomizing subjects to smoke is unethical. Many other situations commonly arise in the public health sector. One common example involves the comparative effectiveness of off-label use of approved therapies (where an RCT must be performed under an IND). Prospective observational cohort studies can provide an opportunity to examine the frequency and characteristics of off-label use of medical products that occur as part of the practice of medicine.

![Diagram of Credibility of Study CE/RE Result](image-url)
SEPARATE CALL OUT BOX: VALUE OF INFORMATION ANALYSIS

Value of information analysis is a useful approach to understanding the trade-offs in selecting different study designs. The objective of this type of analysis is to understand the value information from different types of study can bring (by reducing the uncertainty as to the benefits of using a technology) net of the cost of undertaking different types of study. This cost includes not only the out-of-pocket costs of the study but also the delay (in terms of study duration) involved in resolving the uncertainty, which imposes a burden on patients and the health care system because more time passes before the results can be used to improve treatment choices. Typically retrospective studies are less expensive than RCTs and take a shorter time to complete. Prospective observational studies and PCTs fall somewhere in between with respect to out-of-pocket costs. However, the quality of the information coming from the study will also vary and this needs to be compared with the costs. RCTs typically reduce uncertainty more compared to other study designs, but take longer, cost more, and provide information about fewer comparators. Ultimately the choice of design for a particular issue will involve trade-offs between speed, cost, quality, and relevance to key policy questions.

STUDY DESIGN AND ANALYTICAL STRATEGIES

Specifying the Key Policy Questions – Defining a Successful Study

As the study is designed, it is important to revisit the key policy questions that the study will be conducted to inform. If these questions are not answered with sufficient rigor, the study will not be considered successful. While a well-designed study may answer many questions, there are usually one or two questions that are of greatest interest to health policy makers. The findings of the study must be judged sufficiently valid to inform the policy decisions faced by decision makers. The concept of validity refers to whether a study is able to answer accurately the questions it is intended to answer. Validity requires that measurements accurately reflect the true situations of interest (treatments, outcomes and other characteristics that may influence that likelihood of treatment success), and that the conclusions drawn from analysis of those measurements are well-founded. At a high level, validity is often characterized in terms of “internal validity” the extent to which a measurement accurately reflects what it is intended to reflect, and “external validity” the generalizability of the information to broader settings of patients, physicians, and health care settings.
Part and parcel with assessing validity, policy makers are assessing whether the study permits them to draw causal inferences. Causal inference is a special case of predictive inference where the goal is to first identify subjects who could have received any of the study treatments and then infer treatment effectiveness among these subjects. From the policy decision-maker vantage point, the intent of CER is to understand which treatments “cause” improvement in a condition and what is the relative improvement caused by various treatment options. Making any inferences about causal relationships requires information that is accurate and free from systematic error.

**Specifying the Population, Comparators, and Outcomes of Interest**

**Populations**

Clear specification of the populations, treatments, and outcomes of interest are critical to making inferences regarding causal relationships [21]. The target population of inference can include (1) a very broad group (assuming what would happen if everyone in a given population were exposed to treatment or no treatment, similar to the assumptions underlying a randomized trial), (2) only those who actually receive treatment (e.g. effect of treatment on the treated), or (3) only those who comply with treatment (e.g. the effect of treatment on compliers). In the first situation, if everyone were exposed, the target population is the whole population, and the causal effect is termed the average causal effect. The investigator is interested in determining the causal effect in a population of individuals, regardless of whether they are treatment compliers or not. In other situations – as in scenario (3), the investigator may be focused on determining the treatment effect among those individuals who are likely to comply with their treatment assignment. In this setting, individuals who are thought to always take the treatment or to never take the treatment are eliminated when estimating the causal effect.

There may be more interest in determining the causal effect among those receiving the treatment (scenario 2). For example, in assessing the effectiveness of oral anti-diabetics, a policy maker may be interested in determining the effect of treatment among those who would typically have taken the treatment. In this case, the target population is the population of potential treatment takers. Causal effects of this sort are referred to as the treatment effect on the “treated” and are called local causal effects. Often patients, clinical decision makers, and payers are interested in characterizing the effects of treatment in those who actually have the condition of interest and use the treatment and the researcher is interested in evaluating broader questions of tolerability, adherence and effectiveness.
Finally, in all cases, the target population may consist of subsets of interest, such as the elderly or individuals with specific characteristics, in which the effectiveness of the intervention may be thought to vary. In this case, the investigator believes that treatment heterogeneity may exist.

**SEPARATE CALL OUT BOX: TREATMENT HETEROGENEITY**

*Heterogeneity of treatment effect refers to the fact that individual or groups of patients experience more (or less) benefit from a treatment compared to the average effect of the treatment* [22, 23]. *Reasons for different treatment effects may arise due to biological reasons, patient or provider preferences, or values.*

A statement of the primary hypothesis or research question requires precision and clarity to ensure the design and subsequent analysis provide meaningful evidence.

