Volume 12 • Number 8 • 2009
VALUE IN HEALTH

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

Marc L. Berger, MD,[1] Muhammad Mamdani, PharmD, MA, MPH,[2] David Atkins, MD, MPH,[3] Michael L. Johnson, PhD[4]

[1]Global Health Outcomes, Eli Lilly and Company, Indianapolis, IN, USA; [2]Applied Health Research Centre of the Li Ka Shing Knowledge Institute of St. Michael's Hospital at the University of Toronto, Toronto, ON, Canada; [3]Department of Veterans Affairs, Health Services Research and Development Service, Washington, DC, USA; [4]University of Houston, College of Pharmacy, Department of Clinical Sciences and Administration, Houston, TX, USA; and Senior Scientist, Houston Center for Quality of Care and Utilization Studies, Department of Veteran Affairs, Michael E. DeBakey VA Medical Center, Houston, TX, USA

ABSTRACT

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

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

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

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

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

Background to the Task Force

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

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

Beginning in January 2008, the Task Force conducted monthly teleconferences to develop core assumptions and an outline before preparing a draft report. A face-to-face meeting took place in October 2008 to develop the draft, and three forums took place at the ISPOR meetings to develop consensus for the final draft reports. The draft reports were posted on the ISPOR website in May 2009 and the task forces' reviewer group and ISPOR general membership were invited to submit their comments for a 2-week reviewer period. In total, 38 responses were received. All comments received were posted to the ISPOR website and presented for discussion at the Task Force forum during the ISPOR 12th Annual International Meeting in May 2009. Comments and feedback from the forum and reviewer and membership responses were considered and acknowledged in the final reports. Once consensus was reached, the manuscript was submitted to *Value in Health*.

Introduction

Health insurers, physicians, and patients worldwide need information on the comparative effectiveness and safety of prescription

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

10.1111/j.1524-4733.2009.00600.x

1098-3015/09/1044 1044–1052

drugs in routine care. Although randomized clinical trials (RCTs) are the gold standard to determine a drug's efficacy against placebo, it is well recognized that results of such studies may not accurately reflect effectiveness of therapies delivered in typical practice [1–3]. In addition, clinical decisions usually involve choices among therapies yet sponsors of drug trials have limited motivation to test new drugs against existing therapies [4]. Routinely collected and electronically stored information on health-care utilization in everyday clinical practice has proliferated over the past several decades. Large computerized databases with millions of observations of the use of drugs, biologics, devices, and procedures along with health outcomes may be useful in assessing which treatments are most effective and safe in routine care without long delays and the prohibitive costs of most RCTs.

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

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

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

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

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

Defining the Question

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

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

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

Questions regarding comparative effectiveness are often restructured into operational terms such as "How does this drug

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

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

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

Prospective Specification of Research and Protocol Development

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

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

Relevance and Rationale

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

Specificity

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

Novelty

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

Feasibility

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

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

Once the research question has been defined, a sound study protocol should be developed with this study question in mind [12]. Key components of a study protocol include study background and rationale, research question/objective, study design, study population, data sources and storage where relevant, study timeframe, specific study definitions, one prespecified primary end point, secondary end points, statistical analysis (including

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

Recommendations

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

Selection of Study Design Appropriate to the Study Question

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

Cross-Sectional Designs

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

Cohort Designs

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

Case-Control Designs

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

Case-Crossover Designs

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

Case-Time-Control Designs

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

event and exposure to the treatment of interest in either of the predefined exposure windows and issues of selection bias in comparing case to controls may still be problematic [24].

