VALUE IN HEALTH 15 (2012) 217–230

Available online at www.sciencedirect.com

SciVerse ScienceDirect

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

ISPOR TASK FORCE REPORT

Prospective Observational Studies to Assess Comparative Effectiveness: The ISPOR Good Research Practices Task Force Report

Marc L. Berger, MD[1,], Nancy Dreyer, PhD, MPH[2], Fred Anderson, PhD[3], Adrian Towse, MA[4], Art Sedrakyan, MD, PhD[5], Sharon-Lise Normand, PhD[6]*

[1]*OptumInsight, Life Sciences, New York, NY, USA;* [2]*Outcome, Cambridge, MA, USA;* [3]*Center for Outcomes Research, University of Massachusetts Medical School, Worcester, MA, USA;* [4]*Office of Health Economics, London, UK;* [5]*Comparative Effectiveness Program at HSS and NYP, Weill Cornell Medical College, New York, NY, USA;* [6]*Harvard Medical School, Department of Health Care Policy, Boston, MA, USA*

ABSTRACT

Objective: In both the United States and Europe there has been an increased interest in using comparative effectiveness research of interventions to inform health policy decisions. Prospective observational studies will undoubtedly be conducted with increased frequency to assess the comparative effectiveness of different treatments, including as a tool for "coverage with evidence development," "risk-sharing contracting," or key element in a "learning health-care system." The principle alternatives for comparative effectiveness research include retrospective observational studies, prospective observational studies, randomized clinical trials, and naturalistic ("pragmatic") randomized clinical trials. **Methods:** This report details the recommendations of a Good Research Practice Task Force on Prospective Observational Studies for comparative effectiveness research. Key issues discussed include how to decide when to do a prospective observational study in light of its advantages and disadvantages with respect to alternatives, and the 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. **Recommendations:** The task force emphasizes the need for precision and clarity in specifying the key policy questions to be addressed and that studies should be designed with a goal of drawing causal inferences whenever possible. If a study is being performed to support a policy decision, then it should be designed as hypothesis testing—this requires drafting a protocol *as if* subjects were to be randomized and that investigators clearly state the purpose or main hypotheses, define the treatment groups and outcomes, identify all measured and unmeasured confounders, and specify the primary analyses and required sample size. Separate from analytic and statistical approaches, study design choices may strengthen the ability to address potential biases and confounding in prospective observational studies. The 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 nondifferential follow-up. Similar to the concept of the importance of declaring a prespecified hypothesis, we believe that the credibility of many prospective observational studies would be enhanced by their registration on appropriate publicly accessible sites (e.g., clinicaltrials.gov and encepp.eu) in advance of their execution.

Keywords: comparative effectiveness, prospective observational studies.

Introduction

Context and background

In both the United States and Europe there has been an increased interest in comparative (or relative) effectiveness of interventions to inform health policy decisions. In the United States, the American Reinvestment and Recovery Act 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, CER 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. Furthermore, it noted that CER necessitates the development, expansion,

[*] *Address correspondence to:* Marc L. Berger, OptumInsight, Life Sciences, 301 West 118th Street, Apartment PH 2-C, New York, NY 10026, USA.

E-mail: Marc.Berger@Optum.com.

1098-3015/$36.00 – see front matter Copyright © 2012, International Society for Pharmacoeconomics and Outcomes Research (ISPOR). Published by Elsevier Inc.

doi:10.1016/j.jval.2011.12.010

Table 1 – Categories of intervention effects.

Term	Efficacy	Relative efficacy	Effectiveness	Relative effectiveness
Definition: *Extent to which*	An intervention does more good than harm under *ideal* circumstances	An intervention does more good than harm, under *ideal* circumstances, compared with one or more alternative interventions	An intervention does more good than harm when provided under the *usual* circumstances of health-care practice	An intervention does more good than harm compared with one or more intervention alternatives for achieving the desired results when provided under the *usual* circumstances of health-care practice
Key features	Randomization versus placebo; select patients; high-volume centers	Randomization versus active control; or use of indirect comparisons of trials versus placebos or active comparators	Observational study; heterogeneous patient population; typical treatment environment; comparison typically made to other treatments	Observational study of several competing interventions; or randomized naturalistic pragmatic clinical trial

and use of a variety of data sources and methods to assess comparative effectiveness.

The American Reinvestment and Recovery Act 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, U.S. Health Care Reform legislation—the Patient Protection and Affordable Care Act—created a new entity, the Patient-Centered Outcomes Research Institute, 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 and updating standards on internal validity, generalizability, and timeliness.

In Europe, the European Network for Health Technology Assessment initiative was launched in 2006 following the request of European Union member states in the High Level Group on Health Services with a work program focusing on a pan-European "core model" for health technology assessment in Europe, with initial reports on diagnostics and medical and surgical interventions. The 2011 European Network for Health Technology Assessment work program 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, 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 task force's update of the key features.

The report noted that the aim of a relative effectiveness assessment is to compare health-care interventions in practice to classify them according to their practical additional therapeutic value. It acknowledged that differences between the objectives and priorities of different national health care systems may create differences in the way in which health-care 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 on relative efficacy data to inform their health technology assessments and felt that 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 prelaunch 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 postlaunch. Similarly, health technology assessment and pricing and reimbursement bodies are experimenting with "coverage with evidence development" including risk sharing that involves collection of additional data postlisting. At the same time, concerns about safety have led to augmented postlaunch pharmacovigilance requirements. For most of these initiatives, prospective observational studies have been the vehicle for evidence 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 United States [3] and Europe. Safety and effectiveness evidence for medical devices, along with novel surgical procedures and diagnostics, has typically involved observational studies.

