Methods for Comparative Effectiveness Research/Patient-Centered Outcomes Research: From Efficacy to Effectiveness

1. Introduction

Effectiveness research that assesses the impact of health interventions on patient outcomes is confronted with many methodological challenges. Although designing valid studies of effectiveness may be more complex than determining efficacy in highly controlled research environments, the hope is that comparative effectiveness research (CER) — if done well — will better inform health care decisions to improve patient outcomes.

The Agency for Healthcare Research and Quality (AHRQ), through its Effective Health Care program, sponsored a fourth symposium on research methods for CER and patient-centered outcomes research (PCOR) in June 2012 to address a selection of issues arising in CER, complementing earlier AHRQ-funded CER Methods Symposia [1–3]. The focus of the 2012 symposium was on original research, methodological insights, or advances arising from the conduct of CER, and how CER can better support health care decision-making. Methodologists and other researchers were invited to present innovations in research methods relevant to CER; specifically, methods that can address some of the underlying differences in the results of randomized efficacy trials and observational effectiveness studies conducted in routine health care settings. Research presented at the symposium was subsequently developed by the authors into manuscripts based on relevant feedback from symposium participants. Authors submitted manuscripts for review by the editorial team and qualified manuscripts were subsequently peer reviewed by external experts. The following sections provide an overview of the papers that were presented and accepted for publication in this supplement.

2. Comparative effectiveness research (CER)/Patient-centered outcomes research (PCOR) implementations

Choudhry and Shrank present design and implementation issues from a novel, pragmatic cluster randomized trial embedded in a commercial insurance claims data system [4]. The study compares the effect of a financial incentive (eliminated copayment vs. usual copayment for cardiovascular medications after myocardial infarction) within an existing payment system. The expectation was that financial incentives will improve medication adherence, yielding improved clinical outcomes. The Post-Myocardial Infarction Free Rx Event and Economic Evaluation (MI FREEE) trial provides a valuable exploration of the challenges and benefits of conducting a pragmatic trial. Despite the limitations imposed by the design, studies conducted in this way are potentially powerful approaches to evaluating health care interventions because research is conducted within an existing health care delivery system. As a result, study results are directly applicable to patients and providers in that system, and may be generalizable to other similar systems. Many operational challenges of the approach are described to help researchers understand common issues encountered when applying this design.

Wu et al. underline the importance of adding patient-reported outcomes to electronic health records and illustrate how straight-forward collection of routine data can provide a rich resource for CER [5]. They advocate incorporating patient-reported outcome measures into widely used electronic health record systems and using these measures for both clinical care and effectiveness research. Electronic health record add-on programs can be used with computerized patient-reported outcome measures, including Patient Reported Outcomes Measurement Information System (PROMIS) computer adaptive tests developed by the National Institutes of Health (NIH). Three converging trends are identified and a framework offered for integrating them: greater patient centeredness and adaptability of patient-reported outcomes; personalization of electronic health records; and patient-oriented CER. Given the current national emphasis on patient-centeredness, the authors argue that this is an opportune time to align the many stakeholders in health care to gather support for using electronic health records as a major CER data source.

N-of-1 trials are seen as the ultimate personalization of CER because they offer immediate implementation of trial findings into the process of delivering care. Duan et al. summarize the conduct and value of N-of-1 trials for clinicians, researchers, and health care organizations [6]. The authors discuss how such trials help clarify the roles of individual and collective clinical experience in medical decision-making. These trials become important weapons in the CER and PCOR methodological armamentarium and effective clinical tools to inform personalized treatment...
3. Comparative effectiveness research (CER) challenges in mental health

CER is particularly challenging but rewarding in mental health. Blanco et al. highlight the challenges in CER for psychiatric treatments [7]. Research questions often go beyond simple head-to-head comparisons in the general population, seeking to assess comparative effectiveness in patient populations with comorbid medical conditions, comparing many nuanced non-pharmacological therapies, or evaluating methods to improve access and retention in minority populations. The authors argue that clinical trials should retain their important role in CER in cases of highly prevalent disorders, large expected effect sizes, difficult-to-reach populations, sequential treatments, and stepped-care algorithms. Substantial attention should be paid in choosing appropriately between clinical trials and observational studies because it is important to allocate research resources carefully to inform key treatment decisions. Often a combination of complementary study types is called for.