Interrupted Time Series Designs

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

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

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

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

The Study Question Dictates the Choice of Data Source

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

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

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

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

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

Reporting

Structured Abstract

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

To this end, a standardized approach to reporting of observational studies should be adopted, similar to the CONSORT (Consolidated Standards of Reporting Trials) recommendations [31] which has been modified by STROBE (Strengthening the reporting of observational studies in epidemiology) statement [23]. The CONSORT and STROBE recommendations could be adapted as follows:

SECTION AND TOPIC	ITEM #	DESCRIPTOR
Study Design	1	Description of Study Design
Introduction • Background	2	Scientific background and explanation of rationale including discussion of the suitability of the database employed
Methods • Defining the question • Objectives • Selection of study design • Selection of data source • Definition of treatment cohorts	3	Clearly defined goals of the study with description of specific sub-questions. Description of study design and why it was chosen; Description of strengths and weaknesses of data source and how study groups were identified, including description of critical variables: diagnostic criteria, exposures, and potential confounders
• Measurement of treatment effects • Classification bias	4	Discuss how treatment effects were measured and how classification bias was addressed.
• Measurement of outcomes • Classification bias	5	Discuss how outcomes were measured and how classification bias was addressed.
• Confounding • By indication • Measured vs. unmeasured • Time dependent • Analytic plan to address confounding	6	Discuss the potential for confounding, both measured and unmeasured, and how this was assessed and addressed
Discussion • Internal validity	7	Interpretation of results, taking into account confounding and imprecision of results.
• Generalizability	8	Generalizability (external validity) of the study findings.
• Overall evidence	9	General interpretation of the results in the context of current evidence.

When reporting results, the objective is to provide a complete and transparent record of the conduct of the study. Although there are accepted criteria to assess the quality of randomized controlled clinical trials, there are no widely accepted criteria for comparative effectiveness studies on secondary databases. For these reasons, we believe it is an imperative for researchers to provide a narrative description within the methods section of a manuscript of whether and to what extent the prespecified analytic plan required modification. What did you originally intend to do? If the analytic plan evolved over time—explain what you found that led to the modification of analyses or models. Report—at least in summary terms—the results of the prespecified analytic plan as well as the ultimate study results. If you had to make compromises in the goals of the study, what did you do and why? To the extent possible, provide an estimate of the expected magnitude of potential confounding before study execution, or estimate the level of confounding that would have driven your result back toward the null hypothesis. We acknowledge that journal editors may not allow space for this level detailed reporting; nonetheless, we believe that this will enhance transparency of the research process and could be included in an appendix.

Interpretation of Results

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

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

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

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

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

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

in the system of health-care delivery can have important effects on patient outcomes. Ideally, the investigators should define a priori what would be considered a clinically meaningful difference and interpret their findings in the light of this definition.

Recommendations

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

How Findings Should Be Interpreted in Light of Policy Questions

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

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

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

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

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

At the other end of the spectrum, there are situations where the risks of acting "too late" may look greater than the risk of acting based on imperfect evidence. Many public health interventions are based primarily on observational evidence, in part because of the impracticality of requiring evidence based on randomized studies for interventions such as tobacco restrictions or seat belt laws. Similarly, because the imitations of RCTs for assessing drug safety are well known, most signals about safety

risks are derived from large epidemiologic studies. Responses to safety concerns can fall on a continuum that includes drug or device withdrawal; (e.g., IUDs), restrictions on access (e.g., Accutane), prominent "black box" warnings (Celebrex) to clinical advisories (e.g., Champix/Chantix). These reflect both how strong and consistent the safety signal is, the potential for bias, the risks posed to the public and the consequences of limiting access to a potentially beneficial intervention.

