eGEMs
Generating Evidence & Methods to improve patient outcomes

Analytical Methods for a Learning Health System: 3. Analysis of Observational Studies

Michael Stoto, PhD;[i] Michael Oakes, PhD;[ii] Elizabeth Stuart, PhD;[iii] Randall Brown;[iv] Jelena Zurovac;[iv] Elisa L. Priest, DrPH[v]

ABSTRACT

The third paper in a series on how learning health systems can use routinely collected electronic health data (EHD) to advance knowledge and support continuous learning, this review describes how analytical methods for individual-level electronic health data EHD, including regression approaches, interrupted time series (ITS) analyses, instrumental variables, and propensity score methods, can also be used to address the question of whether the intervention "works."

The two major potential sources of bias in non-experimental studies of health care interventions are that the treatment groups compared do not have the same probability of treatment or exposure and the potential for confounding by unmeasured covariates. Although very different, the approaches presented in this chapter are all based on assumptions about data, causal relationships, and biases. For instance, regression approaches assume that the relationship between the treatment, outcome, and other variables is properly specified, all of the variables are available for analysis (i.e., no unobserved confounders) and measured without error, and that the error term is independent and identically distributed. The instrumental variables approach requires identifying an instrument that is related to the assignment of treatment but otherwise has no direct on the outcome. Propensity score methods approaches, on the other hand, assume that there are no unobserved confounders. The epidemiological designs discussed also make assumptions, for instance that individuals can serve as their own control.

To properly address these assumptions, analysts should conduct sensitivity analyses within the assumptions of each method to assess the potential impact of what cannot be observed. Researchers also should analyze the same data with different analytical approaches that make alternative assumptions, and to apply the same methods to different data sets. Finally, different analytical methods, each subject to different biases, should be used in combination and together with different designs, to limit the potential for bias in the final results.

[i]Georgetown University, [ii]University of Minnesota, [iii]Johns Hopkins University, [iv]Mathematica Policy Research, [v]Baylor Health Care System

Introduction

Learning health systems use routinely collected electronic health data (EHD) to advance knowledge and support continuous learning. Even without randomization, observational studies can play a central role as the nation's health care system embraces comparative effectiveness research and patient-centered outcomes research. However, neither the breadth, timeliness, volume of the available information, nor sophisticated analytics, allow analysts to confidently infer causal relationships from observational data. Rather, depending on the research question, careful study design and appropriate analytical methods can improve the utility of EHD.

This is the second paper in a series (see Box 1) on how learning health systems can use routinely collected electronic health data (EHD) to advance knowledge and support continuous learning, this review summarizes study design approaches, including choosing appropriate data sources, and methods for design and analysis of natural and quasi-experiments. The first paper[1] began by drawing a distinction between big-data style analytics of electronic health data (EHD), with its claims that randomized studies were no longer necessary, and traditionalists who believe that without randomization little can be known with certainty. Of course this is a false distinction; some questions do not involve assessing a cause and effect relationship, but when causal assessment is

Box 1. Series on Analytic Methods to Improve the Use of Electronic Health Data in a Learning Health System

This is one of four papers in a series of papers intended to (1) illustrate how existing electronic health data (EHD) data can be used to improve performance in learning health systems, (2) describe how to frame research questions to use EHD most effectively, and (3) determine the basic elements of study design and analytical methods that can help to ensure rigorous results in this setting.

- Paper 1, "Framing the Research Question,"[2] focuses on clarifying the research question, including whether assessment of a causal relationship is necessary; why the randomized clinical trial (RCT) is regarded as the gold standard for assessing causal relationships, and how these conditions can be addressed in observational studies.
- Paper 2, "Design of observational studies,"[3] addresses how study design approaches, including choosing appropriate data sources, methods for design and analysis of natural and quasi-experiments, and the use of logic models, can be used to reduce threats to validity in assessing whether interventions improve outcomes of interest.
- Paper 3, this paper, describe how analytical methods for individual-level electronic health data EHD, including regression approaches, interrupted time series (ITS) analyses, instrumental variables, and propensity score methods, can be used to better assess whether interventions improve outcomes of interest.
- Paper 4, "Delivery system science,"[4] addresses translation and spread of innovations, where a different set of questions comes into play: How and why does the intervention work? How can a model be amended or transported to work in new settings? In these settings, causal inference is not the main issue, so a range of quantitative, qualitative, and mixed research designs are needed.

necessary observational studies of existing EHD can be a useful complement to RCTs. In particular, when the question is whether an intervention "works" – improves outcomes of interest, causal inference is indeed critical, but appropriately designed and analyzed observational studies can yield valid results that better balance internal and external validity than RCTs.