Kane et al. focus on the comparative effectiveness of long-acting injectable versus oral antipsychotic medications for relapse prevention in schizophrenia, focusing on medication adherence and rehospitalization [8]. This case study summarizes and compares the methodology of different study types and what these studies add to the evidence base. The authors show how the results of long-acting injectable studies differ considerably depending on design, and the variety of questions being asked and answered are carefully dissected here. The authors conclude that the parallel-group randomized controlled trial is not necessarily the gold standard for assessing the effectiveness of long-acting injectables. A large simple trial might be more informative.

4. Studying the effectiveness of health care delivery system changes

Changes in health care delivery systems are themselves subject to effectiveness evaluation, but may also create extra variation in treatment choice that can be used to study the effectiveness of individual products and therapies. Dore et al. utilize a study of erythropoietin stimulating agents in end-stage renal disease patients to compare different methods of effect estimation [9]. They estimate different causal effects of erythropoietin-stimulating agents on 6-month and 1-year survival in end-stage renal disease: a) a "natural experiment" with a policy shift through a switch to bundled payments in end-stage renal disease care, using a “difference-in-difference” metric, b) instrumental variable analysis, as well as c) a propensity score analysis. The authors compare the estimates and, for each analysis, dissect the specific causal contrasts and assumptions on which causal inference rested.

Also in the setting of anemia treatment in end-stage renal disease patients, Ellis and Brookhart examine the challenges of estimating the effect of treatment (iron supplementation) when they are closely correlated with a concurrent treatment (erythropoiesis-stimulating agent) [10]. If only one treatment is of interest, then including the other treatment in the propensity score model as a confounder may yield more stable estimates of the effect of interest. If the joint treatment effect is of interest, extreme inverse probability of treatment weights may need to be addressed by restricting the sample, exploring limited treatment plans, or through other means. This paper demonstrates the value of presenting both relative and absolute effect measures in CER, particularly when studying multiple outcomes with varying incidence rates. It also provides a summary of analytic methods used in comprehensive evaluation of safety and effectiveness in terms of acute, short-term, and long-term events.

Branas et al. report an effectiveness-maximizing approach for geographically locating emergency care resources [11]. The authors use simulations of response time as a proxy for trauma outcomes. The location of trauma centers and helicopter depots affects how rapidly severely injured patients can access trauma center care, which has implications for survival. These methods may be useful in other settings, such as planning for other emergency services, trauma care, or the location of new hospitals.

5. Analytic issues in comparative effectiveness research (CER)

Secondary data describing utilization and outcomes in routine care without perturbing the system may impose limits on many design aspects that investigators seek to control (e.g., standardized measurements). However, they also give rise to unanticipated analytic opportunities that can be exploited for CER.

The paper by Walker on the deleterious effect of match- ing on provider in CER examines an often overlooked difference between randomized trials and observational research—the value of matching at the level of the provider (or block randomization in trials) [12]. Determinants of treatment in observational studies are either instruments or confounders. Control for instruments amplifies the bias induced by unmeasured confounders. The following variables are often considered instruments or near-instruments: physician, practice group, hospital, calendar time, formulary, and other administrative constraints.
Finely stratifying a propensity score by such variables may increase rather than decrease bias from unmeasured confounding. This is not the case in block-randomization, whose purpose is to reduce bias by standardizing measurements within a center in addition to the randomization. Although similar in appearance (non-randomized provider stratification vs. block randomization), their consequences are quite different.

Brooks et al. demonstrate the use of geographical treatment variation for instrumental variable analyses [13]. Local area practice-style measures can be useful, but results may vary with the size of the local area in terms of both the extent of treatment variation used in estimation and the potential risk of confounding. Brooks defines the “local area practice style (LAPS)” as the instrument. Each patient’s local area is defined by a fixed number of N patients surrounding the patient within a 10-minute drive of the patient’s zip code. Results are sensitive to how many patients are required by the definition, and may be subject to residual confounding (when N is too small) or loss of heterogeneity in the instrument itself (when is N too large).

Variable selection in propensity score analyses needs to include all predictors of outcome for valid estimation. Stuart et al. use prognostic-scores to assess the balance achieved in propensity score analyses in CER [14]. Using a prognostic score estimated among the unexposed to compare the underlying outcome risk across treatment groups in a weighted or matched cohort is described. The more similar the outcome risk, the more likely confounding is controlled. Balance measures based on the expected prognosis under one condition (e.g., control) perform particularly well. CER can use the prognostic score-based balance measure to gauge the success of their propensity score approach.

