What Is the Value of Big Data in Comparative-Effectiveness Research and Clinical Decision Making?

William Crown, PhD, Chief Scientific Officer, OptumLabs, Cambridge, MA, USA

William Crown, PhD

KEY POINTS . . .

The availability of patient data to support outcomes research is expanding rapidly, but significant effort needs to be invested in preparing such data before it is research ready.

With strong research design and appropriate statistical, methods database studies have demonstrated substantial success in replicating the average treatment effects from randomized trials.

Health economic models, in combination with simulation methods, can explore a variety of questions that are sometimes very difficult, or impossible, to examine with the data directly.

Evolution of the Health Care Data Landscape

Numerous initiatives are amassing huge repositories of claims, electronic medical records, and other data that can be used to support pharmacoeconomic research [1-3]. Such data are very rich, but linkages across all these sources of data are often limited and create complex data structures that create challenges for traditional multivariate statistical methods. New methods such as machine learning are now starting to be used in health economics and outcomes research. These methods are more facile at handling complex data structures, but traditionally have been focused upon prediction rather than estimating treatment effects. Can we bring those things together?

The availability of electronic medical records (EMR) data in the United States has expanded exponentially—primarily due to the meaningful use provisions of the Affordable Care Act. This increased availability, in combination with the well-known limitations of claims data with respect to clinical outcomes and severity measures, has spurred tremendous interest in conducting research with EMR data. Unfortunately, the state of knowledge regarding the use of EMR data for research is similar to that of the analysis of medical claims 20 years ago. This can lead to frustration when researchers attempt to use EMR data for research for the first time.

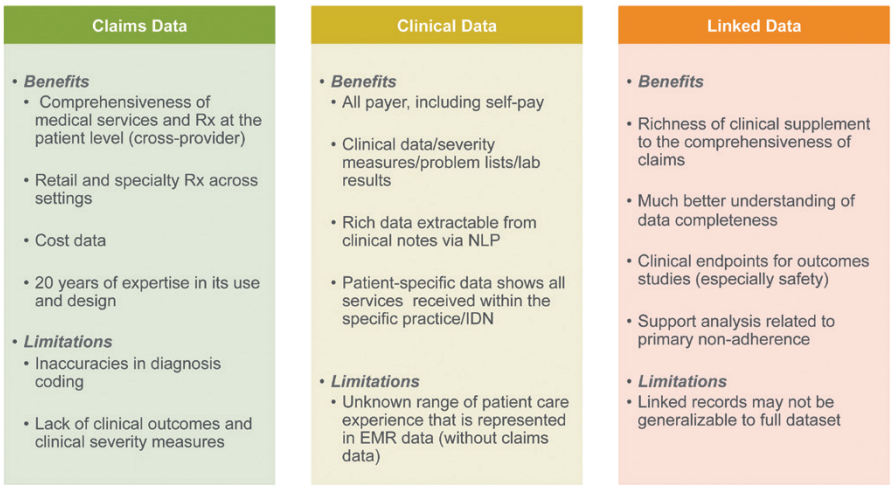
There are several characteristics of EMR data that make it challenging to use for research (see Fig. 1). EMR data tends to be specific to particular clinical settings. As a result, it is often difficult to understand patient comorbidity profiles and the breadth of interaction of patients with the health care system. In contrast, medical claims are very good at capturing the breadth of patient health care utilization across care settings, but lack clinical detail. Large clinical organizations, such as integrated delivery networks (IDNs), may have multiple EMRs and multiple EMR vendors for their different clinical sites. These data may not be linked and, even if they are, may be in different data formats. Finally, despite the availability in EMRs of structured fields for data such as height, weight, blood pressure, common laboratory results, etc., these fields are often empty and the data remain in unstructured notes. This makes the data very difficult to use for research.

PCORnet: An Example of Building Big Data Infrastructure

Most of the data that are traditionally used for health outcomes and epidemiology research come from claims and EMR systems that were not intended to support research. PCORnet goes beyond the aggregation and linking of data collected for another purpose to proactively building a research database that can support observational studies and clinical trials. The notion of creating a reusable infrastructure for comparative-effectiveness research (CER) was part of the Patient-Centered Outcomes Research (PCORI) Board of Governors' vision. PCORnet attempts to blend the capabilities of health care systems and patient-driven organizations to create a sustainable national ecosystem for research that is much more efficient and also more patient-centered than what exists today. In doing so, PCORI hopes to tackle much inefficiency in the research process—with a heavy emphasis on patient engagement.