When the question is whether an intervention improves outcomes of interest, the second paper in this series illustrates how study design methods can help researchers identify valid results that better balance internal and external validity than RCTs. The methods discussed include choosing appropriate data sources, epidemiologic designs, methods for design of natural and quasi-experiments, and the use of logic models. The primary issue addressed by these evaluation designs is how to estimate the counterfactual – what would have happened if the intervention had not been implemented. Even with a strong design, however, the potential for bias remains.

Faced with the need to infer cause and effect when randomization is not feasible, statisticians and econometricians have developed a series of analytical methods for "causal analysis." The current paper complements the second by describing how analytical methods for individual-level electronic health data EHD, including regression approaches, interrupted time series (ITS) analyses, instrumental variables, and propensity score methods, can also be used to address the question of whether the intervention "works." Each of these approaches addresses Cochran's call for methods to adjust for differences in observed characteristics between treatment and control groups in order to isolate the effect of an intervention from other factors.[5,6,7] These methods are routinely used to study health interventions when randomization is not possible, and despite the potential for bias methods researchers from various disciplines agree that each method has merit when implemented with care. The analytical methods discussed in this paper can be used in combination; doing so can limit the potential for bias in the final results if the individual methods are subject to different biases. These methods can also be combined with the design approaches discussed in the second paper in this series[8] to limit threats to validity.

This paper does not attempt to serve as a textbook or describe these approaches in detail. Rather, it presents these methods in a consistent framework rather than provide detailed information on each topic. Because the use of existing EHD is not yet well developed, some of the examples use other types of data but were chosen to illustrate the methods.

The methods discussed in this section primarily involve the use of individual-level EHD. Since each of these paradigms face the same basic inference questions, there is some overlap in the material covered, and throughout we explain how each method relies on assumptions that are often not possible to verify with the existing data. This paper concludes with a discussion of "analyzing observational data like randomized experiments." This is not so much an analytic method *per se*, but rather a general approach or framework that should cut across all the methods.

Causal Inference Framework

As background for the methods described in this section it is useful to clarify a framework for causal inference. The fundamental idea is that, for a given individual, the "effect" of a treatment is based on the difference between that outcome that would be observed if the person receives the treatment and what would be observed if the person receives the comparison condition instead (the counterfactual). The problem, of course, is that no single individual can receive ***both*** the treatment and the comparison

condition at the same time. For instance in Stuart's study of the Medicare Part D program discussed below the treatment group members (dual eligibles) are likely sicker and older than the comparison group. Simply comparing those who are dual eligibles to other Medicare or Medicaid recipients would yield a biased estimate of the intervention's effect. Randomization solves this by creating a control group that, on average, is no different from the treatment group, and in particular there will be no confounding.

The regression-based approaches described assume that all of the factors that differentiate the treatment and control group members are represented in the observed variables and covariates. The instrumental variables approach identifies special variables (the "instruments") that affect treatment but are unrelated to outcomes except through the treatment, and estimates how much of the variation in the treatment variable that is induced by the instrument - and only that induced variation - affects the outcome measure. Propensity score methods model the factors related to the probability of treatment assignment and, typically, match treated and untreated based on such probabilities. These models also assume that the causal model is correctly specified. This can be hard to assess, but directed acyclic graphs (DAGs) can be used to clarify assumptions about causal pathways and use their representation in graphical form to guide selection of covariates for statistical adjustment through structural equation models (SEM) or other approaches,[9] although the details are beyond the scope of this paper.

Regression Approaches

Perhaps the simplest and most intuitive approach to analyzing observational data is to fit a linear statistical model of the form

(1) $Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + e_i$

where

Y_i is the outcome variable for subject i
X_i is an indicator variable for the treatment, e.g. 1 for treatment group and 0 for control
β_1 is the effect of the treatment, conditional on the covariates
Z_i represents other factors that influence the outcome
e_i is an independent and identically distributed (iid) error term.

The parameters in Equation 1 are typically are estimated by ordinary least squares (OLS) methods, and such "OLS estimates" are commonly used to describe this regression approach. In this model, the fitted value of β_1 estimates the effect of the treatment, and can be evaluated using standard statistical hypothesis tests. This approach can be extended as necessary if Y_i is categorical or dichotomous (e.g., logistic regression), there are multiple Z's, or the relationship is non-linear. Another extension is known as a "difference-in-differences" approach, which uses the difference in an outcome variable before and after an intervention as Y_i, which can have the benefit of individuals serving as their own control.

