1-17-2013

Effect Identification in Comparative Effectiveness Research

Michael Oakes
U Minnesota, oakes007@umn.edu

Follow this and additional works at: http://repository.edm-forum.org/egems

Part of the Health Services Research Commons

Effect Identification in Comparative Effectiveness Research

Abstract

The widespread adoption of electronic medical records means there are now vast data resources available for comparative effectiveness research (CER). In concert with conventional randomized controlled trials, CER holds great promise for advancing our understanding of how different therapeutic treatments yield different health outcomes in different settings and with different populations. But in a research culture fixated on estimating correlations and p-values, the threat of misinterpretation of results and improper CER inferences is troubling. Accordingly, this paper aims to shore up the inferential foundations of CER by introducing the fundamentals of effect identification, which is the process of identifying or teasing out empirically defensible causal effects from competing explanations. Three primary requirements of effect identification – positivity, exchangeability, and consistency -- are explained and simple exampled are given. The take home message is that so-called big data from medical records may not yield better or more useful results. Advances will come only when the right question is addressed with the appropriate data and methods.

Disciplines
Health Services Research

eGEMs

Effect Identification in Comparative Effectiveness Research

J. Michael Oakes, PhD[i]

Abstract

The widespread adoption of electronic medical records means there are now vast data resources available for comparative effectiveness research (CER). In concert with conventional randomized controlled trials, CER holds great promise for advancing our understanding of how different therapeutic treatments yield different health outcomes in different settings and with different populations. But in a research culture fixated on estimating correlations and p-values, the threat of misinterpretation of results and improper CER inferences is troubling. Accordingly, this paper aims to shore up the inferential foundations of CER by introducing the fundamentals of effect identification, which is the process of identifying or teasing out empirically defensible causal effects from competing explanations. Three primary requirements of effect identification—positivity, exchangeability, and consistency—are explained and simple exampled are given. The take home message is that so-called big data from medical records may not yield better or more useful results. Advances will come only when the right question is addressed with the appropriate data and methods.

Introduction

Comparative effectiveness research (CER) holds great promise for advancing our understanding of how various therapeutic treatments (e.g., pharmaceuticals, surgeries, suggestions) yield different health outcomes in different settings with different populations. In other words, CER is poised to help us understand what works, where, and for whom. CER, often grounded in randomized clinical trials (RCT) and observational treatment effect studies, may be viewed as part of a larger effort focused on learning health care systems and evidence-based medicine.[1] Whereas the often sanitized conditions of RCTs maximize the internal validity of effect estimates, CER studies typically seek a balance between internal and external validity. This distinction is similar to the differences between efficacy and effectiveness trials; the former aims to maximize the internal validity of a treatment effect estimate in a narrowly defined population with strict adherence to study protocols, and the latter aims to evaluate effects within real-world settings. But CER includes both randomized and observational designs and may involve both prospective and retrospective measurements. Accordingly, CER studies are often conducted with less expense and at a more rapid pace than conventional RCTs and therefore have earned acclaim as more pragmatic than conventional RCTs.[2]

The consolidation of health care systems and increasing use of electronic medical records (EMR) provide CER investigators with access to tens of millions of encounter-level records on millions of persons over extended periods.[1,3,4] In addition, multilevel measures on the characteristics and costs of clinical practices and the sociodemographic characteristics of patients and catchment areas may also be available. While fraught with complications associated with the accuracy of measures, optimal database models, and legal/ethical access to private health information, CER studies—and the unprecedented abundance of information available for the conduct of such studies— offer investigators great promise for new insights and improved population health. But, as with all good things, there are limitations. Among other concerns, the abundance of information means that conventional approaches to scientific inference, all too often grounded in p-value and correlational frameworks, may mislead research consumers (e.g., clinicians and the general public). With millions of records at hand, the threat of misinterpretation of CER investigations is likely to grow.[4]

This paper helps shore up the inferential foundations of CER by explaining the fundamentals of "effect identification." As the words imply, effect identification is the process of identifying empirically defensible causal effects from among competing explanations. Sometimes competing explanations involve, in the jargon of econometrics, an endogenous regressor. But the source of the endogeneity is merely a competing explanation. In any case, some find it useful to conceive of the process of effect identification as similar to the procedures of differential diagnosis employed by physicians. Physicians observe a patient's symptoms and, based on different presentations, rule competing diagnoses in or out. In instances of sufficient differentiation of symptoms, a physician can make a clear diagnosis. In cases of competing possibilities, physicians face an identification problem requiring alternative data (e.g., different tests) to make a

[i]University of Minnesota

clear diagnosis. Ultimately, this paper encourages CER investigators to keep their eye on the prize by asking how well treatments work, under what circumstances, for whom, and compared to what. These questions are at the heart of effect identification.