**Interventions and Comparators**

Determination of the number of treatments and comparators deserves careful consideration in designing a prospective study. In the setting of one treatment (A) and one comparator (B), there is only one possible comparison, namely, A versus B. When there are multiple choices available, beyond treatment B, a better strategy may involve the inclusion of more comparators (Didelez et al. [21]). Factoring into this choice will be an understanding of whether various patient subpopulations are likely to be offered and to receive one or more of the compared therapies. However, as the number of treatments increase, determination of the target population and the causal effects becomes more complicated. For example, suppose a researcher is interested in studying five different drugs. In an RCT where all patients are eligible for each drug (therefore the target population is everyone), all possible pair-wise comparisons could be estimated assuming sample size permits. In a prospective observational study, all patients are not equally likely to receive each of the drugs, because treatments are rarely prescribed randomly, even in conditions of apparent clinical equipoise. Instead, treatment choices are affected by known and unknown factors, some of which may be prognostic. As an illustration, consider when a new asthma treatment is introduced to the market. Generally patients who are doing well on their current treatment would not change to a new drug, especially because new prescription drugs are generally more expensive than products that have been on the market for awhile. Instead, patients who are most likely to switch to the new drug are those who are sicker or who do not tolerate other treatments currently on the market. A simple study of outcomes among people treated
with the new drug compared to other marketed treatments may show a lack of effectiveness of the new
drug, in part, because the new users are sicker and more challenging to cure. Thus it should be
expected that the characteristics of the treated groups will differ. This is sometimes referred to as
channeling by indication, a type of bias, which makes the analyses more difficult and causal
interpretation more challenging. Even if all factors that impacted treatment choice were measured, the
investigator still must determine which comparisons are of primary interest – are all pair-wise
comparisons important or is a comparison of the new drug with each currently marketed drug the focus
of the investigation?

Outcomes

In addition to specifying the population and comparators, it is critical to ensure that the
outcomes of interest are specified and measured. These may include clinical, economic, and
humanistic outcomes; some of this data may be based upon patient response to surveys or
questionnaires. For the latter, validation of these outcomes should be considered through other means
for critical key clinical and economic endpoints if the survey instruments have not been independently
validated. Prospective observational studies also provide the ability to examine a broader range of
clinically relevant clinical outcomes (e.g., disease flares, readmissions, etc.) compared with
retrospective database studies, and have a greater opportunity to specify uniform definitions and data
collection methods for both exposure and outcome.

Potential Study Designs

It is perhaps surprising that there has been little focus in the literature on designing
observational studies to inform health policy decisions despite its importance. If a study is intended to
provide robust-enough information to inform policy decisions, the study must be designed to test a
hypothesis. The Task Force adopted the assumption that an observational study approximates a
randomized study and thus recommended that a necessary first step involves drafting a protocol as if
subjects were to be randomized [24]. This means that investigators should create a formal study
protocol that clearly states the purpose or main hypotheses, defines the treatment groups and
outcomes, identifies confounders (whether measured or not), and specifies the primary analyses and
required sample size.
Choice of a specific design involves balancing the benefits (the internal validity and external validity of the results), and costs of the study. Numerous designs for undertaking observational analyses have been proposed by many researchers across a variety of research fields. Those developed by epidemiologists [25] frequently focus on designs to assess the impact of an exposure (most often a drug or environmental exposure) or an outcome (often an adverse event). For comparative effectiveness research, the Task Force focused on two broad issues – cross-sectional versus longitudinal designs [26], and the Task Force considered only those study designs that utilized at least one comparison group. The comparison group could be comprised of different subjects than those receiving the intervention or the comparison group could consist of the same subjects receiving the intervention measured prior to receiving the intervention (Table 2).

**Single Group, Pre-Test/Post-Test Designs.** These designs are longitudinal studies in a single group of subjects. The pre-test period is defined as the time prior to the intervention or exposure and the post-test period as the time after the exposure. Subjects serve as their own control in that outcomes observed in the post-period are subtracted from the outcomes observed in the pre-period for each subject. In the simple setting of a single pre-intervention outcome measurement and a single post-intervention measurement, the CER estimate is the average of the within-subject differences. The main advantage of this design relates to the benefits of using a subject to serve as his/her own control so that unmeasured time-invariant factors are differenced out. Disadvantages of single-group pre-test/post-test designs involve the inability to control for unmeasured time-invariant confounding and are therefore less valuable in the context of comparative effectiveness research. Another weakness relates to the inability to rule out that changes in the outcomes would have occurred naturally over time.

If outcomes are collected at multiple time points in the pre and the post period, then this design has been referred to as an **interrupted time series design**. These designs have often been utilized to assess policy intervention such as a change in drug benefit. Interrupted time-series are stronger than a single pre-test/post-test design because of their ability to minimize regression to the mean through collection of repeated measurements. However, interrupted time series designs remain vulnerable to time-invariant confounding and to disentangle natural history from treatment effect.

**Multiple Group, Cross-Sectional Cohort Designs.** In this setting, outcomes in multiple groups of subjects are compared. The mean outcome in one group is subtracted from the mean outcome in another
group. A key advantage of these designs includes the ability to quickly measure exposures and outcomes. However, a causal effect is difficult to establish unless the investigator has some assurance the exposure preceded the outcomes and the various groups are comparable along all the dimensions except the one under study. Lack of longitudinal data make this design vulnerable to regression to the mean issues.

**Multiple Group, Pre-Test/Post-Test Designs.** These longitudinal studies involve multiple groups of subjects in which the average change in an outcome from baseline for one group of subjects is compared to the average change in another group of subjects. Most often, the average change in outcome in the exposed group is subtracted from the average change in outcome in the unexposed or comparison group. For this reason, these designs have been referred to as difference-in-difference designs or quasi-experimental designs. The main advantage of this design is that each subject serves as his/her own control so that unmeasured subject-specific confounders are eliminated.