Despite the simplicity of this approach, there are many ways that things can go wrong when applied to observational data. Most basically, regression approaches assume that the actual causal relationship between the treatment, outcome, and other variables is properly specified, all of the variables are available for analysis and measured without error, and that the error term is independent and identically distributed. In particular, the two groups may be on different trajectories, and would not have exhibited the same difference after the intervention that they did before. The relationship could be improperly specified; the functional form could be incorrect or a variable omitted from the model may have a relationship with Y, X and/or Z. In

addition, X and/or Z may depend, in part, on Y, for example if the treatment received (X) is dependent on Z (confounding bias) or Y (endogeneity or selection bias). Also, some of the Z may not be available for analysis or measurement errors may affect X and/or Z. These problems could result in bias in the estimated treatment effect (β_1). They could also cause the error term e_i to not be iid, which would lead to incorrect confidence intervals and hypothesis tests.

It is standard practice in econometrics to assess omitted variable bias by identifying the available variables that are most closely related to the missing variable and seeing how the results change when these variables are dropped from the model. One never knows, however, how well these variables capture the effect of the missing variable, or of additional missing factors that may exist but are unknown to the researcher.

Regression Discontinuity (RD) Method

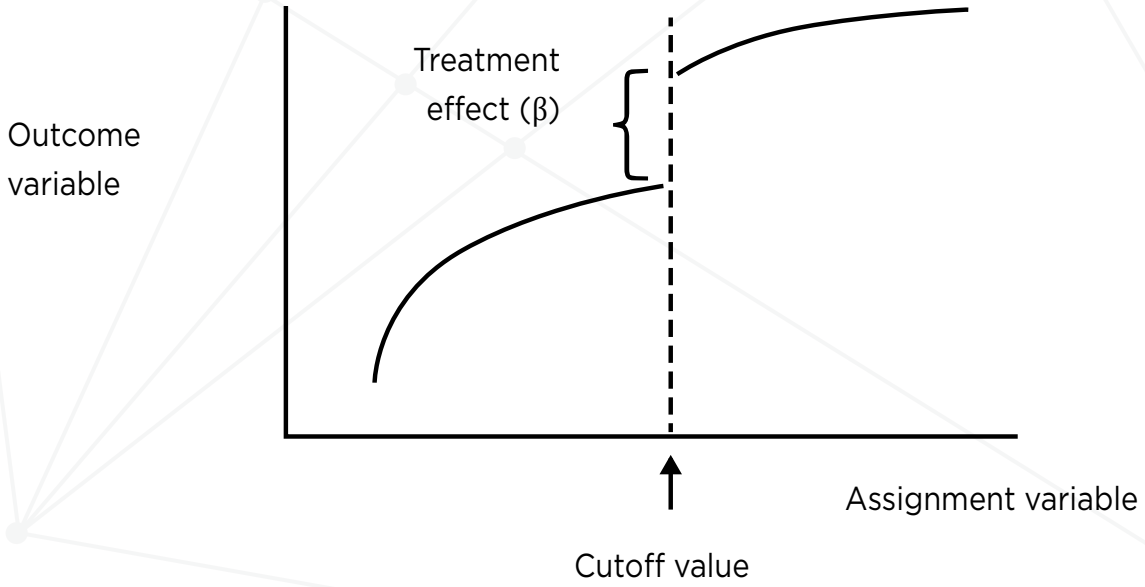
A variant of this approach, the regression discontinuity method,[10,11] is used when assignment to treatment is based on a cutoff value of a continuous variable Z. The relationship between Y and Z is modeled as

$$(2) \quad Y_i = \beta_0 + \beta_1 X_i + f(Z_i) + e_i.$$

The functional form of $f(Z_i)$ could simply be linear (which would be modeled as $\beta_2 Z_i$) or alternatively Y or Z could be transformed to fit this approach. As in other regression approaches, the fitted value of β_1 estimates the effect of the treatment, and can be evaluated using standard statistical hypothesis tests. Figure 1 illustrates this approach.

The within-study comparisons literature has shown that RD analyses of observational data generally replicate RCT results well despite the use of different

Figure 1. Regression Discontinuity Method



Source: Adapted from Dowd & Oakes.[10]

statistical methods to estimate the RD effect.[12] This model assumes that subjects just to either side of the cutoff value are likely to be similar in all relevant respects, including those unobserved. The challenge is ensuring that the relationship between Y and Z is completely and correctly modeled; otherwise the effect parameter β_1 will be biased. Furthermore, the effect estimated, is considered valid only for observations close to the cutoff of the assignment variable Z, not more generally. Another challenge, relating to implementation, is ensuring that assignment to both the treatment and comparison conditions adheres strictly to the cutoff value of Z.