Effect Identification

Any explanation of effect identification must first place the discussion in the broader landscape of health research and more specifically within the class of research studies that set out to understand the direction and magnitude of a treatment (i.e., an effect) on some defined outcome. For purposes here, treatment may be a drug or procedure that is hypothesized to improve a given health measure, the outcome. Treatment effect studies always involve causal questions. What is the impact of X on Y? If we change X by two units, what is the expected change in Y? And, finally, does X1 or X2 have a bigger impact on Y in target population P at time T? By contrast, through possibly worthy of investigation, the mechanisms by which X effects Y are not necessarily part of a treatment effects study. Further, the correlation (sometimes called association) between X and Y is not the goal of a treatment effects study. Correlations are not necessarily the goal of CER, unless we believe that they summarize the effect of causal relationships. Nonetheless, while we could ask several important questions, treatment effect studies are limited to the causal relationship between X and Y.

It is worth acknowledging that serious scholars have long debated the concept of cause .[5–17] In fact, scholars might never reach consensus on the definition of causality or the methods employed to identify causal relationships. Final judgments about the causal effects of treatments typically depend on the accumulation of evidence from a series of studies.[18] For practical purposes, this article defines a causal effect as the change in an outcome variable that is attributable to an exogenous treatment. The outcome is most simply described as a counterfactual contrast of outcome Y when X is set to different values.[8] Let us imagine an outcome measure Y and call it an (imperfect) measure of body fat, such as the body mass index (BMI). We next imagine a treatment, X, that is thought to prevent the unhealthy accumulation of body fat. The treatment X = 1 might be an exercise program, and the lack of treatment, X = 0, is no exercise program. A given person's counterfactual contrast may be calculated by subtracting the observed value of Y when X = 1 from the observed value of Y when X = 0. If the difference is, say, 4, the causal effect of the treatment, X, on Y is 4. The problem is that, for any given time, a person either did or did not exercise. We cannot observe Y under both levels of X. Indeed, the fundamental problem of causal inference is the problem of missing data. Given that space constraints preclude further elaboration on the issue of missing data, interested readers are encouraged to examine the well-developed literature.[8,10,12,19] The important message is that the counterfactual (sometimes called the potential outcomes) framework illuminates the core aspect of effect identification.

Defining Identification

It is worth beginning with a formal definition. Borrowing from Hsiao, we assume a structured system, S, that produces health as measured by a random variable y.[20] We imagine that S is a complete specification of the probability distribution function of y, P(y). The set of all a priori possible structures, s, is called a model. The identification problem consists of making judgments about health-producing structures, S, given model s and observations y. In less formal and perhaps more useful terms, identification means that an effect is discernible and that no other effect could explain the observed data, even as sample size approaches infinity. Formal or not, the definition contains two critically important elements.

First, it makes clear that if two or more explanations (i.e., models) yield the same results, then the effect associated with any given model is *not* identified. Identification means that one and only one explanation or model explains the data, with no competing explanations for the same results. For example, let us imagine that a Martian observed data on a man looking at a mirror and moving his head. A Martian would have no information about optics or human behavior and therefore would not know whether the movements of the real head were causing the image to move or whether the image was causing the head to move. The observed data would not be causally informative because the effects are not identified: both models explain the same data. Manski named such an identification problem the reflection problem.[21]