**Single or Multiple Group Prospective Cohort Designs:** Like pre-test/post-test designs, these longitudinal studies involve multiple groups of subjects, often starting with first treatment or diagnosis. Rates of the outcomes of interest are compared, often using relative risks and risk differences. In contrast to the multiple groups, cross-sectional design, the use of multiple groups in the longitudinal setting provides protection against external population factors for which confounding effects are relatively constant across time, e.g., time-invariant residual confounding [27]. However, these designs are vulnerable to time-varying unmeasured confounding as well as systematic unmeasured differences in subjects belonging to different groups.

| Table 2. Comparative Effectiveness Research: Sample Designs for Prospective Observational Studies |
|----------------------------------|---------------------------------|-------------------------------------------------|
| DESIGN                           | DEFINITION                       | ADVANTAGES/DISADVANTAGES                        |
| Single Group, Pre-Post Test Design | Outcomes collected before and after an exposure for a single group (longitudinal) | Subject serves as own control Secular trends confounded with introduction of exposure (subjects or conditions may change over time naturally) Vulnerable to time-invariant and time- |
### Addressing Confounding and Bias in Study Design

There are many types of bias that should be considered when designing an observational study but, fortunately, several tools can be folded into the design to minimize their impact. Major types of bias include channeling bias (discussed earlier), loss to follow-up, and misclassification of treatments and outcomes. For studies of comparative effectiveness, phenomena of episodic use, complex treatment patterns, and drug switching are examples of real-world situations that cannot be avoided and must be considered. Bias can occur in the context of observational studies because alternative
therapies are frequently accompanied by differing management plans. For example, the selection of monitoring tests and the frequency of visits are out of control of the investigator especially when some products under study have mandated periodic safety tests, such as liver chemistries. Such differential management may significantly impact patient outcomes and therefore may confound attempts to understand the relative contributions of the treatments per se. Capturing a larger number of providers with variation in their management choices provides a better opportunity to address this issue.

The choice and the effectiveness of treatments may also be affected by practice setting, the health care environment (e.g., single payer system, fee-for-service), the experience of health care providers, as well as the prior medical history of patients (i.e., inception cohort vs. chronic users). For example, often the convictions of clinicians, whether or not based on evidence, will result in strong treatment preferences related to their training, healthcare system requirements, or individual economic considerations. Researchers may need to conduct some preliminary investigations to understand these preferences. For example, different health plans formularies may not include all the treatment alternatives that one wants to compare or they may not place the alternatives in similar formulary tiers; tier placement is important because it can encourage or discourage the use of certain products. In another instance, surgeons trained to use newer and less invasive surgery may apply this technique in their practice while others are only comfortable with the older procedure. These situations result in variations in care that can be quite significant [28]. Identifying these practice patterns can bolster the validity of a large scale (all inclusive) prospective observational study.

A related bias results from studying prevalent users, rather than new users of a given treatment [29]. When prevalent users are included in a study, it is important to recognize this study design will exclude people who are noncompliant, cannot tolerate the treatment and many people for whom the treatment did not work, because those people will no longer be using the treatment.

Study design choices provide important tools for controlling confounding and various forms of bias by making study groups more similar [30]. Tools that are often used include inception cohorts which focus on people with newly diagnosed disease, or may start with patients when they first require treatment with medication (e.g., first oral anti-diabetic [new users]). Incident and new user designs (Ray, WA [29]) facilitate drug-drug comparisons for people who are initiating a pharmacotherapy, combination medications, new classes of drugs, etc. For example, re-examination of the Woman’s Health Initiative (WHI) observational hormonal replacement treatment (HRT) data were analyzed only for treatment initiators, the results were inferentially the same as for the WHI randomized trial data,
which enrolled treatment naïve patients [31]. The previous analyses of the WHI observational data included both currently treated patients and treatment initiators, causing the differences in results from the randomized trial data. The prior task force on retrospective database analysis has addressed the incident user design and its alternatives (Berger et al [4]).

The goal of these designs is to facilitate comparisons of people with similar chance of benefiting from the treatment, or experiencing harm. For example, a well described bias can result when frail elderly people are included in studies, because this population is treated differently not simply by virtue of their age, but also because their infirmity and co-morbidities. Differences may be so extensive that physicians choose not to prescribe seemingly reasonable treatments for fear of complications or unknown drug interactions [32]. A more challenging comparison presents itself when trying to study medications in comparison to surgery, because the patients considered good surgical candidates frequently differ significantly from those to whom medical treatment is offered. Without an extensive, heterogeneous pool of comparators to draw from, observational studies may not be able to address the intractable bias that would result from such comparison.

The collection of additional outcomes thought not to be impacted by choice of intervention can bolster findings from observational studies. These “control” outcomes are outcomes believed not to be associated with treatment. The Task Force recommends the usefulness of such outcomes at study onset with pre-planned collection. If a clinically meaningful difference is observed between treatment and comparison groups for the control outcomes, then this provides evidence of unmeasured confounding. For example, Mauri and colleagues [33] examined the effectiveness of drug eluting coronary stents compared to bare metal stents using mortality two days from stent implant as a control outcome. Mortality differences between those implanted with a drug eluting coronary stent compared to those implanted with a bare metal stent at two days are not plausible and so if observed would indicate residual confounding.