The ISPOR Board of Directors approved on May 16, 2010, the formation of the Prospective Observational Clinical Studies Good Research Practices Task Force to develop good research practices for prospective observational clinical studies that focus on the effectiveness and/or comparative effectiveness of health-care interventions. Researchers, experienced in biostatistics and outcomes research working in academia, government health organizations, contract research organizations, and hospitals from the United States and the United Kingdom, were invited to join the Task Force Leadership Group. The task force met about once a month to develop the topics to be addressed and outlined and to prepare the first draft report. A face-to-face meeting was held on March 23, 2011, to debate and finalize any contentious issues in the draft report. The draft report was presented for comment at the ISPOR 13th European Congress in Prague, Czech Republic, in October 2010 and the ISPOR 16th International Meeting in Baltimore, MD, USA, in May 2011. The draft report was sent for comment to the Task Force Reviewer Group (82 invited and self-selected individuals interested in this topic) on October 12, 2011. Comments were then considered. The final draft report was sent for comment to the ISPOR membership via the ISPOR eBulletin October 2011. Collectively, there were 11 written comments. All written comments are published at the ISPOR Web site. All comments (many of which are substantive and constructive) were considered, and once consensus was reached by all authors of the article, the final report was submitted to *Value in Health*.

Definitions

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

Observational Study: A study in which participants are not randomized or otherwise preassigned 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, often longitudinal in nature, for which the consequential outcomes of interest occur *after* study commencement (including creation of a study protocol and analysis plan, and study initiation). Patient exposure to any of the interventions being studied may 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, and decision made to treat. 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 and retrospective observational studies undertaken for consideration by health policy decision makers. The ISPOR Task Force for Good Research Practices for Retrospective Databases Analysis completed its work in 2009 with a focus on CER [4–6]. Our report is focused on setting out good practices in the design, conduct, analysis, and reporting of prospective observational studies that will enhance their value to policymakers. A number of issues addressed by the 2009 task force also apply 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 that will require appropriate protocol review by institutional/ethical review boards and provisions must be made to protect confidentiality of patient-identifiable information.

There have been some prior recommendations on the conduct 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 postauthorization safety studies conducted under European Network of Centres for Pharmacoepidemiology and Pharmacovigilence (ENCePP) [7]. The ENCePP has itself also created a checklist of methodological research standards for ENCePP studies [8]. The Agency for Healthcare Research and Quality first 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 published update guidelines for good pharmacoepidemiology practices [10]. The STrengthening the Reporting of OBservational studies in Epidemiology (STROBE) guidelines focus on the reporting of observational studies [11]. In our view, none of the existing recommendations explain or provide general guidance on how to design a strong prospective observational study to address comparative effectiveness or relative effectiveness research questions.

Choice of study design

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 pragmatic clinical trials (PCTs). For the purposes of this report, we define a PCT as a randomized prospective clinical study in which, following randomization to an intervention, patient care is left to the practitioners according to their typical practice. PCTs are intended to maintain the advantages of randomization while examining outcomes in routine care [12]. The definition of what is a PCT, however, falls on a spectrum [13], and we note that PCTs are not recognized as a separate study design from RCTs in all jurisdictions. Clinical trialists and others have questioned whether all effectiveness studies could employ randomization; therefore, we felt it necessary to address the relative merits of PCTs compared with those of prospective observational studies in our report.

Choice of a CER study design follows from the research question, but optimal design must consider issues of the expected value of the information to be generated, clinical equipoise, timing, feasibility, cost, ethics, and legality. If an observational study is to assess comparative effectiveness, the first choice to be made is whether the design should be retrospective or prospective. Retrospective studies are typically performed by using existing data sets and usually offer advantages in terms of cost and speed of execution. The data sets, however, may not contain all the information desired and therefore definition of exposure and outcomes may not be ideal. Prospective studies could also use existing databases, but they also offer the opportunity to collect additional desired information, and so they are usually more costly and take longer to complete. Formal or informal value of information analysis is useful in making this choice.

When developing an observational study design, it is important to consider whether there is clinical equipoise for the treatments of interest and whether the proposed study design and analysis plan will be sufficient to address confounding and bias. Clinical equipoise has been defined by Freedman [14] as existence of "genuine uncertainty within the expert medical community—not necessarily on the part of the individual investigator—about the preferred treatment". Equipoise is defined at both the individual physician/patient level and the population level [15]. 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 frequently 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 confounding and bias. In these situations, a randomized study design may be preferable and one might explore a naturalistic randomized PCT study design.

When the clinical situation does not involve very strong treatment preferences, either a prospective observational study or a PCT can be of value and provide useful information. PCTs offer randomization that may be considered important when the principle study question focuses on the relative contributions of different treatments to the observed outcomes. Randomization in this context provides an additional tool to mitigate concerns about potential confounding. In PCTs, however, randomization is typically not enforced and a significant amount of treatment switching has been observed to occur [16,17] with consequent decrease in the value of randomization.