Second, the definition makes it clear that identification has nothing to do with statistical inference per se. Sampling distributions, confidence intervals, p-values, and hypothesis testing are not relevant because sample size is assumed to be infinitely large. And standard errors, confidence intervals, and p-values all approach zero with infinite sample sizes. This point is particularly important in CER studies with large databases. Researchers working with large sample sizes may estimate correlations with a particularly small p-value or narrow confidence intervals, but the correlations may have no causal (i.e., clinical) meaning, which is critical to CER investigations.[4] Further, the definition makes clear that the collection of more data (of the same type) will not solve an identification problem. While empirical research requires both identification and statistical estimation, identification precedes estimation because, unlike statistical estimation, it does not depend on sampling distributions. Negative identification implies that statistical inference does not contribute meaningfully to understanding the cause of a particular outcome and is therefore fruitless (Manski 1995).[22] If nothing else, a study riddled with identification problems needs to embrace strong assumptions if it is to make meaningful claims. There is nothing wrong with making strong assumptions, but we must acknowledge that with strong assumptions come strong requirements for empirically defending them.[6,7] Future or auxiliary studies must evaluate the veracity of such assumptions. Absent such support, the original claims, supported by unexamined assumptions, become less useful.

Figures 1 and 2 illustrate the point about statistical inference (based on Manski 1995). Let us imagine that the analytic task is to estimate the value of some outcome variable y conditional on some predictor or causal variable x. Figure 1 graphs means (dots) and 95 percent confidence intervals (lines) estimated at a sample size of N = 100. Examination of the plot shows wide confidence

intervals for points in the region [1–6] and [9–11]. This statistical problem may be solved by increasing the sample size to N = 10,000, as depicted in Figure 2. Increasing the sample size increases the precision of our estimates and is therefore desirable. But we face a more difficult problem if we wish to estimate values of y over the region of x from (6–9). In the absence of observed data, the confidence interval is infinitely wide and thus uninformative. This is an identification problem. Many will be tempted to interpolate the values in the region, but efforts at interpolation involve a model/theory and associated assumptions. Without additional information or defensible modeling assumptions, we cannot identify the values of y. The missing region y could take any value or, in fact, not even be possible given the system under investigation. An increase in sample size will not solve the problem.

Figure 1. Illustrative plot of fictitious means and 95% confidence intervals, N = 100
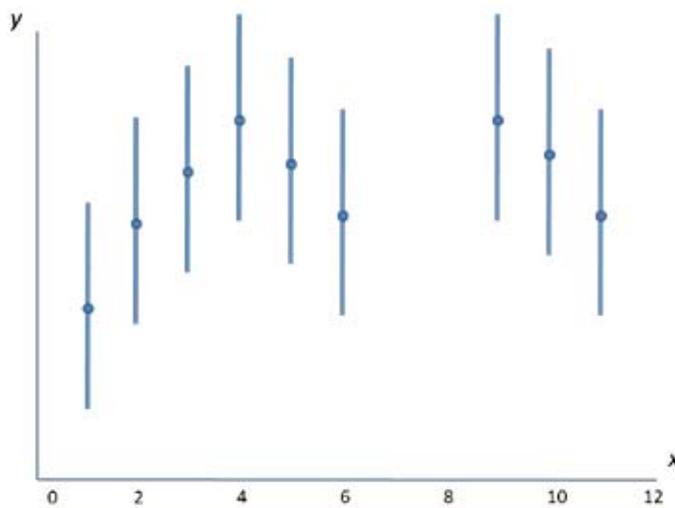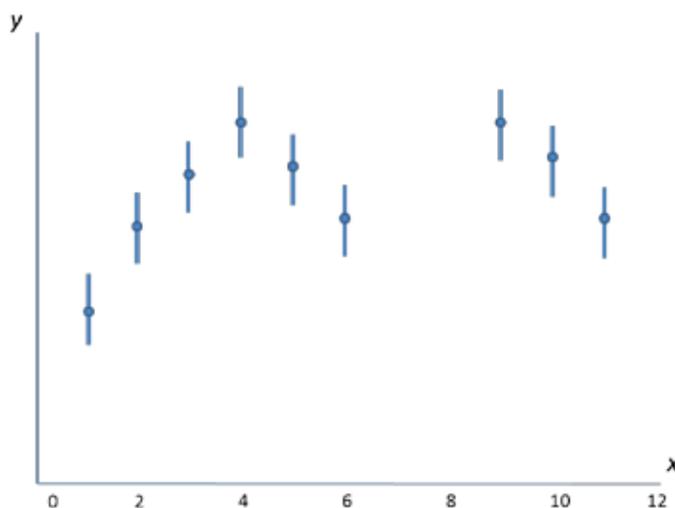

Figure 2. Illustrative plot of fictitious means and 95% confidence intervals, N = 10,000