Similarly, the use of multiple comparison groups can bolster findings. In some settings, two comparison groups that differ on a confounder the researcher knows a-priori cannot be collected, may still permit some comparisons. For example, in examining the effectiveness of two drugs, A and B, some subjects may receive drug B and participate in the registry, while other subjects may not participate in the registry and also receive drug B. The investigator may believe those that participate in a registry are different from those that do not in terms of compliance and life-style. If some relevant outcomes data are available for the non-participants, then both cohorts who received drug B could
serve as comparison groups for drug A. If a similar effect of treatment A compared to B is observed using either comparison group, then concern about unmeasured bias due to compliance or life-style is reduced. Yoon et al. used multiple control groups to examine effectiveness of the passing of the Mental Health Parity Act [34].

**SEPARATE CALL OUT BOX: DESIGN TOOLS**

<table>
<thead>
<tr>
<th>Problem</th>
<th>Design Solution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unmeasured differences in study groups</td>
<td>Outcomes not impacted by intervention</td>
</tr>
<tr>
<td></td>
<td>Multiple comparison groups (patients, providers and/or institutions)</td>
</tr>
<tr>
<td></td>
<td>Inception cohorts</td>
</tr>
<tr>
<td></td>
<td>New user designs</td>
</tr>
<tr>
<td>Differing patient management and monitoring plans</td>
<td>Increase number of providers participating in study</td>
</tr>
<tr>
<td>Skill or experience of provider</td>
<td>Stratify sample within provider groups defined by experience</td>
</tr>
</tbody>
</table>

**SEPARATE CALL OUT BOX: Managed Entry Agreements**

*When there is uncertainty regarding the effectiveness and cost-effectiveness of a new therapy, payers have the option to provide coverage for them with restrictions. The major strategies applied have included coverage with evidence development, coverage only in research, and managed entry agreements (e.g. risk-sharing schemes). Managed entry schemes have been defined as arrangements between manufacturers and payers/providers that enable access to (coverage/reimbursement of) health technologies subject to specified conditions; these arrangements can use a variety of mechanisms to address uncertainty about the performance of technologies or to manage the adoption of technologies in order to maximize their effective use, or to limit their budget impact [35].*

*In the two well-discussed cases of outcome based risk sharing schemes, the parties opted to create a patient registry or a single-arm prospective cohort observational study. In the case of Bosentan in Australia, a registry was established for patients receiving the new treatment, and in the case of treatments for Multiple Sclerosis in the UK, a single-arm prospective cohort observational study was conducted. For both of these risk-sharing schemes, the objective was to use observational data to*
improve the model-based estimate of the incremental cost effectiveness of the treatments compared to relevant alternatives. The MS scheme has been highly controversial in part because of difficulties with the analysis of effectiveness [36-38].

It is the view of this taskforce, that these studies should where possible be designed following good practice recommendations for comparative effectiveness research. Ideally studies should either be conducted as prospective observational studies using appropriate concurrent comparator groups, or they should be conducted as pragmatic clinical trials. For the former, patients would need to be enrolled who do not receive the therapy under investigation. In a single payer system it may be expected that few patients would be denied the new treatment. In this case a single arm study would provide information but understanding of the nature of the historic control would be needed and the potential for bias would need to be recognized.

Other alternatives such as a pragmatic clinical trial may be more appropriate. In a PCT, one arm may receive a new therapy under the same rigorous set of monitoring conditions as in the pivotal RCT, and the other arm may receive the therapy under ordinary practice conditions. This may require that such studies be implemented as cluster randomized protocols.

It may also be appropriate to collect evidence of the comparative effectiveness of the treatment in another jurisdiction where treatment choices are more mixed. Clearly issues of the transferability of results from one area to another would need to be addressed in this situation. These issues have been addressed by another ISPOR taskforce in a reported titled “Transferability of Economic Evaluations Across Jurisdictions: ISPOR Good Research Practices Task Force Report” [39]. We also note that the ISPOR Performance Based Risk Sharing Task Force is preparing a paper on GRP for risk sharing and related schemes.

**Analytical Approaches to Address Potential Bias and Confounding**

The analytic approaches and tools recommended by the ISPOR Task Force on Good Research Practices for Retrospective Database Analysis to mitigate threats to validity from bias and confounding in measurement of exposure and outcome, apply equally well to both retrospective and prospective observational studies. Its recommendations included the need for data analysis plan with causal diagrams (Cox et al. [5]), detailed attention to classification bias in definition of exposure and clinical outcome, careful and appropriate use of restriction and extreme care to identify and control for confounding factors, including time-dependent confounding. This Task Force Report (Johnson et al.
[6]) also recommended general analytic techniques and specific best practices including use of stratification analysis before multivariable modeling, multivariable regression including model performance and diagnostic testing, propensity scoring, instrumental variables, and structural modeling techniques including marginal structural models, as well as rigorous sensitivity analysis. These recommendations will not be further discussed in this report.

One type of bias not discussed in the Retrospective Database taskforce report is immortal time bias. Immortal time is a span of cohort follow-up during which, because of exposure definition, the outcome under study could not occur [40]. This can confound the results of a prospective observational study in two situations: when the time between cohort entry and date of first exposure is not accounted for in the analysis; and when a decision is made to switch treatment between the time that treatment is initially planned and when treatment is actually initiated. In the first situation, consider the challenge of understanding time to treatment with prescription medications, where people get sick before coming to medical attention but can only be treated once they come to attention. People who die before coming to medical attention would not be included and effects that occur before coming to medical attention cannot be studied. For this situation, restriction or matching methods and time-dependent covariate analyses have been proposed [41]. For the second situation, intention-to-treat methodology may be applicable, depending on the research question.