Interrupted Time Series Analyses

To motivate the need for interrupted time series (ITS) methods,[13] Ross-Degnan[14] considers the Rational Prescribing in Primary Care (RaPP) cluster randomized trial of a tailored intervention to improve the use of antihypertensive and cholesterol medicines for primary prevention of cardiovascular disease (CVD) in Norway.[15] The intervention, an educational outreach by pharmacists with audit and feedback, and computerized reminders in the EMR was implemented in 70 practices including a total of 257 physicians. The control subjects received passive dissemination of evidence-based guidelines (69 practices; 244 physicians). Outcomes were measured monthly for all eligible patients in participating practices one year before and after intervention.

Figure 2 illustrates the traditional "difference-in-differences" analysis of RaPP study with prescribing of low-dose diuretics as the outcome variable. For the sake of comparison, the difference between change in the treatment and control groups is 9.0 percent (95% C.I. 4.9% - 13.1%), a significant improvement. Figure 3 displays the ITS analysis of the RaPP study based only on the intervention group data. Displaying the results by month vs. an annual basis shows that adherence to guidelines changed immediately after the intervention began, which is less likely to be due to some other cause than if the change had occurred at some other time. The monthly data also suggest that the effect not only didn't drop off in time, but might have increased. In addition, even though the control group data were not used, the ITS estimate of the effect, 11.5 percent (95% C.I. 9.5% - 13.5%), is consistent with the randomized trial analysis.

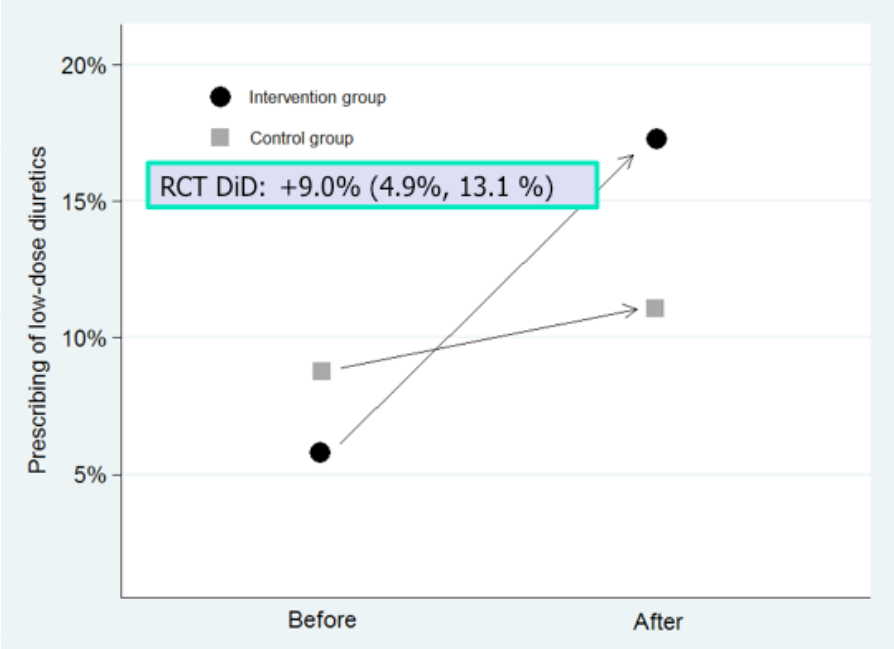
Figure 4 summarizes the logic of ITS analysis and shows how parameters can be estimated by segmented linear regression. Figure 5 shows how this approach can be extended to multiple time segments. In this example, New Hampshire Medicaid data on 860 multiple drug recipients show the effect of the implementation a reimbursement cap in August 1981, which was replaced by a $1 copay the following year. The statistical model is as follows:

(3) $Y_t = \beta_0 + \beta_1 * time_t + \beta_2 * policy1_t + \beta_3 * time\ after\ policy1_t + \beta_4 * policy2_t + \beta_5 * time\ after\ policy2_t + e_t$

In this model, β_2 is the effect of implementing policy 1 and β_4 is the effect of implementing policy 2, conditional on the covariates.