Accounting for switching by using various statistical approaches that are typically applied in observational studies may enable assessment of the relative contribution of the originally assigned therapies to the observed relative effectiveness. For different policy questions, such as when the focus is on the relative outcomes associated with various treatment strategies (that initiate with alternative treatment options), or the relative outcomes associated with a device versus a drug, observational studies may be preferred because the extra burdens associated with randomization may not be worthwhile. Prospective observational studies are generally less costly than PCTs and may pose fewer feasibility

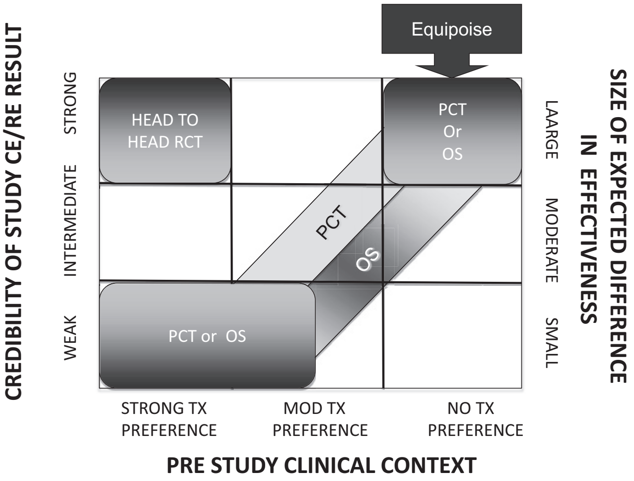

Fig. 1 – Consideration of prestudy clinical context and the size of expected differences in effectiveness are critical to the choice of study design for comparative effectiveness research studies. CE, comparative effectiveness; MOD, moderate; RCT, randomized controlled trial; RE, relative effectiveness; OS, observational study; PCT, pragmatic clinical trial; TX, treatment.

issues (including the 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 PCT may be able to provide useful information because of 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 (Fig. 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 [18]. 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 [19]. Following these results and similar findings from several other RCTs, HDC-ABMT fell out of practice.

Typically, there is a relative paucity of head-to-head RCTs that answer questions of interest to policymakers. On occasion, large multicenter RCTs such as the Antihypertensive and Lipid-lowering Treatment to Prevent Heart Attack Trial [20] are funded in the United States by the National Institutes of Health or by the Veterans Administration, and these provide important insights into the relative outcomes obtained with various treatments. These studies, however, 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 is limited. Observational studies will be an important component of the CER evidence base for the foreseeable future.

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 experience have important impacts on the safety and effectiveness of the product. RCTs are limited in addressing issues of experience because 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 [21]. 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 or device is compared to nonintervention. Other intervention characteristics also make blinding difficult. Evaluation of interventions in different delivery settings, such as in-home as compared with in-hospital, is one such example. 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 because it has a shorter timeline.

Sometimes the study goal may involve questions not answerable by randomized trials—either standard RCTs or PCTs. 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 investigational new drug application). 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.

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 that information from different types of studies can bring (by reducing the uncertainty as to the benefits of using a technology) net of the cost of undertaking different types of studies. 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. The quality of the information coming from the study, however, will also vary, and this needs to be compared with the costs. RCTs typically reduce uncertainty more than do 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 is intended to inform. If these questions are not answered with sufficient rigor, the study will not be considered successful. The findings of the study must be judged sufficiently valid to inform the policy decisions faced by decision makers. 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 concept of validity refers to whether a study is able to accurately answer the questions it is intended to answer. Validity requires that outcome measurements accurately reflect the true situations of interest (treatments, health outcomes, and other characteristics that may influence the likelihood of treatment success) and that the conclusions drawn from the analysis of those measurements are well founded. 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.

Along with assessing validity, policymakers 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. Well-conducted RCTs are generally considered the "gold standard" to establish causal relationships, and it remains a matter of dispute whether one can ever establish casual relationships in observational studies. It is the position of this task force that rigorous well-designed and well-executed observational studies can provide evidence of causal relationships.

Specifying the population, comparators, and outcomes of interest populations

Clear specification of the populations, treatments, and outcomes of interest is critical to making inferences regarding causal relationships [22]. 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 a particular policy or not) or 2) only those who would participate in the policy or only those who receive a treatment (e.g., effect of treatment on the treated). In the first situation, if everyone was exposed, the target population is the whole population, and the causal effect is termed the *average causal effect*. Similarly, in a clinical trial setting, if everyone who was randomized to the experimental treatment complied and everyone who was randomized to the comparator complied, then the causal effect is the average causal effect and the target population is the population of *potential compliers*.

There may be more interest in determining the causal effect among those who would likely receive the treatment (scenario 2). For example, in assessing the effectiveness of oral antidiabetics, a policymaker 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 while researchers are interested in evaluating broader questions of tolerability, adherence, and effectiveness. Potential treatment takers may include treatment-naive patients and not just those who have been observed to use the treatment. Similarly, policymakers may be interested in the effect of participation for those who actually participate in a policy. Subjects who are potential treatment takers are likely to have specific characteristics and needs that would make them different from a randomly selected subject.

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.

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

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 with the average effect of the treatment [23,24]. Reasons for different treatment effects may arise because of biological reasons, patient or provider preferences, or values.

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 [22]. 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. As the number of treatments increases, however, determination of the causal effects on the target population becomes more complicated. For example, suppose a researcher is interested in studying five drugs. In an RCT in which all patients are eligible for each drug (therefore, the target population is everyone), all possible pairwise 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 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 treatments currently on the market. A simple study of outcomes among people treated with the new drug compared with 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 that 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 pairwise 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 on patient response to surveys or questionnaires. For the latter, validation of these outcomes should be sought by using other means for key clinical and economic end points if the survey instruments have not been independently validated. Prospective observational studies also provide greater potential to examine a broader range of clinically relevant outcomes (e.g., disease flares and readmissions) 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

Table 2 – Sample designs for prospective observational comparative effectiveness research studies.