The ideas behind effect identification are hardly new, but, for reasons scholars have yet to elucidate, correlational frameworks and statistical inference (e.g., sampling distributions, standard errors, and p-values) have dominated modern empirical health research.[19] The work of contemporary scholars such as Greenland, Robins, Freedman, Heckman, Manski, and Angrist has clarified and popularized the ideas and importance of identification.[5,6,13,23–26]

Requirements

Appreciating that different disciplines approach the identification issue from different perspectives, we can set forth three primary requirements of effect identification: positivity, exchangeability, and consistency. If all three requirements are met, an investigator could possibly identify an effect. If any of the requirements goes unmet, however, identification is technically impossible. Of course, in any real-world analysis, assumptions and data are subjected to examination and together provide support (or not) for drawing defensible scientific conclusions.

Positivity means that a subject under investigation has a positive (i.e., non-zero) probability of treatment or exposure[27]—something that may seem obvious. In non-randomized studies, however, some persons may have no chance of treatment or exposure. As discussed below, persons with no theoretical chance of treatment or exposure violate the positivity criterion and are thus literally uninformative to a given analysis. Positivity incorporates the often stringent eligibility and inclusion assumptions of most RCTs.

Exchangeability means that subjects who are compared to one another in a study may be swapped between treatment and control groups without changing the overall value of the estimated treatment effect.[28,29] That is, if subjects actually treated were instead part of the control condition and those in the control condition were actually treated, we should see no change in an estimated effect. In this way, exchangeability is about confounding or, more technically, about the perfect substitution between observed subjects and their counterfactual substitutes. If the case of an imbalance in variables between treatment conditions related to outcomes (e.g., one group is more educated), then effect estimates may be biased or unidentified; swapping the controls for the treated would yield a different answer.

Consistency is the idea that treatments are the same across subjects.[30] In other words, effects may be identified only when consistent treatments are administered to subjects. Consistency also implies no interference between subjects. Sometimes called interference or the stable unit treatment value assumption (SUTVA), the idea behind interference is that the treatment administered to one subject has no impact on another subject.[31,32] Interference is not contamination; contamination involves treatments crossed between treatment conditions.

Identification in Randomized Trials

Ever since Hill's 1947 streptomycin study, RCTs have held a special place in the epistemology of medical research.[33] Many view RCTs as the gold standard in research. Few seem to recognize, however, that, by definition, RCTs fulfill all the core requirements for effect identification and thus, at least theoretically, yield clear causal conclusions about some treatment effect. Simply put, randomization is a mechanism that fulfills both the positivity and exchangeability requirements of effect identification. **Positivity** is ensured when the investigator randomizes subjects/groups from the target population, though the probability of assignment to a treatment need not be 0.50. In fact, sophisticated RCTs often have a variety of assignment probabilities, but all are greater than zero. The **exchangeability** requirement is met as sample sizes increases. It is has long been established that, in the long run, randomization eliminates confounding in both measured and unmeasured variables.[34] The upshot is that a study's control condition serves as the counterfactual substitute for the desired but unobservable counterfactual of the treatment group.[35] Finally, properly conducted RCTs meet the **consistency** requirement because investigators evaluate the same treatment administered in the same way per a written protocol.

Briefly stated, effect identification is straightforward in randomized trials, at least theoretically. In practice, of course, methodological issues such as adherence, external validity, and follow-up often complicate an investigation. Still, mastering the principles of counterfactual thinking and experimental design are essential for understanding and designing useful observational studies.

While CER investigators tend not to rely on conventional RCTs,[36] they nonetheless need to think about the desired counterfactuals and the ideal RCT that would conclusively answer the question at hand. It is useful to set aside considerations of budgets and ethics during such an intellectual exercise and instead consider which identification requirements are/are not met in a proposed (or actual) study.[37] The thought experiment also helps investigators appreciate that, if they cannot imagine an experiment that answers their questions in an "anything goes" world of no budget or ethical constraints, then the odds of generating useful results with a modest budget and non-experimental survey data are slim. In making a serious and important point, Angrist and Pischke wryly say that research questions that cannot be answered by any experiment are actually fundamentally unidentified questions.[5]

Identification in Observational Studies

However, in the constraints posed by the "real world", observational studies are often the design of choice for CER studies. Observational studies are treatment effect studies that do not randomly assign subjects to conditions.[12] For this reason and as already discussed, observational studies require more effort and careful thinking if they are to meet identification requirements. To begin, in the absence of random assignment, observational studies are prone to positivity violations. For example, CER investigations into a new pharmaceutical treatment must ensure that all persons not treated with the medication *could* have been treated, at least

theoretically. Such investigations raise the question, for example, of why a particular patient eligible for a therapeutic treatment did not receive it. Yet, despite the ethical discomfort in asking the question, CER investigators must recognize that, because of physiology, insurance plans, social discrimination, or even patient choice, some subjects occasionally have no chance of treatment or exposure to a given treatment.