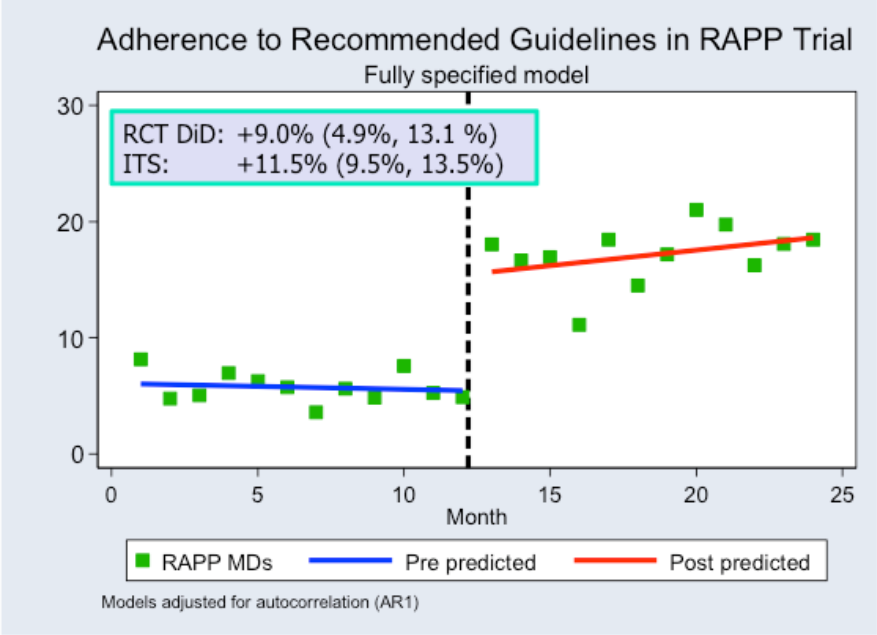
One key assumption is that the baseline trend correctly reflects what would have happened after the intervention time point, had the intervention not occurred. This in turn depends on the trends within segments being linear, and that the autocorrelation structure of errors is correctly modeled. The New Hampshire example in Figure 5 illustrates how simple assumptions about linearity can lead to misleading conclusions. The most basic ITS model also depends on the assumption that there is no lag between when the intervention occurs and when its effects are reflected in the outcome measures. With sufficient time points, more complex and flexible non-linear models could also be used to relax these assumptions.

Figure 2. Traditional Difference in Differences Analysis of RaPP Study



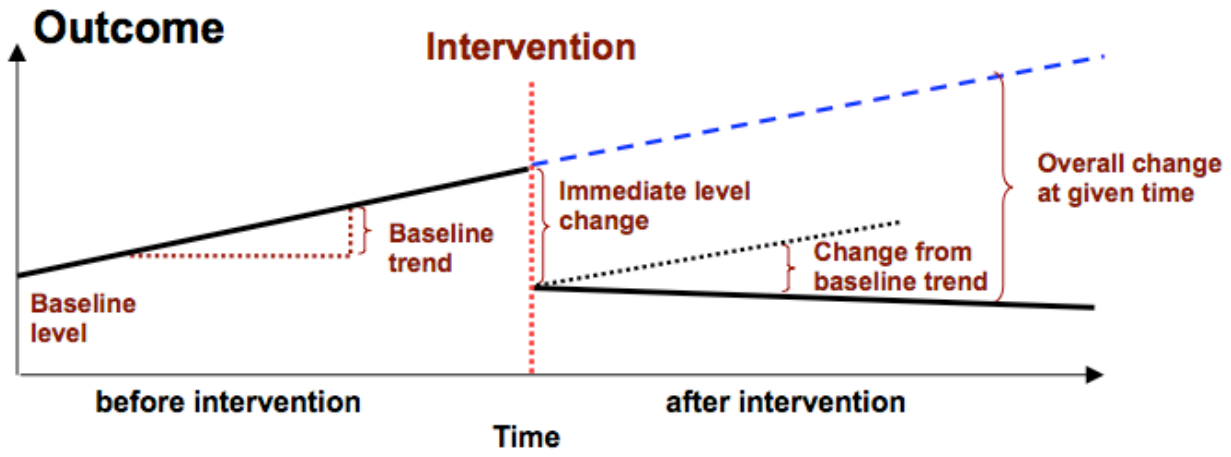
Source: Ross-Degnan and colleagues.[14]