Design	Definition	Advantages/disadvantages
Single group, pretest/posttest design (also called interrupted time series)	Outcomes collected before and after an exposure for a single group (longitudinal)	Subjects serve as own control Secular trends confounded with introduction of exposure (subjects or conditions may change over time naturally) *Vulnerable to time-invariant and time-varying unmeasured confounding and to confounding treatment effect with natural history*
Multiple group, cross-sectional cohort design	Subjects defined as having a particular condition who underwent one of multiple different interventions/exposures (cross-sectional)	*Vulnerable to time-invariant and time-varying unmeasured confounding* *Vulnerable to regression to the mean*
Multiple group, pretest/posttest designs (also called quasi-experimental, difference-in-difference design; regression discontinuity design)	Outcomes collected before and after an exposure or intervention for at least two groups, e.g., an "intervention" group and a "comparison" group (longitudinal)	Subjects serve as own control Disentangle secular trends from intervention or exposure Robust to time-invariant unmeasured confounding *Vulnerable to time-varying unmeasured confounding*
Prospective cohort studies	Outcomes collected after an exposure, diagnosis, or intervention, ideally for at least two groups—an intervention group and a comparison group	Broad enrollment criteria may enhance ability to evaluate treatment heterogeneity Disentangle secular trends from intervention or exposure *Vulnerable to time-varying unmeasured confounding*

decisions despite its importance. If a study is intended to provide sufficiently robust 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 [25]. 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 [26] frequently focus on designs to assess the impact of an exposure (most often a drug, medical device, or intervention) or an outcome (often an adverse event). For CER, the task force focused on two broad issues— cross-sectional versus longitudinal designs [27]—and considered only those study designs that utilized at least one comparison group. The comparison group could be composed of different subjects than those receiving the intervention or the comparison group could consist of the same subjects receiving the intervention measured before receiving the intervention (Table 2).

Single group, pretest/posttest designs

These designs are longitudinal studies in a single group of subjects. The pretest period is defined as the time before the intervention or exposure and the posttest period as the time after the exposure. Subjects serve as their own control in that outcomes observed in the posttest period are subtracted from the outcomes observed in the pretest period for each subject. In the setting of a single preintervention outcome measurement and a single postintervention 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 or her own control so that unmeasured time-invariant factors are differenced out. Disadvantages of single-group pretest/posttest designs involve the lack of a comparison group and the inability to control for unmeasured time-varying confounding. The absence of a comparison group results in 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 pretest and the posttest periods, then this design has been referred to as an *interrupted time series design* [28]. 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 pretest/posttest design because of their ability to minimize regression to the mean through collection of repeated measurements. Interrupted time series designs, however, remain vulnerable to time-varying confounding and to confounding natural history with 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. A causal effect, however, is difficult to establish unless the investigator has some assurance that the exposure preceded the outcome and the various treatment groups are comparable along all the dimensions except the one under study. The lack of longitudinal data makes this design vulnerable to regression-to-the-mean issues. Moreover, unmeasured differences between treatment groups may confound the observed treatment effects.

Multiple group, pretest/posttest 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 with the average change from baseline 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 or her own control so that unmeasured subject-specific time-invariant confounders are eliminated.

Single or multiple group prospective cohort designs

Like pretest/posttest 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, for example, time-invariant residual confounding [29]. These designs, however, are vulnerable to time-varying unmeasured confounding as well as systematic unmeasured differences in subjects belonging to different groups.

Time-varying and time-invariant confounding

When an outcome and treatment/exposure are both influenced by a third variable and when the distribution of the third variable is different across treatment or exposure groups, that variable is called a confounder. Time-varying confounding refers to the setting in which the outcome and treatment/exposure are influenced by new values of a third variable [30]. Disease severity changes over time, influences the decision to initiate therapy, and relates to outcome. Time-invariant confounding refers to a confounder that does not change values over time. For example, a patient's socioeconomic status may be related to treatment selection and functional status, and, assuming a short observational period, does not change over 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 the 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 the practice setting, the health-care environment (e.g., single-payer system and fee for service), the experience of health-care providers, as well as the medical history of patients (i.e., inception cohort vs. chronic users). As an example, it is frequently the case that the clinicians have strong preferences that may not be based on evidence but related to their training, health care system requirements, or individual economic considerations. Researchers may need to conduct some preliminary investigations to understand these preferences. Another example is that 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. A third example is that surgeons trained to use newer and less invasive surgery may apply this technique in their practice while others are comfortable only with the older procedure. These situations result in variations in care that can be quite significant [31]. Identifying and collecting data on 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 [32]. When prevalent users are included in a study, it is important to recognize that 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 [33]. Tools that are often used include inception cohorts that focus on people with newly diagnosed disease, or may start when patients first require treatment with medication (e.g., first oral antidiabetic [new users]). Incident and new user designs [32] facilitate drug-drug comparisons for people who are initiating a pharmacotherapy, combination medications, new classes of drugs, and so on. This is illustrated by the report of the re-examination of the Woman's Health Initiative when observational hormonal replacement treatment data were analyzed only for treatment initiators; in this analysis, the results were inferentially the same as for the Woman's Health Initiative randomized trial data, which enrolled treatment-naive patients [34]. Earlier analyses of the Woman's Health Initiative observational data included both currently treated patients and treatment initiators, causing 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 [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 of their infirmity and comorbidities. Differences may be so extensive that physicians choose not to prescribe seemingly reasonable treatments for fear of complications or unknown drug interactions [35]. 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 the 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 preplanned 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 [36] examined the effectiveness of drug-eluting coronary stents compared with bare metal stents by using mortality 2 days from stent implant as a control outcome. Mortality differences between those implanted with a drug-eluting coronary stent and those implanted with a bare metal stent at 2 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 that those who participate in a registry are different from those who do not in terms of compliance and lifestyle. If

Table 3 – Potential solutions for problems frequently encountered in the design of prospective observational studies.