The paper by Lendle et al. on targeted maximum likelihood estimation (TMLE) introduces this innovative and comprehensive approach, adapted to CER [15]. The paper makes a case for moving beyond the use of simplifying parametric models for estimating causal parameters. The flexibility and efficiency of using super-learner techniques combined with targeted maximum likelihood estimation may be useful, particularly for the analysis of secondary data, where the choice and measurement of covariates and censoring is largely out of the investigator’s control. In addition to the flexibility in estimating treatment choice and outcomes, the approach provides efficient estimation of the effect size. The approach is doubly robust, i.e., even if either treatment or the outcome model (but not both) are misspecified, the effect size is validly estimated. An example study and appropriate references to the literature will allow the reader to explore this exciting and highly promising data-driven approach to CER analyses.

Neugebauer et al. used super-learning to avoid incorrect inference from arbitrary parametric assumptions in marginal structural modeling [16]. Inferences from marginal structural modeling based on inverse probability weighting from electronic health record data are sensitive to parametric decisions for modeling treatment selection and right-censoring mechanisms. Super-learning can effectively harness confounding and selection bias by flexibly bundling multiple existing machine learning algorithms. Erroneous inferences about clinical effectiveness because of arbitrary and incorrect parametric assumptions may be reduced through the use of the super-learning algorithm.

Nelson et al. demonstrate the incorporation of detailed covariate information in a patient subsample in the setting of influenza vaccination effectiveness in reducing mortality [17]. This association is known to be subject to very strong confounding. The authors augmented self-controlled analyses in longitudinal administrative data with detailed information on health state and frailty in a subset of patients. The validation methods to correct for bias caused by unmeasured confounding will somewhat reduce confounding via imputation or inverse probability weighting. Selectively missing vaccination exposure information may explain parts of the residual bias.

6. Design issues in comparative effectiveness research (CER)

Engaging stakeholders including patients in the design of CER may ensure that the research will answer questions relevant to decision-makers. Devine et al. describe how a PCOR infrastructure was built [18]. The project administers patient-reported outcome instruments for research and clinical care and engages patients through advisory groups, thus improving patient-centeredness in their research. The authors apply a recently published conceptual framework for conducting PCOR to their inaugural pragmatic trial in peripheral artery disease to describe the usefulness of their approach.

The paper by Connor et al. on Bayesian Adaptive Trials for CER makes a strong case for the usefulness of adaptive trial designs for more efficiently using study resources, maximizing patient benefit and minimizing patient risk in CER [19]. The example study compares three treatment options in an adaptive trial of anti-epileptic agents. The authors’ innovative approach identifies not just the single superior treatment but also the least effective treatment. At the price of additional planning and closer monitoring of interim findings required by adaptive designs, they will minimize the number of subjects needed to answer this question and get patients on the best treatment as soon as possible.

Time-related biases are an important threat to validity in the analysis of secondary databases. Mi et al. evaluate the impact of immortal person-time and time scale in effectiveness studies of medical devices using implantable cardioverter-defibrillators as an example [20]. The authors compared Mantel-Byar, Landmark, and the exclusion method for handling immortal time bias. The authors also examined the effect of different time scales (time from
birth, time on study, time since heart failure) on the application of the various methods. Immortal time bias can be corrected through design or analysis, but the exclusion method is biased in favor of the treatment group and should be avoided in this setting. Most importantly, comparative effectiveness researchers need to be aware of time-related biases. Mi et al. demonstrate appropriate ways to address immortal time.

Overall, methods for CER and PCOR are in a dynamic phase of development and testing, particularly methods for non-randomized treatment comparisons. As new methodologies begin to penetrate the CER community, including algorithm-based approaches to causal inference (targeted maximum likelihood estimation, Super Learner), long-known principles of bias control may present themselves in new light. In these exciting times it is important to be vigilant for both new opportunities and old fallacies.

Sebastian Schneeweiss a,*, John D. Seeger a, John W. Jackson b, Scott R. Smith b

a Division of Pharmacoepidemiology and Pharmacoeconomics
Brigham and Women’s Hospital
Harvard Medical School
Boston, MA, USA

b Center for Outcomes & Evidence Agency for Healthcare Research and Quality
Boston, MA, USA

*Corresponding author. Division of Pharmacoepidemiology and Pharmacoeconomics
Brigham & Women’s Hospital
Harvard Medical School
1 Brigham Circle, Suite 3030
Boston, MA 02120, USA

E-mail address: Schneeweiss@post.harvard.edu

References