Figure 3. ITS Analysis of RaPP Study: Intervention Group Only



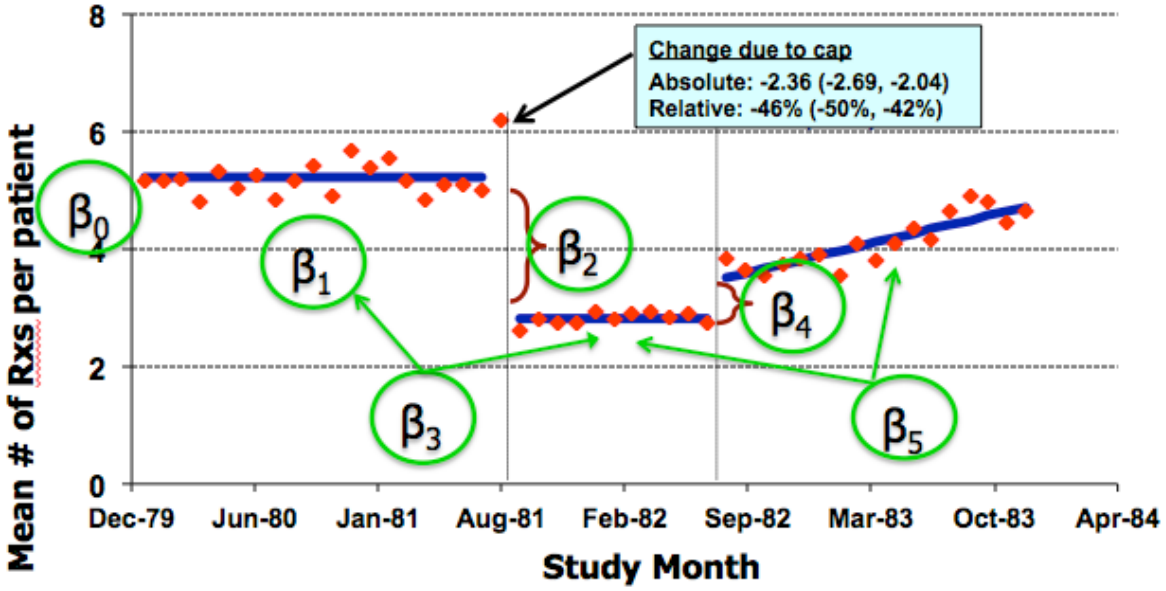
Source: Ross-Degnan and colleagues.[14]

Figure 4. ITS Logic and Parameters Estimated by Segmented Linear Regression



Source: Ross-Degnan and colleagues.[14]

Figure 5. Parameters of ITS Model



Source: Ross-Degnan and colleagues.[14]

The major threats to the validity of the ITS design are: confounding (i.e. a co-occurring intervention), selection (pre-intervention factors that affect inclusion in the intervention, such as volunteers), regression to the mean (groups selected on baseline values), instrumentation (changes in measurement or ascertainment), and history or maturation (some other event or natural process explains the observed effect).

The major threats to the reliability of ITS estimates are: unstable data and wild data points, low frequency outcomes (e.g., deaths), boundary conditions (e.g., percentages which are bounded between 0 and 100), short segments that inaccurately reflect the trend, changing denominator populations, and non-linear trends.

Ross-Degnan[16] suggests the following approaches for strengthening ITS studies:

- check data quality: identify and remove outliers and implausible data, impute missing data;
- contrast multiple outcomes or groups such as high-risk subgroups or differential response;
- account for policy phase-in including anticipatory effects or post-intervention lag; and
- test model assumptions including normality of errors, linearity of segments.

Although the description to this point assumes that there is only one group being followed, ITS analysis can also be used to compare two or more comparison groups. Indeed, the results of an ITS analysis are strengthened if the comparison groups are matched by standardizing or using propensity scores (see below) or chosen using principles of natural and quasi-experimental design as discussed in Stoto.[17]

In summary, the advantages of ITS analysis include an intuitive visual display, direct estimate of effects, and the controls it provides to common threats to validity. The limitations are that the ITS method requires reasonably stable series and relatively long segments. There can also be boundary problems and sensitivity to points near end of segments. Also, because ITS analysis uses aggregate data, there is no opportunity for patient-level adjustment, but one can use risk-adjusted rates.

Instrumental Variables

The instrumental variables approach addresses the causal inference problem by identifying special variables ("instruments") that affect the treatment that research subjects receive, but are unrelated to the outcomes they experience except through the treatment, and estimating how much of the variation in the treatment variable that is induced by the instrument - and only that induced variation - affects the outcome measure. The idea is that the instrument can be thought of as more plausibly randomly assigned than the treatment of interest, and that the instrument affects whether an individual takes the treatment but does not directly affect outcomes. A classic example of an IV is the distance subjects lived from a health care facility offering two types of emergency procedure.[18] The Physician prescribing preference example below demonstrates how instruments can be found in commonly available EHD.