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

some relevant outcomes data are available for the nonparticipants, then both cohorts who received drug B could serve as comparison groups for drug A. If a similar effect of treatment A compared with treatment B is observed by using either comparison group, then concern about unmeasured bias due to compliance or lifestyle is reduced. Yoon et al. [37] used multiple control groups to examine the effectiveness of the implementation of the Mental Health Parity Act.

Design Tools

Several frequently encountered design problems and potential design solutions are shown in Table 3.

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 including 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 to maximize their effective use, or to limit their budget impact [38].

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 United Kingdom, a single-arm prospective cohort observational study was conducted. For both 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 with relevant alternatives. The multiple sclerosis scheme has been highly controversial in part because of difficulties with the analysis of effectiveness [39–41].

It is the view of this task force that these studies should, where possible, be designed following good practice recommendations for CER. Ideally, studies should either be conducted as prospective observational studies using appropriate concurrent comparator groups or be conducted as PCTs. For the former, patients who do not receive the therapy under investigation would need to be enrolled. 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 for comparison with the model-based estimate 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 PCT may also be 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 an appropriate alternative to collect evidence on 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 task force in a reported titled "Transferability of Economic Evaluations across Jurisdictions: ISPOR Good Research Practices Task Force Report" [42]. We also note that the ISPOR Performance Based Risk Sharing Task Force is preparing an article on Good Research Practices 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 [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 (in one of its three articles [6]) also recommended general analytic techniques and specific best practices including the 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. We endorse these recommendations, and they will not be further discussed in this report.

One type of bias not discussed by the Retrospective Database Task Force 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 [43]. 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 be treated only when 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 [44]. 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 [45], as well as other documents intended to help readers evaluate good quality in observational studies of comparative effectiveness [46].

Addressing treatment heterogeneity in the analytic plan

Most RCTs are underpowered to test for treatment heterogeneity given such designs effectively require 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 cotreatments).

Intention-to-treat compared with other analyses

The impact of longer follow-up periods with higher dropout and treatment nonadherence, and switching depends on the primary planned treatment effect analysis. Intent-to-treat analyses in PCTs estimate the treatment effects that would occur in populations with similar levels of treatment nonadherence and switching and are considered useful for policy decisions. 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. Other approaches, such as the "adequate" and "completer" methods, classify subjects according to randomization but analyze data only for those who receive an adequate or full amount of the treatment, respectively. All approaches nonetheless remain vulnerable to unmeasured confounding [47] depending on the degree of treatment adherence and of follow-up.

More rigorous approaches based on instrumental variable methodology make use of randomization in both RCTs and PCTs to protect as-treated analyses against unmeasured confounding. Specifically, using the randomization assignment as the instrumental variable yields a valid estimator of treatment effectiveness [48]. In this way, PCTs may be superior to observational studies. This protective property of randomization is reduced, however, with longer follow-up and larger magnitudes of treatment nonadherence, such as treatment discontinuation and switching, and study dropout. These reduce the advantage of PCTs over prospective observational studies in practice. Stated another way, randomization is most protective against residual confounding with intent-to-treat analyses, but it is less useful for as-treated analyses, or when there is an expectation of significant treatment nonadherence. When a prospective observational study is focused on comparing *initial* treatment decisions by 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 [34]. Such an analytic approach played a role in resolving the conflicting hormonal replacement treatment evidence from observational and randomized trials. Absence of treatment heterogeneity and no residual confounding are assumed in these situations unless an instrumental variable that satisfies the necessary assumptions can be identified [47].

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 because the distinction between the safety and effectiveness end points is artificial—at its simplest, an adverse event is negative effectiveness. The efficacy comparisons between treatments, however, can entail either noninferiority or superiority hypotheses (e.g., is drug A equivalent or better than drug B in reducing the blood pressure). In contrast, the safety assessments are much less likely to have superiority hypotheses, particularly when the safety end point is relatively rare. The infrequency of events results in larger sample size requirements for the assessment of safety and for assessing superiority as compared with noninferiority. In addition, 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 because enrollment of large numbers of patients without scientific justification undermines the study's scientific credibility and may suggest that study sponsors have primarily a nonscientific agenda, such as building experience with a new product. There is debate in general, both for observational and for randomized studies, about whether effect sizes should come from pilot study data, the literature, or a general consensus about clinically significant effect sizes [49]. Estimating effect sizes requires large enough pilot sample sizes to ensure adequate precision. Obtaining effect sizes from the 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). Design of PCTs must account for the expectation of treatment switching and therefore may require larger sample sizes to provide adequate power to understand the specific contribution of the assigned treatments to outcomes. Expectation of treatment switching is integral to sample size calculations for prospective observational studies.

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 dropout rates, treatment nonadherence 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 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 takes 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 [50]. It is important to recognize, however, that even small observational CER studies may provide important insights, recognizing the limitations described above.

Study Execution

Given the aim of CER is to inform health policy decisions, the recommendations of both the ISPOR Retrospective Database Task Force and GRACE (Good ReseAch for Comparative Effectiveness) [46] include a requirement for a formal written protocol to 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 apply to prospective observational studies.