To be research-ready, most data sources require significant cleaning and preparation before they can be used for research. A good example of this is the challenge of using EMR data for research when so much of the data content still sit in the notes. Of course, this is fine from the standpoint of the clinician treating the patient; the physician can simply look at the notes. But unstructured data present a real problem for the traditional statistical methods commonly used in research. To remedy the problem, natural language processing is required to pull content out of the notes and place it in structured fields. This is the case even when EMR systems have structured fields for vital statistics and laboratory results and so forth, because these fields are often empty and the data still sits in the notes.

Figure 1. Strengths of Claims, Clinical, and Linked Data



Note: EMR indicates electrical medical record; IDN indicates integrated medical network; NLP indicates Neuro-linguistic programming' and Rx indicates medical prescriptions.

The data elements included in PCORnet are extensive—including demographics, vital statistics, encounter data, lab data, diagnoses, procedures, and several patient-reported outcomes measures. In the future, PCORnet hopes to include health data from mobile devices, FitBits and trackers, biological specimen data, and social media. The intent is that all 71 health systems participating in PCORnet will map their data to a common model and create a rich resource for CER, both observational and interventional.

Drawing Causal Inferences from Big Data

Randomized trials offer the highest level of internal validity for estimating a treatment effect because randomization balances comparators on both observed and unobserved variables that may be associated with both treatment and outcomes. In the absence of randomization, statistical methods (e.g., propensity score methods) can be used to balance comparators on the basis of observed characteristics. Will big data—particularly linking variables that were previously unavailable—allow us to reduce the bias introduced by unobserved variables and get closer to the estimates we would have gotten with a randomized study? This question is crucial because, for a variety of reasons, we do not conduct trials for every treatment or policy question. Instead, we often use existing observational (and big) data to try to emulate a randomized trial—the target trial—that would answer our treatment and policy questions of interest. Any comparative effectiveness or safety analysis using a large database can be seen as an attempt to emulate a target trial.

An explicit specification of the target trial is helpful to conduct a sound analysis and avoid common methodological mistakes. To specify the target trial, we have to define the eligibility criteria, treatment strategies, outcome, statistical methods and analysis plan for the study, just as we would with a randomized trial. To emulate a trial using observational data, we need to emulate each of those components.

There are several examples in which observational analysis of preexisting data have failed to replicate the results of comparator Randomized Clinical Trials (RCTs) (e.g., postmenopausal estrogen plus progestin hormone therapy and risk of coronary heart disease). This failure is generally ascribed to inherent limitations in observational data that preclude appropriate confounding adjustment. However, these randomized-observational discrepancies can often be partly explained by a failure of the observational analysis to emulate the target trial. Finally, despite general perceptions to the contrary, careful replications of target trials using observational data tend to generate similar average treatment effect estimates with surprising regularity [4].

The problem of replicating a target trial is ultimately a problem of causal inference. But what if we're trying to analyze systems and systems of care where we have interactions, non-linearities, and feedback loops? Traditional statistical methods don't work very well for those kinds of methods. This is where modern causal inference methods, (i.e., Robins' g-methods [5], and health economic, and simulation modeling) step in [6].

Is There a Role for Health Economic Modeling with Big Data?

Traditionally, health economics models have coupled efficacy data from clinical trials with cost data culled from anywhere it can be found—often from the published literature or retrospective database studies. However, as noted earlier, the linkage of datasets has improved the ability of researchers to control for an increasing number of covariates which should, in turn, improve our ability to estimate the average treatment effects that we would have gotten had we done the target trial to answer the same question for the same patient population.

Nevertheless, there are some significant challenges that may be difficult to overcome even in the presence of all these data. The issue of causal feedback loops is challenging to address with traditional epidemiological or econometric methods. Moreover, people who get different interventions—whether they be treatments or tests—may not be comparable to each other. Sometimes they may differ along dimensions that can be observed and measured—such as co-morbidity profiles, gender, or age—but people receiving different interventions can also differ along dimensions that are not measured in the data. If these unmeasured factors are also correlated with outcomes, it will introduce bias into treatment effect estimates.

Economists refer to this problem as sample selection bias and it is a specific example of a broader set of issues that relate to what epidemiologists refer to as confounding and economists refer to as endogeneity. From a statistical standpoint, confounding or endogeneity results from any measurement issue that creates a correlation between the treatment variable and the error term of the outcome equation.