Specifically, this approach centers on two regression equations, generally fit by two stage least squares:

(4) $X_i = \alpha_0 + \alpha_1 Z_i + \alpha_2 IV_i + f_i$

(5) $Y_i = \beta_0 + \alpha_1 Z_i + \beta_1 \hat{\square}_i + e_i$

where

IV_i is the instrumental variable for subject i

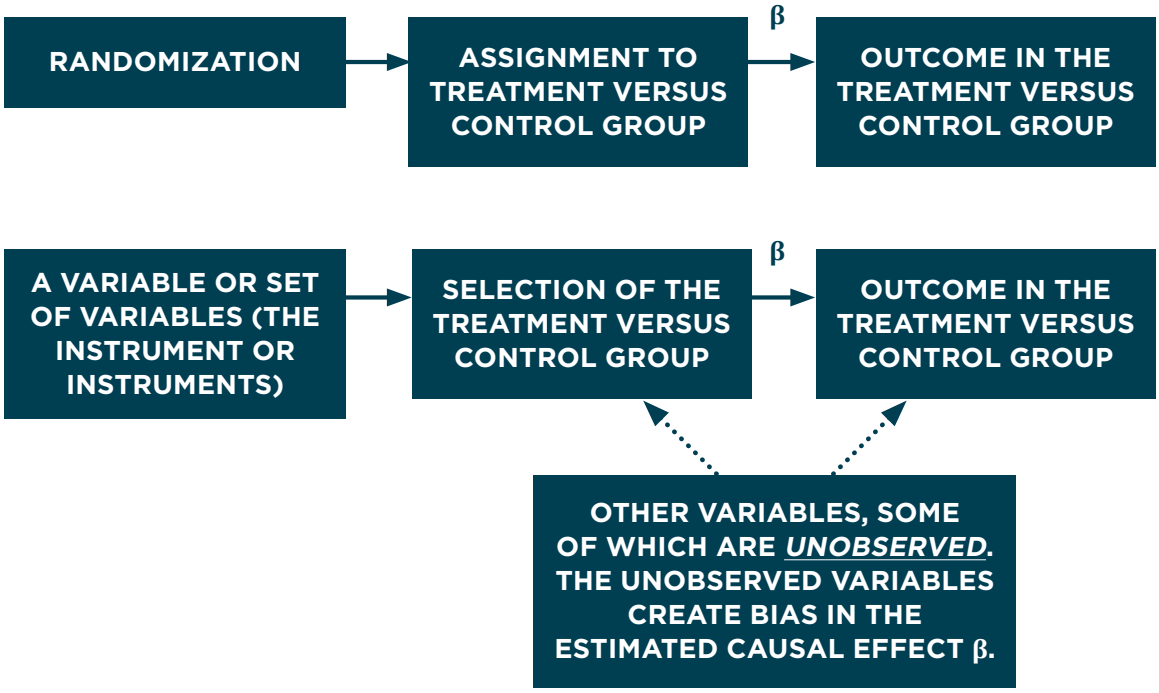
$\hat{\square}_i$ is the predicted value of X_i after fitting equation 2

β_1 is the effect of the treatment, conditional on the covariates

e_i and f_i are iid error terms

and the rest of the variables are as defined above. Figure 6 compares this approach to an RCT.

Figure 6. Causal Diagrams for RCTs and Instrumental Variables



Source: Adapted from Dowd & Oakes.[10]

For the instrumental variables approach to be effective, two critical assumptions must be true. The first is known as relevance, i.e. that there is a strong association between the IV and the treatment variable X. Using weak or poor instruments (those for which this association is not strong) can lead to biased and imprecise estimates of the treatment effect. The second assumption, exogeneity, is that the correlation between the IV and e_i, the error term in equation 5, must be zero. Leaving an important Z_i out of equation 4, perhaps because it was not available, can bias the estimated treatment effect (β_1). Unfortunately, there is no way to be sure that these conditions are met in any particular situations. There are tests to identify the best of multiple instruments conditional on having a good one, but none that test whether a particular instrument is truly exogenous, or whether any of the instruments are "good enough" to yield reasonably precise estimates. Baiocchi and colleagues demonstrate how the strength of an instrument relates to observed and unobserved covariates and discussion approaches for building stronger instruments and testing how well they work.[19]

Example of Instrumental Variables: Physician Prescribing Preference

Rassen and colleagues[20] illustrate the instrumental variables approach with an example study about the risk of short-term mortality associated with the initiation of antipsychotic medication (APM). The

premise is that the IV in this example, physician prescribing preference (PPP), predicts which drug a patient will be treated with but is otherwise unrelated to the outcome. This premise could be incorrect if, say, some of the patients had already been tried on APM and had a bad reaction.