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 and 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 and 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 institutional review board 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 to the ability not only to a priori implement common definitions of covariates and outcomes but also to collect potential confounders, control outcomes, and additional comparison groups. Like an RCT, specifics of the study inclusion and exclusion criteria (again, before collection of outcomes) should be listed. Because subjects will not be randomized to treatments, a list of possible confounders should be composed and compared with 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 callbacks) 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 [51]. 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 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 postmarketing 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 a larger sample size is needed 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., SDs 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 fieldwork are the distinguishing characteristics of successful studies. In observational studies, the challenges are greater than in RCTs, 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 in which safety may not be the main objective. Study samples may be selectively enriched to include the exposures of interest through strategies where enrollment in comparators may be limited so as to be sure that sufficient numbers of patients with the treatment(s) of interest are recruited for study. Researchers are further challenged by practical issues including 1) limiting the length of questionnaires to match the time and patience of physicians and patients who rarely receive much, if any, compensation for their cooperation and 2) managing to collect the data of interest within regularly scheduled encounters with health-care providers. Other challenges include addressing issues of missing data and data quality, because prospective observational studies have traditionally channeled any additional resources into recruiting larger numbers of patients and into longer follow-up, rather than into source data verification, which is a rarity for most of these studies. Other questions of interest related to reporting and publication include whether observational studies should be registered and how to work with coauthors on publications such that they ethically can attest that they have participated in data analysis and whether this requires providing all coauthors with copies of the data sets, and if and when data access should be granted to nonparticipating 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 ongoing guidance about operational issues that may influence validity. For example, sequential enrollment plans may not always be feasible and investigators may propose alternatives that require scrutiny in terms of protection from bias. The most successful advisory boards have written charters that lay out their roles and responsibilities, including rules for voting at formal meetings and in special situations that may arise between meetings.

Incentives for participation

Institutional/ethical review boards 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, and gynecological examinations). 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. Incomplete reporting, however, 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. This payment strategy provides a more proximate incentive to complete case report forms and may be more effective at ensuring data collection.

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 the development of a registry of patient registries in the United States; recommendations are under development. In Europe, the ENcEPP recently created a voluntary registry for observational pharmacoepidemiology and pharmacovigilance studies, and it 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 CER, 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 article. "Full access" to study data, however, is not defined. In major RCTs, some funding agencies and journal editors have required that authors include a statement in the published article that readers may contact the first author to request access to the full study data set (e.g., Hull et al. [52]). While the rationale for the Food and Drug Administration requirement that a pharmaceutical company supply the raw data used in its analysis of clinical trials to Food and Drug Administration statisticians in support of a new drug application is compelling [53], 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 patients' 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., The Global Registry of Acute Coronary Events [54], Coronary Artery Risk Development in Young Adults [55], The Global Longitudinal Study of Osteoporosis in Women [56], and Can Rapid Risk Stratification of Unstable Angina Patients Suppress Adverse Outcomes With Early Implementation of the ACC/AHA Guidelines [57].) A number of limited data sets have been shared with bona fide investigators from outside the scientific advisory boards of these studies [58]. Furthermore, 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 [59]. Also, some journal editors have taken the position that study data should be made available to others, for example, the *British Medical Journal* [60], 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 articles 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. They have received summary data and appropriate statistical findings for all the questions they have proposed in preparing the manuscript for submission to a 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 for both 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 by the health insurance provider that 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 [61].

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 [62].

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 [4–6,9,63]. We also note that Consolidated Standards of Reporting Trials can be useful for the reporting of nonrandomized studies [64] while the STROBE guidelines [11] and the Meta-analysis Of Observational Studies in Epidemiology guidelines [65] are designed specifically for observational studies. Reporting should be consistent in a way that enables it to be used in meta-analysis [66]. To this end, some journals have required that a "Reproducible Research Statement" be included in the article [67]. Thus, specific reporting features should include the following:

- 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 as follows:

- 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 subpopulation 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 [4,68]).

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 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 the following:

- 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 PCTs. 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 and value of answering the question vs timeliness and cost), and any specific considerations with respect to the technology (e.g., devices and 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. The 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 the 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 nondifferential follow-up.
- We believe that similar to the concept of the importance of declaring a prespecified hypothesis, 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 and encepp.eu) in advance of their execution.

Acknowledgments

The task force thanks Rebecca Corey for her indefatigable efforts in providing support for this effort. Tom Ten Have was a member of the task force and made important contributions. Tom passed away before the task force report was completed. He will be greatly missed.

Source of financial support: This task force was supported by ISPOR. The views expressed in this article are Dr. Normand's and do not necessarily reflect those of the Patient Center Outcomes Research Institute.

REFERENCES

[1] U.S. Department of Health and Human Services, Federal Coordinating Council for Comparative Effectiveness Research. Report to the President and the Congress, 2009. Available from: www.hhs.gov/recovery/programs/cer/cerannualrpt.pdf. [Accessed September 6, 2011].

[2] European Union Pharmaceutical Forum, High Level Pharmaceutical Forum 2005–2008. Final report, 2008. Available from:

http://ec.europa.eu/pharmaforum/docs/ev_20081002_frep_en.pdf. [Accessed September 7, 2011].

[3] Institute of Medicine of the National Academies. Medical Devices and the Public's Health: The FDA 510(k) Clearance Process at 35 Years, Committee on the Public Health Effectiveness of the FDA 510(k) Clearance Process, Board on Population Health and Public Health Practice, Washington, D.C.: Institute of Medicine; 2011.