The analytic plan for a prospective observational study of comparative effectiveness should consider issues of treatment complexity (switching, combination, sequencing, dosing, and adherence/compliance), as may be expected to present themselves in the study. Various guides are available to help researchers examine and anticipate such bias, notably the ENcEPP Methods guidance, [42] as well as other documents intended to help readers evaluate good quality in observational studies of comparative effectiveness [43].

**Addressing Treatment Heterogeneity in the Analytic Plan**

Most RCTs are underpowered to test for treatment heterogeneity given such designs effectively requires testing for interaction terms. Post-hoc analyses of treatment interactions are vulnerable to false discovery rates. However, to achieve the goals of CER rigorous methods to estimate treatment heterogeneity will be needed. Prospective observational studies may be better positioned to assess treatment heterogeneity given the larger numbers of subjects that can be accrued. The absence of treatment heterogeneity is a crucial assumption for virtually all analytical approaches including the use
of differences in means, propensity scoring, structural mean models, and instrumental variables. These analytical approaches as well as standard analyses can accommodate interactions involving baseline modifiers, but not modifiers that are measured after treatment is selected (e.g., changing disease status, comorbidities, or co-treatments).

**Intention-to-Treat Compared to As-Treated Treatment Analyses**

The impact of longer follow-up periods with higher drop-out and treatment non-adherence/switching depends on the primary planned treatment effect analysis: intent-to-treat or as-treated (on-treated). Intent-to-treat analyses in randomized pragmatic studies estimate the treatment effects that would occur in populations with similar levels of treatment non-adherence and switching. As-treated (on-treated) analyses are more useful in evaluating the effect of time-varying treatments and switching. Standard as-treated approaches ignore randomization and remain vulnerable to unmeasured confounding.

More rigorous approaches based on instrumental variable methodology make use of randomization in protecting as-treated analyses against unmeasured confounding. That is, one of the ideal instrumental variables is randomization [44]. Hence, in this way randomized pragmatic trials may be superior to observational studies. However, this protective property of randomization is reduced with longer follow-up and larger magnitudes of treatment non-adherence, such as treatment discontinuation and switching, and study drop-out. These reduce the advantage of randomized pragmatic trials over prospective observational studies in practice. Stated another way, randomization is most protective against residual confounding with intent-to-treat analyses, but is less useful for as-treated analyses, or when there is an expectation of significant treatment non-adherence. When a prospective observational study is focused on comparing initial treatment decisions using intent-to-treat analyses (with propensity scores or other techniques to adjust for observed factors influencing treatment selection), comparisons among follow-up treatment decisions or lack of adherence to the initial treatment decisions (with marginal structural models) may be presented as as-treated or on-treated analyses (Hernan et al. [31]). Such an analytic approach played a role in resolving the conflicting HRT evidence from observational and randomized trials. Absence of treatment heterogeneity and no residual confounding are assumed in these situations; unless an instrumental variable can be identified that satisfies the necessary assumptions (Small et al. [44]).
Sample Size Calculations

The analytic and design strategies for benefit in terms of relative effectiveness conflict sometimes with those for risk in terms of relative safety. In theory, relative effectiveness and relative safety could be addressed with the same analytical methodology since the distinction between the safety and effectiveness endpoints is artificial – at its simplest, an adverse event is negative effectiveness. However, the efficacy comparisons between treatments can entail either non-inferiority or superiority hypotheses (e.g., is Drug A equivalent/better than Drug B regarding reduction in blood pressure). In contrast, the safety assessments are much less likely to have superiority hypotheses, particularly when the safety endpoint is relatively rare. The infrequency of events results in larger sample size requirements for assessment of safety and for assessing superiority as compared to non-inferiority. Additionally, the importance of as-treated analyses may be greater for assessing safety than effectiveness. Thus, for safety assessment the role of randomization is limited and given larger sample size requirements prospective observational studies are more suitable to address safety concerns, especially in real world use of a medical technology.

Sample size specification also needs to accommodate issues of confounding and treatment heterogeneity, but care should be taken to ensure that a study is not overpowered as enrollment of large numbers of patients without scientific justification undermines the study’s scientific credibility and may suggest that study sponsors have primarily a non-scientific agenda, such as building experience with a new product. There is debate in general, both for observational or randomized studies, about whether effect sizes should come from pilot study data, the literature, or from a general consensus about clinically significant effect sizes [45]. Estimating effect sizes require large enough pilot sample sizes to ensure adequate precision. Obtaining effect sizes from literature requires adequate synthesis based on relevant past studies with enough information for quantitative synthesis methods such as meta-analysis (including mixed treatment comparisons and network meta-analysis). Ideally, clinically significant effect sizes should be recognized either formally as a consensus among clinical researchers and/or societies of practitioners, or informally in the medical literature.

Regardless of how effect sizes are obtained, confounding increases the number of patients required to study because of the need to identify groups of patients who are similar on the observed covariates and because of other design considerations (such as the inclusion of additional comparison groups to bolster the causal interpretation). The design of PCTs must account for the expectation of treatment switching. Pilot studies or previous studies with a similar focus may inform the magnitude of
confounding by selected observed factors that can be accounted for in effect size determination. Additional information from previous studies include drop-out rates, treatment non-adherence rates (which affect the effect size), within-subject correlations for longitudinal data analyses, and within-provider/within-site correlations for studies involving many providers or sites randomized at either the patient, provider, or site level.