Building on Cochran (1957), Oakes went so far as to say that when there is no chance of exposure, the treatment effect is structurally confounded. In other words, violations of positivity may make it impossible to rule out competing explanations tied up in confounding variables.[23,27,29] To see the problem of structural confounding, it is useful to consider Rosenbaum's classic investigation into early studies on the effect of Head Start, the early childhood education program.[12] Rosenbaum explains how early efforts to estimate the impact of the program failed because researchers compared children in Head Start to children not in Head Start. Simply put, only poor students are eligible for Head Start. Yet, the inclusion in the research of wealthy and thus ineligible children would confounds the effect estimate, especially given that socioeconomic status is strongly linked to educational outcomes. The proper design would compare outcomes of students in Head Start to those who are eligible for the program but who, for seemingly random reasons, do not participate in it. When there are structural reasons, such as policy or eligibility criteria as a condition for inclusion in a program or treatment, regression adjustment or matching cannot overcome problems of confounding. In any case, the complexities of both individual physiology and the health care system lead some investigators to eschew the use of observational data for CER, though doing so may limit progress in understanding the experience of treatment from patients' point of view.

Even when researchers believe that positivity assumptions hold, exchangeability is extremely difficult to achieve in observational CER studies. Absent random assignment (or with small sample sizes), it is highly probable that treatment and comparison groups differ systematically in ways related to outcome measures; in other words, effects are often confounded. Importantly, without randomization, confounding in unmeasured variables is always a threat. In addition, exchangeability in non-experimental CER studies is often undermined by the fact that sicker patients are almost always more likely to be treated and more likely to have poor outcomes compared to less sick patients.[4] With illness severity often at least somewhat related to genetics and socioeconomic background, the potential for violations of exchangeability is great. Finally, the data available to many CER investigators (often billing records) may not contain sufficient information (e.g., variables) to measure or mitigate confounding. Missing or erroneously measured information on the characteristics of the patient, clinician, office environment, and so forth may undermine exchangeability because of residual confounding.[38,39]

Consistency in non-experimental CER studies is often difficult to ensure because of patient- tailored dosing, technological change, and variation in clinical practice, among other factors. Meeting

the consistency requirement is often especially challenging in retrospective CER studies because definitions of treatments may be post hoc; furthermore, the information in a given data resource may be insufficient. For example, it is often not clear from available data exactly what procedures a physician performed on a given patient at a particular time. It is widely recognized that EMRs are often tailored to financial or clinical needs; research needs are a rare consideration.

Identification in Practice

CER investigators are likely to produce their most internally valid effect estimates from randomized trials—whether at the patient or clinic level (for example, cluster RCTs).[19] Again, at least theoretically, randomized trials, by design, meet the core identification requirements. Accordingly, CER investigators should conduct more randomized trials; for good reason, they are the gold standard in health research. Nonetheless, observational CER studies, especially those that are well designed and able to exploit large data sets, promise a better balance between internal and external validity while yielding timely and pragmatic results.

Faced with millions of records and a pressing need to find answers quickly, researchers need to recognize that effect identification has important implications for ensuring the optimal conduct of CER with observational data. To begin, investigators should imagine the ideal experiment in the absence of all practical constraints. They should develop an experimental protocol that details every aspect of the ideal but hypothetical experiment. For example, for a study of the independent and joint effects of two or more treatments (as is often the case), investigators should envision factorial experiments.

Once researchers have fully articulated the idealized experiment, they should identify and consider what they are able to study in view of available resources and ethical constraints. Such considerations will illuminate the differences between the ideal study and the study to be conducted. Identification problems will often surface when researchers compare the ideal study to the actual study just as fundamentally unidentifiable questions will likewise surface. Other differences between the ideal and actual study will relate to methodological limitations of one type or another. Researchers should note and fully address such differences in a report's methods and/or discussion section.