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

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

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

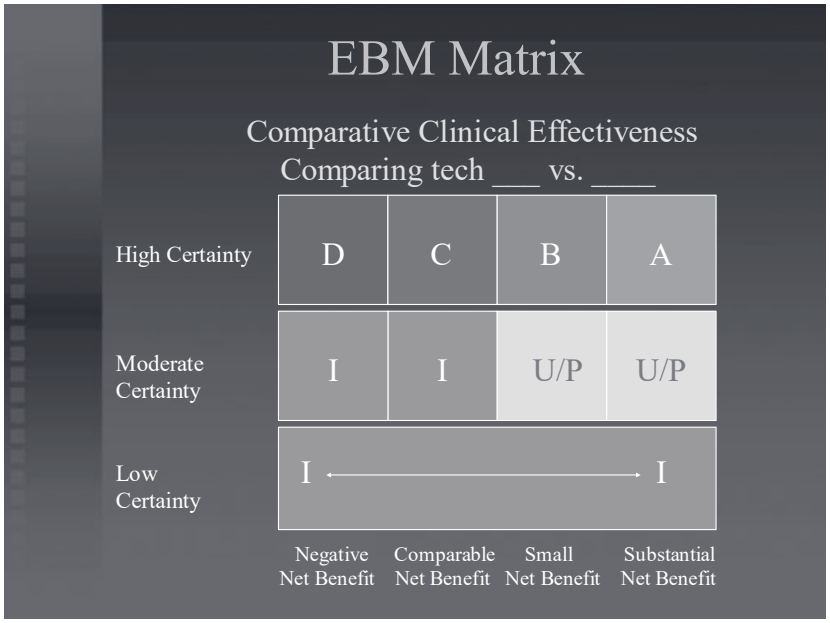

Figure 1 Framework for describing judgments of comparative clinical effectiveness evidence.

A = "Superior"	[High certainty of a substantial comparative net health benefit]
B = "Incremental"	[High certainty of a small comparative net health benefit]
C = "Comparable"	[High certainty of a comparable comparative net health benefit]
D = "Inferior"	[High certainty of a negative comparative net health benefit]
U/P = "Unproven with Potential"	[Moderate certainty of a small or substantial comparative net health benefit] This category is intended to represent bodies of evidence that provide a best estimate in comparative net health benefit as small or substantial but without enough precision to judge which is more likely. The U/P category also implies that there is a relatively small possibility that future evidence would demonstrate that the true net comparative benefit is inferior to other alternatives for many or all patients.
I = "Insufficient"	The evidence does not provide high certainty that the net health benefit of the technology is at least comparable with that provided by the comparator(s).