The planned assessment of treatment heterogeneity based on a priori stratification factors requires sufficient sample sizes for each combination of factors to facilitate tests of treatment heterogeneity with model-based interactions. For estimation of sample size that take into account treatment heterogeneity due to a single effect modifier, a general rule of thumb is that roughly a 50% increase in sample size is required relative to the sample size for detecting a simple treatment effect [46]. However, it is important to recognize that even small observational CER studies may provide important insights, recognizing the limitations described above.

STUDY EXECUTION

Given the aim of comparative effectiveness research is to inform health policy decisions, the recommendations of both the ISPOR Retrospective Database Taskforce and GRACE (Good ReseAch for Comparative Effectiveness) (Dreyer NA et al [43]) included that a formal written protocol specify the a priori research questions and study design to assure end-users that the results were not the product of data-mining. These recommendations similarly apply to prospective observational studies.

SEPARATE CALL OUT BOX: Sample Study Protocol Outline

1. Purpose: What is the key health policy question that the study is designed to answer.
2. Background: What is the current state of knowledge?
3. Hypotheses: What is the primary hypothesis? What are the secondary hypotheses (if any)?
4. Study Design:
   a. Study design and rationale
   b. Definition of population (patients, providers, sites) that will be studied (target of inference)
   c. Definition of treatment cohorts to be compared
   d. Definition of outcome measures to assess treatment effects
   e. Definition & justification of control outcome (if any)
5. **Data Analysis Plan:**
   a. Sample size justification
   b. Primary planned analyses, secondary planned analyses
   c. Definition of relative effectiveness measure or causal effect (average causal effect, local causal effect)
   d. Planned approaches to deal with bias, confounding, missing data, and multiple outcomes (if secondary outcomes)
   e. List confounders (whether collected or not)

6. **Study Governance and Implementation:**
   a. Governance and sponsorship
   b. Incentives for participation (if any)
   c. Safety reporting
   d. Informed consent and IRB approval (if required)
   e. Data processing and quality control
   f. Sample listing of data elements
   g. Plan for dissemination of results and publication planning
   h. If the study is designed to support a policy decision, explanation of decision to register study or not
   i. Anticipated timing of dissemination and publication of study results

Strengths of a prospective observational study relate not only to the ability to *a priori* implement common definitions of covariates and outcomes, but also the ability to collect potential confounders, control outcomes, and additional comparison groups. Like an RCT, specifics of the study inclusion and exclusion criteria (again, prior to collection of outcomes) should be listed. Because subjects will not be randomized to treatments, a list of possible confounders should be composed and compared to those that are feasible to collect. Protocols for minimizing missing information should be specified -- these may involve a fixed number of attempts to retrieve the missing data (e.g., up to five call-backs) or use of proxies when feasible.

If the cost of collecting clinical outcomes on many subjects is prohibitive or when a large number, relative to the sample size, of observed confounders have been selected *a priori*, the researcher could utilize a matched design in which treated and comparison subjects are selected
through matching or stratification on the basis of a balancing score, such as the propensity score, and then followed [47]. This would permit more detailed data collection on a smaller sample of subjects. Matched designs provide a mechanism to minimize confounding through the selection of subjects for whom the researcher believes *a priori* are similar with the exception of the treatment received.

Various issues related to study execution can influence the validity of the results including selection bias in recruitment of patients, of health care providers and of sites. Because of the increased likelihood of treatment heterogeneity with post-marketing studies and the need to assess it, every effort should be made to complete studies, even with sample size imbalance across treatment arms. The reason for strengthening the likelihood of study completion in the presence of treatment heterogeneity relates to the fact that heterogeneity implies more variability, and hence more sample size to preserve power. Moreover, in the presence of imbalance across treatment arms, the threat of residual confounding due to unmeasured factors such as treatment preferences increases. The statistical design and analysis approaches to account for measured and unmeasured confounders apply here, unless the imbalances are severe (e.g., standardized differences of 20% or larger between treatment groups). In such cases, early termination with a futility analysis may then be needed. However, it should be noted that with such severe imbalances, a futility analysis may not be informative.

No matter how good a study concept, protocol, and intentions of sponsors and collaborators, the details of good field work are distinguishing characteristics of successful studies. In observational studies, the challenges are greater than in randomized clinical trials, because the investigators need to work within existing health systems, patient populations, and physician groups to conduct exacting science within the constraints of an unruly “real world.” Study management and implementation includes issues relating to governance, including the involvement of expert advisors to provide guidance (advisory boards), questions about appropriate remuneration for researchers and subjects/participants, and safety reporting, especially for studies where safety may not be the main objective. Questions of interest relating to reporting and publication include whether observational studies should be registered, and how to work with co-authors on publications such that they ethically can attest that they have participated in data analysis and whether this requires providing all co-authors with copies of the data sets, and if and when data access should be granted to non-participating investigators.

**Good Governance**
Advisory boards can be useful to promote avoidance of conflicts of interest, and appropriate study execution and reporting. Large observational clinical studies often include scientific advisory boards comprising clinician disease experts, statisticians and epidemiologists. While the role of advisory boards varies from project to project, these boards can provide on-going guidance about operational issues that may influence validity. For example, sequential enrollment plans may not always be feasible and investigators may propose alternatives which require scrutiny in terms of protection from bias. The most successful advisory boards have written charters which lay out the roles and responsibilities, including rules for voting at formal meetings and in special situations that may arise between meetings.