Indeed, CER investigators will benefit by recognizing that all study designs suffer methodological shortcomings, at least in comparison the ideal study design. Yet, such shortcomings are not a major threat to scientific advancement. The real threat arises when investigators do not recognize and address shortcomings (especially fundamental identification problems) in ways that research consumers can understand. By contrast, the clear articulation of the limitations of research translates into opportunities for future researchers to advance our understanding of treatments of interest. Simply put, scientific progress is not necessarily impeded by shortcomings and assumptions in current methods and data but rather by the lack of recognition and disclosure of those limitations.

While planning an investigation, CER researchers should imagine and then record all possible competing explanations for both expected and unexpected results. This important step in effect identification, especially in observational designs, gives rise to concerns about positivity, confounding, and consistency but may be understood as concerns about selection, confounding, measurement, and actual treatments applied. It may also point to the need for the collection of auxiliary data in order to rule various explanations in or out. Further, the development of a data analysis plan aimed at ruling out competing explanations can mitigate the pull of confirmation bias.[40] Many find it challenging to imagine alternative explanations for data that conform to a favorite hypothesis, but confidence in findings from observational designs is dependent on ruling out alternative explanations and the leverage of theories about the same. For this reason, R.A. Fisher, arguably the father of randomized experiments, famously said that investigators should make their theories elaborate.[23] Elaborate or complex theories offer many instances for testing empirical veracity; simple theories tend to fit lots of data sets and are thus less helpful because of identification problems.

Although not usually a matter for researchers to decide, prospective rather than retrospective observational studies are preferable. Defining outcome measures, treatments/exposures, and potential confounders ex ante (1) force investigators to ensure that they have the data needed to answer research questions and (2) limit any post hoc rationalization of results (i.e., confirmation bias). It may be a surprise that highly detailed research protocols are at least as important in observational CER studies as in their experimental kin.

Further Issues

CER is likely to pose problems that require careful thinking about effect identification. Three slippery problems merit attention. First is the problem of dynamic processes, whereby treatments at one time are a function of treatments and health outcomes at earlier times. The prototypical example is the identification of effects from pharmaceutical treatments administered to patients based on how well previous (pharmaceutical) treatments worked. This problem challenges assumptions about positivity and exchangeability in that strong if not determinant selection forces are usually at work. While the literature on dynamic processes is beyond the scope of this paper (Robins et al. 2000; Sturmer et al. 2011), CER investigators need to be wary of the issue.[4,41]

The second problem is heterogeneous treatment effects, a term that describes different outcomes from the same treatment. Owing to often unobservable aspects such as immune system responses and perhaps even psychological states from placebo effects, treatments affect different people differently. The vast and growing literature on heterogeneous treatment effects merits careful study.[19,42,43] For now, it is important to state that (1) the calculation of the average causal effect of a treatment may mask important differences and inhibit progress and that (2) instrumental variable models appear well suited to addressing heterogeneous treatment effects.

Finally, the problem of technological and/or therapeutic change requires careful consideration of the effect to be identified, estimated, and interpreted. Tied directly to the meaning of any effect, such considerations revolve around the nature of treatments and the research comparison/control groups used to evaluate them. The canonical RCT compares a new treatment to a placebo or perhaps the current standard of care. Such an approach is perfectly acceptable if the goal is to determine the degree of impact of a mature (as opposed to evolving) treatment on the outcome as compared to the placebo or standard of care. CER, however, often focuses on potentially helpful treatments that may be evolving. Therefore, CER investigators must not necessarily view assessments of novel treatments as conclusive; in practical terms, it is difficult to assess a moving target. Even worse, while potentially accurate, null or negative findings may undermine efforts to improve novel technologies or therapies until they are effective. [44–46]

Conclusion

Comparative effectiveness research holds great promise to improve our understanding of what therapy works, where, and for whom. Efforts such as the Observational Medical Outcomes Partnership (OMOP) experiment, which take a systematic approach to exploring how specific methods work to address particular questions, expand our understanding of how observational data and methods may be used for surveillance and outcomes research. To the extent that the CER and Patient-Centered Outcomes Research (PCOR) communities participate in these efforts, they will benefit from greater insights into the strengths and limitations of inference from observational data. In this way, CER may be viewed as part of a larger effort centered on learning health care systems. The coupling of agglomerated health systems with electronic medical records translates into an unprecedented abundance of data for CER. The question, however, is whether the data are suited to answering a given question.