Another limitation with observational data is that you can't observe what didn't happen. There may be interventions or clinical strategies that people aren't using >

or they're not using very widely. Also, many innovations in clinical care come from innovative clinical strategies involving existing tests and treatments, having to do with questions such as when to stop a treatment, when to switch treatments, or how to monitor a treatment. You may not observe every possible clinical strategy, including some that might be good ideas. And there are always new interventions coming along. For example, suppose you wished to compare a new two-drug antiretroviral regimen for HIV with a commonly used three-drug regimen. Since the two-drug regimen isn't being used, we can't use big data to observe it. How would we use health economic modeling to examine this question?

Unfortunately, the state of knowledge regarding the use of EMR data for research is similar to that of the analysis of medical claims 20 years ago.

One important feature of health economic models is the ability to evaluate alternative "what if" scenarios. Parameters estimated from traditional statistical models are incorporated as parameters in health economics models, along with the probability distributions of treatment options, costs of care, etc. These parameters can all be modified to see what the impact would be on patient outcomes, health care costs, etc., under different treatment regimens. The ability to shift patient and physician behaviors with alternative policies such as benefit design can also be evaluated. A particular strength of health economic simulation models is their ability to account for feedback loops between the intervention and the response to the treatment.

Measurement of patient preferences is a key component of cost-effectiveness analyses. At the moment, health-related quality-of-life data are not generally available in medical claims and electronic medical record databases. PCORnet is an example of a big data research infrastructure starting to capture patient-reported outcomes. To the extent that patient preferences become part of medical care delivery or quality measurement (such as the Medicare STARS program in the US), such data will become increasingly linkable to other administrative data. Other sources of data on patient preferences are accumulating in patient communities organized around specific medical conditions and in disease registries. So there is reason to think that patient preferences will gradually become available in big data. Finally, even in the absence of direct measures of patient preferences, we do have measures of "revealed preferences" based upon the choices that they make.

One area where big data (particularly medical claims) are very strong is in measuring costs. However, even here we should be cautious. Patients with different therapeutic strategies may be different in observable or unobservable ways. And, even for otherwise similar patients, we may observe treatment variation due to provider prescribing differences. Again, some of the reasons for prescriber behavior may be observable, while others may not.

There are always going to be pieces of information that we're not going to have from the existing data. Big data is helping to shrink these gaps but they will never go to zero. Some issues, particularly non-linear feedback loops, are so complex that they cannot be handled with traditional statistical methods. And there will always be a need to be able to ask "what if" questions to assess patient outcomes and costs associated with alternative treatment regimens and health care policies. On balance, however, it seems clear that big data has much to contribute to the generation of new evidence about the value of medical and pharmacological interventions, health care policy evaluations, and many other areas. We are rapidly moving to the point where big data can be used to recommend alternative treatment options for doctors in clinical decision making, but still have more work to do in this area. It is here where the ability to infer causality is especially critical. Clinicians and regulatory authorities will be loath to go this last mile until there is consistent and robust evidence regarding the conditions under which it is safe to infer causality in observational research.

References

[1] Curtis L, Brown J, Platt R. Four health data networks illustrate the potential for a shared national multipurpose big-data network. Health Aff 2014;33:1178-86. [2] Etheredge L. Rapid learning: a breakthrough agenda. Health Aff 2014;331155-61. [3] Wallace P. Shah N, Dennon T, et al. 2014. Optum labs: Building a novel node in the learning health care system. Health Aff 2014;33:1187-94. [4] Anglemyer A, Horvath HT, Bero L. Healthcare outcomes with observational study designs compared to those assessed in randomized trials. The Cochrane Collaboration 2014. John Wiley & Sons, Ltd. [5] Robins JM. A new approach to causal inference in mortality studies with sustained exposure periods—Application to control of the healthy worker survivor effect. Mathematical Modelling 1986;7:1393-512. [6] Marshall D, Ijzerman M, Crown W, et al. Applying dynamic simulation modeling methods in health care delivery research: The SIMULATE checklist. Value Health 2015;18:5-16. ■

Additional information:

The preceding article is based on a workshop given at the ISPOR 20th Annual International Meeting, May 16-20, 2015.

To view Dr. Crown's presentation, go to: http://www.ispor.org/Event/ReleasedPresentations/2015Philadelphia