[4] 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.

[5] 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 non-randomized studies of treatment effects using secondary data sources: the ISPOR Good Research Practices for Retrospective Database Analysis Task Force Report–part II. Value Health 2009;12:1053–61.

[6] 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.

[7] European Medicines Agency. The ENCePP Code of Conduct for Independence and Transparency in the Conduct of Pharmacoepidemiological and Pharmacovigilance. May 7, 2010. Available from: http://www.encepp.eu/documents/encepp_studies/ENCePP%20Code%20of%20Conduct_20100507.pdf. [Accessed January 3, 2012].

[8] European Medicines Agency. Checklist for Study Protocols (Revision 1); adopted by ENCePP Steering Group (Doc. Ref. EMEA/540136/2009). August 19, 2011.

[9] Gliklich RE, Dreyer NA, eds. Registries for Evaluating Patient Outcomes: A User's Guide. Prepared by Outcome DEcIDE Center [Outcome Sciences, Inc. dba Outcome] under Contract No. HHSA29020050035I TO1. AHRQ Publication No. 07-EHC001-1. Rockville, MD: Agency for Healthcare Research and Quality, 2010.

[10] International Society for Pharmacoepidemiology. ISPE guidelines for good pharmacoepidemiology practices (GPP). Pharmacoepidemiol Drug Saf 2008;17:200–8.

[11] von Elm E, Altman DG, Egger M, et al. STROBE Initiative guidelines for reporting observational studies in epidemiology (STROBE) statement. BMJ 2007;335:806–8.

[12] Zwarenstein M, Treweek S, Gagnier JJ, et al. Improving the reporting of pragmatic trials: an extension of the CONSORT statement. BMJ 2008; 337:a2390.

[13] Thorpe KE, Zwarenstein M, Oxman AD, et al. A pragmatic-explanatory continuum indicator summary (PRECIS): a tool to help trial designers. J Clin Epidemiol 2009;62:464–75.

[14] Freedman B. Equipoise and the ethics of clinical research. N Engl J Med 1987;317:141–5.

[15] 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. Gen Hosp Psychiatry 2003;25:115–23.

[16] Buesching D, Luce B, Berger ML. A review of industry-sponsored pragmatic trials highlight methodological and policy challenges to becoming more common. J Comparative Effectiveness. In Press, 2012.

[17] Price D, Musgrave SD, Shepstone L, et al. Leukotriene antagonists as first-line or add-on asthma-controller therapy. N Engl J Med 2011;364: 1695–707.

[18] Mello MM, Brennan TA. The controversy over high-dose chemotherapy with autologous bone marrow transplant for breast cancer. Health Aff (Millwood) 2001;20:101–17.

[19] Stadtmauer EA, O'Neill A, Goldstein LJ, et al. Conventional-dose chemotherapy compared with high-dose chemotherapy plus autologous hematopoietic stem-cell transplantation for metastatic breast cancer. Philadelphia Bone Marrow Transplant Group. N Engl J Med 2000;342:1069–76.

[20] ALLHAT Offices and Coordinators for the ALLHAT Collaborative Research Group. Major outcomes in high-risk hypertensive patients randomized to angiotensin-converting enzyme inhibitor or calcium channel blocker vs. diuretic: the Antihypertensive and Lipid-lowering Treatment to Prevent Heart Attack Trial (ALLHAT). JAMA 2002;23: 2981–97.

[21] Sedrakyan A, Marinac-Dabic D, Normand S-LT, et al. A framework for evidence evaluation and methodological issues in implantable device studies. Med Care 2010;48(6, Suppl.):S121–8.

[22] Didelez V, Meng S, Sheehan NA. Assumptions of IV methods for observational epidemiology. Statist Sci 2010;25:22–40.

[23] Kraemer HC, Frank E, Kupfer DJ. Moderators of treatment outcomes: clinical, research, and policy importance. JAMA 2006;296:1286–9.

[24] Kravitz RL, Duan N, Braslow J. Evidence-based medicine, heterogeneity of treatment effects, and the trouble with averages. Milbank Q 2004;82:661–87.

[25] Rubin DB. The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials. Stat Med 2007;26:20–36.

[26] Rothman KJ, Greenland S, Lash T. Modern Epidemiology (3rd ed.). Philadelphia, PA: Lippincott Williams & Wilkins, 2008.

[27] Sadish WR, Cook TC, Campbell TD. Experimental and Quasi-Experimental Designs for Generalized Causal Inferences. Berkeley, CA: Houghton-Mifflin, 2001.

[28] Hartmann DP, Gottman JM, Jones RR, et al. Interrupted time-series analysis and its application to behavioral data. J Appl Behav Anal 1980;13:543–9.

[29] Wooldridge JM. Econometric Analysis of Cross Section and Panel Data. Cambridge, MA: The MIT Press, 2002.

[30] Samore MH, Shen S, Greene T, et al. A simulation-based evaluation of methods to estimate the impact of an adverse event on hospital length of stay. Med Care 2007;45:S108–5.

[31] Sedrakyan A, van der Meulen J, Lewsey J, et al., Variation in use of video assisted thoracic surgery in the United Kingdom. BMJ 2004;329: 1011–2.

[32] Ray WA. Evaluating medication effects outside of clinical trials: new-user designs. Am J Epidemiol 2003;158:915–20.

[33] Weiss NS. The new world of data linkages in clinical epidemiology: are we being brave or foolhardy? Epidemiology 2011;22:292–4.