**Incentives for Participation**

Institutional review boards (IRB) are appropriately sensitive to the threat of coercion that can stem from excessive compensation for participation in research. Payment to health care providers, if any, should be commensurate with work performed. In some instances, it may be acceptable to provide a modest bonus for completion of a long term study or for studies that involve low risk procedures (such as endoscopy, multiple blood samples for pharmacokinetic studies, gynecological examinations, etc.). Patients are sometimes offered gift cards; once again, the value of the gift should be related to the length of time and effort required to participate in the study.

Some studies do not offer any payment to health care providers for time spent preparing and submitting data for a study. However, incomplete reporting is more common when there are no payments or other incentives for participation. Payments are often made per patient visit, rather than using an annual payment system. Paying on a per patient visit basis encourages health care providers to complete case report forms for each visit for which they will be paid.

**Is it worth “registering” an observational study?**

Similar to the concept of the importance of declaring a prespecified hypothesis, some feel that validity is enhanced by registering studies in advance of their execution. Many groups register their studies on clinicaltrials.gov, although there is no mandate or requirement for that reporting. The Agency for Healthcare Research and Quality in the United States has recently initiated a contract for development of a Registry of Patient Registries in the US; recommendations are under development. In Europe, the ENcEPP recently created a voluntary registry for observational pharmacoepidemiology
and pharmacovigilance studies, and has instructed that any study that wishes to receive the ENcEPP Seal of Approval be registered before it commences. The ENcEPP program also requires that the protocol be provided and provides a mechanism for the protocol to be kept confidential for some period of time. For comparative effectiveness research, registration of prospective observational studies will enhance their credibility with key decision makers and we encourage researchers to register their studies.

**Data Access**

Routinely, the editors of scientific journals ask authors to sign a statement declaring that they had full access to all study data used in the development of the paper. However, “full access” to study data is not defined. In major randomized clinical trials, some funding agencies and journal editors have required that authors include a statement in the published paper that readers may contact the first author to request access to the full study data set (e.g. Hull, et al. EXCLAIM study) [48]. While the rationale for the FDA requirement that a pharmaceutical company supply the raw data used in their analysis of clinical trials to FDA statisticians in support of a NDA (new drug application) is compelling, [49] the appropriate balance of interests in access to raw study data between study sponsors, journal editors and interested individuals who wish to conduct an independent examination of raw study data, is less clear for observational clinical studies.

In general, proprietary observational disease registries have not routinely shared their raw data with investigators outside of the study coordinating center team. This caution in allowing broad access to data reflects scientific, ethical concerns about data ownership, patient privacy, and contractual obligations to sponsors, physicians and hospitals, who have agreed to share their patient’s data in a proscribed, limited manner. However, a number of observational clinical registries have developed written publication guidelines for data access and approval of publication topics and authorship, which at least clarify how interested parties may approach investigators for collaborative research, if not providing direct access to the data of interest (e.g., GRACE [50], CARDIA [51], GLOW [52], and CRUSADE [53].) A number of limited data sets have been shared with bona fide investigators from outside of the scientific advisory boards of these studies [54]. Further, federally sponsored multicenter data sets often include a mandatory requirement allowing individuals unrelated to the leadership of the study to request access to all or major portions of the raw study data [55]. Also some journal editors have taken the position that study data should be made available to others, e.g., the British Medical
Journal, [56] provided that such release has been guaranteed in the patient informed consent, and, we would add, in the agreements with participating investigators.

Most often, however, authors of scientific papers that examine data from industry-sponsored multicenter voluntary observational clinical databases, for which the data are held and analysis is performed by a central study coordinating center (e.g. a university or contract research organization) have limited access to the raw study data, but have received summary data and appropriate statistical findings for all of the questions they have proposed in preparing the manuscript for submission to scientific journal. In this model, the coordinating center conducts the analysis and provides collaborating authors with copies of analytic tables and figures during the course of analysis and writing. For observational studies of comparative effectiveness that use de novo data collection, this practice is preferred both for practical and professional reasons. The investigators who created and operate the study have the best understanding of the data definitions and the circumstances under which they were collected. Thus, the absence of a critical data element may be interpreted differently according to how the data were collected. For example, in claims data a missing element may indicate that a test was not done; yet depending on the nature of the data source, it may also reflect that those tests are not covered under by the health insurance provider which provided the data for the study. In other situations, missing treatment data more likely may be assumed to indicate no such treatment, such as when data are obtained through chart abstraction [57].

**Safety Reporting**

Operational issues may occur because of safety reporting obligations of funding sources. In this situation, selective recording of reportable adverse events and serious adverse events should be conducted uniformly across all comparators, and not simply for the products for which the funders have market authorization. Whether or not safety events are required to be reported to federal authorities depends on funding source [58].

**REPORTING STUDY FINDINGS**

Reporting should be in line with good practice in pharmacoepidemiology studies and retrospective database analysis and not specific to prospective observational studies (Berger et al. [4], Cox et al. [5], and Johnson et al. [6] [59-61]. We also note that CONSORT can be useful for the reporting of non-randomized studies [62] while the STROBE guidelines [63] and the MOOSE guidelines are designed
specifically for observational studies. Reporting should be consistent in a way that enables it to be used in meta-analysis [64]. To this end, some journals have required a “Reproducible Research Statement” included in the manuscript [65]. Thus, specific reporting features should include

- A clear statement of the study question
- Outline of the research design [and the reasons for choice of design] and methods [source population, target population, selection of subjects, data collection methods, statistical methods]
- Sources of data
- Any circumstances that may have affected the quality and integrity of the data
- Analysis including adjustments for confounding
- Results including quantitative estimates of effect, confidence intervals, and sensitivity analyses
- Statement of conclusions and implications
- Reproducibility research statement

**INTERPRETATION**

The GRACE principles set out a set of principles that can help in recognizing high quality observational studies of comparative effectiveness. The three key principles are:

- Were the study plans specified before conducting the study, including the outcomes and comparisons?
- Was the study conducted consistent with good practice and reported in sufficient detail for evaluation and replication?
- How valid was the interpretation of the results for the population of interest?