Conclusion

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

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

References

1 Concato J, Shah N, Horwitz RI. Randomized, controlled trials, observational studies, and the hierarchy of research designs. N Engl J Med 2000;342:1887–92.
2 Concato J. Observational versus experimental studies: what's the evidence for a hierarchy? NeuroRx 2004;1:341–7.
3 Avorn J. In defense of pharmacoepidemiologic studies: embracing the yin and yang of drug research. N Engl J Med 2007;357:2219–21.
4 Schneeweiss S. Developments in post-marketing comparative effectiveness research. Clin Pharmacol Ther 2007;82:143–56.
5 Cox E, Martin B, 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. ISPOR TF Report 2009—Part II. Value Health 2009; doi: 10.1111/j.1524-4733.2009.00601.x.
6 Johnson ML, Crown W, Martin B, et al. Good research practices for comparative effectiveness research: analytic methods to improve causal inference from non-randomized studies of treatment effects using secondary data sources. ISPOR TF Report 2009—Part III. Value Health 2009; doi: 10.1111/j.1524-4733.2009.00602.x.
7 Haynes B. Can it work? Does it work? Is it worth it? BMJ 1999;319:652–3.
8 Teutsch SM, Berger ML, Weinstein MC. Comparative effectiveness: asking the right question. Choosing the right method. Health Affairs 2005;24:128–32.
9 Berger M, Teutsch S. Cost-effectiveness analysis: from science to application. Med Care 2005;43(Suppl.):S49–53.
10 Atkins D. Creating and synthesizing evidence with decision makers in mind. Med Care 2007;45(Suppl.):S16–22.
11 Vandenbroucke JP. Alvan Feinstein and the art of consulting: how to define a research question. J Clin Epidemiol 2002;55:1176–7.
12 Bordage G, Dawson B. Experimental study design and grant writing in eight steps and 28 questions. Med Educ 2003;37:376–85.
13 Vandenbroucke JP. Observational research, randomized trials, and two views of medical science. PLoS Med 2008;5:e67.
14 Austin PC, Mamdani MM, Juurlink DN, Hux JE. Testing multiple statistical hypotheses resulted in spurious associations: a study of astrological signs and health. J Clin Epidemiol 2006;59:964–9.
15 Etminan M, Samii A. Pharmacoepidemiology I: a review of pharmacoepidemiologic study designs. Pharmacotherapy 2004;24:964–9.
16 Etminan M. Pharmacoepidemiology II: the nested case-control study—a novel approach in pharmacoepidemiologic research. Pharmacotherapy 2004;24:1105–9.
17 Mamdani M, Sykora K, Li P, et al. Reader's guide to critical appraisal of cohort studies: 2. Assessing potential for confounding. BMJ 2005;330:960–2.
18 Normand SL, Sykora K, Li P, et al. Readers guide to critical appraisal of cohort studies: 3. Analytical strategies to reduce confounding. BMJ 2005;330:1021–3.
19 Rochon PA, Gurwitz JH, Sykora K, et al. Reader's guide to critical appraisal of cohort studies: 1. Role and design. BMJ 2005;330:895–7.
20 Schneeweiss S, Stürmer T, Maclure M. Case-crossover and case-time-control designs as alternatives in pharmacoepidemiologic research. Pharmacoepidemiol Drug Saf 1997;6(Suppl. 3):S51–9.
21 Schneeweiss S, Maclure M, Walker AM, et al. On the evaluation of drug benefits policy changes with longitudinal claims data: the policy maker's versus the clinician's perspective. Health Policy 2001;55:97–109.
22 von Elm E, Altman DG, Egger M, et al. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. Ann Intern Med 2007;147:573–7.
23 von Elm E, Altman DG, Pocock SJ, et al. for the STROBE Initiative. Strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies. BMJ 2007;335:806–8.
24 Greenland S. Confounding and exposure trends in case-crossover and case-time-control designs. Epidemiology 1996;7:231–9.
25 Mamdani M, McNeely D, Evans G, et al. Impact of a fluoroquinolone restriction policy in an elderly population. Am J Med 2007;120:893–900.
26 Juurlink DN, Mamdani M, Kopp A, et al. Drug-drug interactions among elderly patients hospitalized for drug toxicity. JAMA 2003;289:1652–8.
27 Juurlink DN, Mamdani MM, Lee DS, et al. Rates of hyperkalemia after publication of the Randomized Aldactone Evaluation Study. N Engl J Med 2004;351:543–51.
28 Hennessy S. Use of health care databases in pharmacoepidemiology. Basic Clin Pharmacol Toxicol 2006;98:311–13.
29 Kush RD, Helton E, Rockhold FW, Hardison CD. Electronic health records, medical research, and the Tower of Babel. N Engl J Med 2008;358:1738–40.
30 Tu JV, Bowen J, Chiu M, et al. Effectiveness and safety of drug-eluting stents in Ontario. N Engl J Med 2007;357:1393–402.
31 Moher D, Schulz KF, Altman D for the CONSORT Group. The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomized trials. JAMA 2001;285:1987–91.
32 Lomas J, Culyer T, McCutcheon C, McAuley L. Law for the Canadian Health Services Research Foundation. Conceptualizing and Combining Evidence for Health System Guidance. Canadian Health Services Research Foundation, 2005. (http://www.chsrf.ca/other_documents/pdf/evidence_e.pdf, accessed August 26, 2009)
33 Teutsch S, Berger M. Evidence synthesis and evidence-based decision making: related but distinct processes. Med Decis Making 2005;25:487–9.
34 Park-Wyllie LY, Juurlink DN, Kopp A, et al. Outpatient gatifloxacin therapy and dyglycemia in older adults. NEJM 2006;354:1352–61.
35 Kastelein JJ, Akdim F, Stroes ES, et al. Simvastatin with or without ezetimibe in familial hypercholesterolemia. NEJM 2008;358:1431–43.
36 Guyatt GH, Oxman AD, Vist GE, et al. An emerging consensus on rating quality of evidence and strength of recommendations. BMJ 2008;336:924–6.
37 Sawaya GF, Guirguis-Blake J, LeFevre M, et al. Update on the methods of the U.S. Preventive Services Task Force: estimating certainty and magnitude of net benefit. Ann Intern Med 2007;147:871–5.
38 Health Industry Forum: Comparative Effectiveness Forum Executive Summary. Washington, DC. (Available at http://healthforum.brandeis.edu/meetings/materials/2006-30-Nov./ExecBrief.pdf Accessed August 21, 2009).