This paper aimed to shore up the inferential foundations of CER by explaining the fundamental issue of effect identification. Effect identification precedes statistical estimation in the logic of empirical research, although, in practice, the core principles of identification are too rarely considered. Effect identification involves the discernment of causal effects from among competing explanations. CER that does not embrace the principles of identification will yield limited if not misleading results.

The standards of effect identification are demanding. Few CER investigations will meet the requirements for a fully identified effect, especially in observational designs. In fact, the search for competing hypotheses to explain a given result is endless, but it does not mean that "unidentified" effects are useless. Rather, the practice of identification illuminates the assumptions needed for drawing credible conclusions and highlighting the need for greater transparency in studies designed to assess treatment effects. As advocated here, recognition and forthright communication of the challenges associated with CER will advance comparative effectiveness research.

In the end, this paper maintains that, armed with an understanding of effect identification, CER is a promising strategy for advancing scientific understanding of what works best under what conditions. So-called "big data" from medical records may not yield better or more useful results. As in all scientific endeavors, what is required is the coupling of the right question to the appropriate data and methods.

References

1. Holve E et al. Opportunities and Challenges for Comparative Effectiveness Research (CER) with Electronic Clinical Data: A Perspective from the EDM Forum. Med Care. 2012 July;50:S11-8.
2. Sullivan P and Goldmann D. The Promise of Comparative Effectiveness Research. JAMA 2011;305(4):400-1. 3. Johnson, BH Electronic Data Methods (EDM) Forum Building the Infrastructure to Conduct Comparative Effectiveness Research and Patient-Centered Outcomes Research Using Electronic Clinical Data. Med Care. 2012;50:ii.
4. Stürmer, T. et al. "Nonexperimental Comparative Effectiveness Research Using Linked Healthcare Databases." Epidemiology. 2011;22(3):298 V5. Angrist JD and S Pischke JS. Mostly Harmless Econometrics: An Empiricist's Companion. Princeton, NJ: Princeton University Press; 2008.
6. Berk RA. Regression Analysis: A Constructive Critique. Thousand Oaks, CA: Sage; 2004.
7. Freedman D. Statistical Models and Shoe Leather (with discussion). In: Marsden, P.V., editor. Sociological Methodology. Vol. 21. San Francisco: American Sociological Association; 1991, pp 291-358.[Jossey-Bass??]
8. Hernan MA. A Definition of Causal Effect for Epidemiological Research. J Epidemiol Community Health. 2004;58(4):265-71.
9. Moffitt R. Remarks on the Analysis of Causal Relationships in Population Research. Demography. 2005;42(1):91-108.
10. Pearl J. Causality: Models, Reasoning, and Inference. New York: Cambridge University Press; 2000.
11. Robins JM Data, Design, and Background Knowledge in Etiologic Inference. Epidemiology. 2001;12(3):313-20.
12. Rosenbaum PR Observational Studies. Second ed. New York: Springer Verlag; 2002.
13. Sobel ME. Causal Inference in the Social and Behavioral Sciences. In: Arminger, G. et al., editors. Handbook of Statistical Modeling for the Social and Behavioral Sciences. New York: Plenum; 1995, pp. 1-38.
14. Susser M. Causal Thinking in Health Sciences: Concepts and Strategies of Epidemiology. New York: Oxford; 1973.
15. Rubin DB. Statistics and Causal Inference: Comment: Which Ifs Have Causal Answers. Journal of the American Statistical Association. 1986;81(396):961-2.
16. Morabia A. On the Origin of Hill's Causal Criteria. Epidemiology, 1991 Sep;2(5):367-9.
17. Evans AS. Causation and Disease: A Chronological Journey. New York: Plenum; 1993.

18. Rosenbaum PR. From Association to Causation in Observational Studies: The Role of Tests of Strongly Ingorable Treatment Assignment. Journal of the American Statistical Association. 1994;79(385):41-8.

19. Dowd BE. Separated at Birth: Statisticians, Social Scientists, and Causality in Health Services Research. Health Services Research. 2011;46(2):397-420.