The key challenge in the interpretation of prospective observational studies is the potential for confounding. How much could unmeasured confounders have influenced the results? Unmeasured confounding is much more likely to explain small effects (odds ratios) than large ones. Beyond using statistical techniques, as discussed above, there may be other ways to improve confidence in the results. For example, are there other studies in this sub-population and for any of the treatment comparators? Consistency of findings would enhance the validity of the findings. If there are RCTs conducted on the same comparators, then confidence in the study results (lack of confounding) can be improved if the results for the relevant subgroups of the prospective observational study match those of the RCT (Berger ML, et al. [4] [66].

28
Interpretation may also need to take account of the funding source for the study and authors’ right to participate in the study design, select hypotheses, pose queries on the data, and exercise editorial control over the final publication.

RECOMMENDATIONS AND CONCLUSIONS

Prospective observational studies will undoubtedly be conducted with increased frequency to assess the comparative effectiveness of different treatments, whether as part of the overall increase in interest and funding for CER, or as a tool for “coverage with evidence-development”, “risk-sharing contracting”, or as a key element in a “learning health care system.” This report summarizes the challenges and approaches to the appropriate design, analysis, and execution of prospective observational studies to make them most valuable and relevant to health care decision makers. Some of the important points made in this report include:

- Precision and clarity in specifying the key policy questions to be addressed is paramount and studies should be designed with a goal of drawing causal inferences whenever possible.
- If a prospective observational CER study is being performed to support a policy decision, then it should be designed to test a hypothesis – the protocol should clearly state the main hypothesis (or research questions), define the treatment groups and outcomes, identify measured and unmeasured confounders, and specify the primary analyses and required sample size.
- Careful consideration should be taken in choosing to perform a prospective observational study over alternative CER designs: retrospective observational studies, rigorous randomized controlled clinical trials, and pragmatic clinical trials. Researchers should provide their rationale for opting for a prospective observational study and discuss critical issues such as the nature of the key research question, feasibility, value of information (likelihood of answering the question versus timeliness and cost), and any specific considerations with respect to the technology (ex. Devices, procedures) being evaluated.
- Separate from analytic and statistical approaches, study design choices may strengthen the ability to address potential biases and confounding in prospective observational studies. Use of inception cohorts, new user designs, multiple comparator groups, matching designs, and assessment of outcomes thought not to be impacted by the therapies being compared are
several strategies that should be given strong consideration recognizing that there may be feasibility constraints.

- The reasoning behind all study design and analytic choices should be transparent and explained in study protocol.
- Execution of prospective observational studies is as important as their design and analysis in ensuring that results are valuable and relevant, especially capturing the target population of interest, having reasonably complete and non-differential follow-up.
- Similar to the concept of the importance of declaring a pre-specified hypothesis, we believe that the credibility of many prospective observational studies intended to be used for decision-support would be enhanced by their registration on appropriate publicly accessible sites (e.g. clinicaltrials.gov, encepp.eu) in advance of their execution.

References


14 Ten Have TR, Coyne J, Salzar M, Katz I. Research to improve the quality of care for depression: alternatives to the simple randomized clinical trial, General Hospital Psychiatry, 2003;25:115-123.


30 Weiss N. The New World of Data Linkages in Clinical Epidemiology. Are We Being Brave or Foolhardy? Epidemiology 2011; 22:292-29-4

31 Hernan MA, Alonso A, Logan R, et al., Observational studies analyzed like randomized experiments: an application to postmenopausal hormone therapy and coronary heart disease, Epidemiology, 2008;19:766-79


35 Klem M, Fronsdal KB, Facey K, What principles should govern the use of managed entry agreements?, Int J Technol Assess Health Care, 2011;27:77-83. 1

36 Lilford RJ, MS risk sharing scheme. Response from chair of scientific advisory committee, BMJ, 2010;341:c3590.
50 Fox KA, Anderson FA, Jr., Dabbous OH, et al., Intervention in acute coronary syndromes: do patients undergo intervention on the basis of their risk characteristics? The Global Registry of Acute Coronary Events (GRACE), Heart, 2007;93:177-82.

53 Mudrick DW, Chen AY, Roe MT, Newby LK, Gibler WB, Ohman EM, Peterson ED, Alexander KP. Changes in glycoprotein IIb/IIIa inhibitor excess dosing with site-specific safety feedback in the Can Rapid risk stratification of Unstable angina patients Suppress ADverse outcomes with Early implementation of the ACC/AHA guidelines (CRUSADE) initiative. Am Heart J. 2010 Dec;160(6):1072-8

54 Miller JD. Sharing clinical research data in the United States under the Health Insurance Portability and Accountability Act and the Privacy Rule. Trials. 2010 Nov 19;11:112


56 British Medical Journal, Resources for Authors: Research [criteria for published articles], 2011;http://resources.bmj.com/bmj/authors/types-of-article/research, accessed: 7 September 2011.


Schneeweiss S, Patrick AR, Sturmer T, et al., Increasing levels of restriction in pharmacoepidemiologic database studies of elderly and comparison with randomized trial results, Med Care, 2007;45:S131-42.