[34] 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] Sturmer T, Glynn RJ, Rothman KJ, et al. Adjustments for unmeasured confounders in pharmacoepidemiologic database studies using external information. Med Care 2007;45:S158–65.

[36] Mauri L, Silbaugh TS, Wolf RE, et al. Long-term clinical outcomes after drug-eluting and bare-metal stenting in Massachusetts. Circulation 2008;118:1817–27.

[37] Yoon FA, Huskamp HA, Busch AB, Normand S-LT. Using multiple control groups and matching to address unobserved biases in comparative effectiveness research: an observational study of the effectiveness of mental health parity. Stat Biosci 2011;3:63–78.

[38] Klemp 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.

[39] Lilford RJ. MS risk sharing scheme: response from chair of scientific advisory committee. BMJ 2010;341:c3590.

[40] Williamson S. Patient access schemes for high-cost cancer medicines. Lancet Oncol 2010;11:111–2.

[41] Wlodarczyk JH, Cleland LG, Keogh AM, et al. Public funding of bosentan for the treatment of pulmonary artery hypertension in Australia: cost effectiveness and risk sharing. Pharmacoeconomics 2006;24:903–15.

[42] Drummond MF, Barbieri M, Cook J, et al. Transferability of economic evaluations across jurisdictions: ISPOR Good Research Practices Task Force Report. Value Health 2009;12:409–18.

[43] Suissa S. Immortal time bias in pharmaco-epidemiology. Am J Epidemiol 2008;167:492–9.

[44] Dafni U. Landmark analysis at the 25-year landmark point. Circ Cardiovasc Qual Outcomes 2011;4(3):363–71.

[45] European Network of Centres for Pharmacoepidemiology and Pharmacovigilance. Guide on methodological standards in pharmacoepidemiology. May 13, 2011. Available from: http://www.encepp.eu/public_consultation/index.html. [Accessed September 5, 2011].

[46] Dreyer NA, Schneeweiss S, McNeil B, et al; on behalf of the GRACE Initiative. GRACE principles: recognizing high-quality observational studies of comparative effectiveness. Am J Manag Care 2010;16:467–71.

[47] Small D. Sensitivity analysis for instrumental variables regression with overidentifying restrictions. J Am Stat Assoc 2006;102:1049–58.

[48] Kraemer HC, Mintz J, Noda A, et al. Caution regarding the use of pilot studies to guide power calculations for study proposals. Arch Gen Psychiatry 2006;63:484–9.

[49] Horvitz-Lennon M, O'Malley AJ, Frank RG, Normand S-LT. Improving traditional intention-to-treat analyses: a new approach. Psychol Med 2005;35:961–70.

[50] Stuart EA. Matching methods for causal inference: a review and a look forward. Stat Sci 2010;25:1–21.

[51] Byar D. Identification of prognostic factors in cancer clinical trials: Methods and practice. In: Buyse M, Staquet M, Sylvester R, eds. Cancer clinical trials - methods and practice. Oxford: Oxford University Press, 1984.

[52] Hull RD, Schellong SM, Tapson VF, et al. Extended-duration venous thromboembolism prophylaxis in acutely ill medical patients with recently reduced mobility: a randomized trial. Ann Intern Med 2010; 153:8–18.

[53] US Food and Drug Administration. New drug application (NDA) 2010. Available from: http://www.fda.gov/Drugs/DevelopmentApprovalProcess/HowDrugsareDevelopedandApproved/ApprovalApplications/NewDrugApplicationNDA/default.htm. [Accessed September 7, 2011].

[54] 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.

[55] Gidding SS, Carnethon MR, Daniels S, et al. Low cardiovascular risk is associated with favorable left ventricular mass, left ventricular relative wall thickness, and left atrial size: the CARDIA study. J Am Soc Echocardiogr 2010;23:816–22.

[56] Hooven FH, Adachi JD, Adami S, et al. The Global Longitudinal Study of Osteoporosis in Women (GLOW): rationale and study design. Osteoporos Int 2009;20:1107–16.

[57] Mudrick DW, Chen AY, Roe MT, et al. 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;160:1072–8.

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

[59] U.S. Department of Health and Human Services. Data sharing policy and implementation guidance, 2003. Available from: http://grants.nih.gov/grants/policy/data_sharing/data_sharing_guidance.htm. [Accessed September 7, 2011].

[60] *British Medical Journal*. Resources for authors: research [criteria for published articles], 2011. Available from: http://resources.bmj.com/bmj/authors/types-of-article/research. [Accessed September 7, 2011].

[61] Adisasmito W, Chan PKS, Lee N, et al. Effectiveness of antiviral treatment in human influenza H5N1 infections: analysis from a global patient registry. J Infect Dis 202:1154–60.

[62] Dreyer NA, Sheth N, Trontell A, et al. Good practices for handling adverse events detected through patient registries. Drug Inf J 2008;42:421–8.

[63] 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.

[64] Andrews EB, Arellano FM, Avorn J, et al. Guidelines for good pharmacoepidemiology practices (GPP). Pharmacoepidemiol Drug Saf 2008;17(2):200–8.

[65] 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.

[66] Laine C, Goodman SN, Griswold ME, et al. Reproducible research: moving toward research the public can really trust. Ann Intern Med 2007;146:450–3.

[67] Stroup DF, Berlin JA, Morton SC, et al. Consensus statement. JAMA 2000;283:2008–12.

[68] 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(Suppl.): S131–42.