20. Hsiao C. Identification. In: Griliches Z and Intriligator MD, editors. Handbook of Econometrics: Vol. 1. New York: Elsevier; 1983, pp. 223-83.

21. Manski CF. Identification Problems in the Social Sciences. In: Marsden, P.V., editor. Sociological Methodology. San Francisco: Jossey-Bass; 1993.

22. Manski CF. Identification Problems in the Social Sciences. Cambridge, MA: Harvard University Press; 1995.

23. Cochran WG. The Planning of Observational Studies of Human Populations. Journal of the Royal Statistical Society, Series A. 1965;128(2):243-65.

24. Freedman D. From Association to Causation via Regression (with comments). In: McKim VR and Turner SP, editors. Causality in Crisis? Statistical Methods and the Search for Causal Knowledge in the Social Sciences. Notre Dame, IN: Notre Dame University Press; 1997, pp. 113-82.

25. Manski, C.F. Identification for Prediction and Decision. Cambridge, MA: Harvard University Press; 2007.

26. Robins JM and Greenland S. Identification of Causal Effects Using Intrumental Variables. Journal of the American Statistical Association. 1996;91(434):456-8.

27. Oakes JM et al. Messer et al. Respond to "Positivity in Practice." American Journal of Epidemiology. 2010;171(6):680-1.

28. Greenland S and Robins JM. Identifiability, Exchangeability, and Epidemiological Confounding. International Journal of Epidemiology. 1986;15(3):413-9.

29. Oakes JM. The (mis)Estimation of Neighborhood Effects: Causal Inference for a Practicable Social Epidemiology. Soc Sci Med. 2004;58(10):1929-52.

30. VanderWeele TJ. Concerning the Consistency Assumption in Causal Inference. Epidemiology. 2009;20(6):880.

31. Sobel ME. What Do Randomized Studies of Housing Mobility Demonstrate? Journal of the American Statistical Association. 2006;101(476):1398-1407.

32. Tchetgen EJT. and VanderWeele TJ. On Causal Inference in the Presence of Interference. Statistical Methods in Medical Research. 2012;21(1):55-75.

33. Hill, AB. Suspended Judgment. Memories of the British Streptomycin Trial in Tuberculosis. The First Randomized Clinical Trial. Controlled Clinical Trials. 1990;11(2):77.

34. Fisher RA. The Design of Experiments. New York: Hafner Publishing Company; 1935.

35. Maldonado G and Greenland S. Estimating Causal Effects. International Journal of Epidemiology. 2002;31(2):422-38.

36. Fiore LD et al. A Point-of-Care Clinical Trial Comparing Insulin Administered Using a Sliding Scale versus a Weight-Based Regimen. Clinical Trials. 2011;8(2):183-95

37. Hernán MA et al. Observational Studies Analyzed Like Randomized Experiments: An Application to Postmenopausal Hormone Therapy and Coronary Heart Disease. Epidemiology. 2008;19(6):766.

38. Brookhart MA et al. Confounding Control in Healthcare Database Research: Challenges and Potential Approaches. Med Care. 2010;48(6):S114.

39. Becher H. The Concept of Residual Confounding in Regression Models and Some Applications. Stat Med. 1992;11(13):1747-58.

40. Nickerson RS. Confirmation Bias: A Ubiquitous Phenomenon in Many Guises. Review of General Psychology. 1998;2(2):175.

41. Robins JM et al. Marginal Structural Models and Causal Inference in Epidemiology. Epidemiology. 2000;11(5):550-60.

42. Florens JP et al. Identification of Treatment Effects Using Control Functions in Models with Continuous, Endogenous Treatment and Heterogeneous Effects. Econometrica. 2008;76(5):1191-1206.

43. Rhodes W. Heterogeneous Treatment Effects: What Does a Regression Estimate? Evaluation Review. 2010;34(4):334-65.

44. Cutler DM and McClellan M. Is Technological Change in Medicine Worth It? Health Affairs. 2001;20(5):11-29.

45. Lilford RJ et al. Trials and Fast-Changing Technologies: The Case for Tracker Studies. BMJ. 2000;320(7226);43-6 Vol., 320, No. 7226, 2000.

46. Mowatt G et al. When and How to Assess Fast-Changing Technologies: A Comparative Study of Medical Applications of Four Generic Technologies. Health Technol Assess. 1997;1(14):i.