The study uses data from two sources: Pennsylvania's (PA) Pharmaceutical Assistance Contract for the Elderly (PACE) program from 1994 to 2003 as well as for British Columbia (BC) residents aged 65+ from 1996 to 2004. The comparison is between initiators of conventional vs. atypical APM therapy, and the outcome: mortality within 180 days of initiation (the index date). The available covariates reflect baseline patient characteristics (coexisting illnesses and use of health services) in the 6 months before the index date. Frailty, cognitive impairment, ability to perform activities of daily living are all potentially important but are not available.

The study examined 25 different variants of PPP as an IV. The "base case" was the approach used in the original analysis: an indicator variable based on the physician's current preference for conventional vs. atypical APM therapy. If the physician's previous APM prescription was for a conventional APM, then for the next patient, the physician was classified as a "conventional APM prescriber." Otherwise, the physician was classified as an "atypical APM prescriber." Rassen and colleagues considered variants based on (1) preference assignment algorithm (e.g. the number of conventional APM prescriptions out of the previous 2-4 prescriptions), (2) cohort restrictions based on physician and patient characteristics, and (3) stratification criteria (e.g. patient of a similar age). Eventually they determined that restricting the analysis to primary care physicians produced the best instrument.[21]

Table 1 displays the estimated differences in the risk of all-cause mortality within 180 days of initiation of conventional versus atypical APM treatment based on four different statistical models; the first three are based on different versions of OLS regression, and the fourth is the IV analysis estimate. The IV estimates are substantially different from the OLS estimates, which themselves vary. In the Pennsylvania base case (including all prescribing physicians), for instance, the IV estimate of excess risk is 7.69 per 100 patients, compared to the OLS estimates that range from 2.47 to 3.91 per 100 patients.

The estimates from an IV analysis often have larger standard errors, and this effect can be seen in Table 1. But are the IV estimates sufficiently less biased to justify this? Rassen and colleagues[22] address this by examining the key assumptions of the IV approach. First, they assess instrument strength using a partial F test from the first-stage regression, which predicts treatment as a function of instrument and covariates, and find that it is significant at the 5 percent level for all cohort definitions. They also report that the partial r^2 between instrument and treatment conditional on other covariates in model, comparing across cohort definition, ranged from 0.028 to 0.099. Thus they conclude that the instrument strength is high.

Rassen and colleagues[23] also find that using PPP as an IV generally alleviated imbalances in non-psychiatry-related patient characteristics, making the instrument more plausibly randomly assigned than are APMs themselves. In Pennsylvania, for instance, the proportion of the patients who were male was 15.1 percent in the atypical APM group and 20.1 percent in the conventional APM group for a difference of 5.0 percent. When stratified by IV, the difference was reduced to 1.8 percent. Averaging over the 25 different variants of PPP, the overall imbalance was reduced by an average of 36 percent over the two cohorts.

Table 1. Differences in Risk of All-cause Mortality Within 180 Days of Initiation of Conventional Versus Atypical APM Treatment

POPULATION AND VARIATION	EVENTS IN CONVENTIONAL APM GROUP	EVENTS IN ATYPICAL APM GROUP	UNADJUSTED OLS ESTIMATE	AGE/SEX-ADJUSTED OLS ESTIMATE	FULLY ADJUSTED OLS ESTIMATE[a]	IV ANALYSIS ESTIMATE
BRITISH COLUMBIA						
Base case (unrestricted)	1,806	2,307	4.46 (3.69, 5.23)	4.49 (3.75, 5.22)	3.55 (2.74, 4.37)	4.00 (0.94, 7.06)
Restricted to PCPs (R6)	1,735	2,115	4.24 (3.41, 5.06)	4.48 (3.68, 5.28)	3.59 (2.70, 4.48)	3.11 (-0.57, 6.79)
PENNSYLVANIA						
Base case (unrestricted)	1,307	1,628	2.69 (1.65, 3.73)	2.47 (1.46, 3.49)	3.91 (2.68, 5.13)	7.69 (1.26, 14.12)
Restricted to PCPs (R6)	960	1,129	2.39 (1.07, 3.71)	2.29 (0.98, 3.60)	4.32 (2.71, 5.93)	5.34 (-3.53, 14.21)

Adjusted for age, sex, race, year of treatment, and history of diabetes, arrhythmia, cerebrovascular disease, myocardial infarction, congestive heart failure, hypertension, other ischemic heart disease, other cardiovascular disorders, dementia, delirium, mood disorders, psychotic disorders, other psychiatric disorders, antidepressant use, nursing home residence, and hospitalization. See text for description of the base case and restriction to PCPs. NOTE. The values within brackets are 95 percent confidence intervals. Risk differences are expressed per 100 patients. Abbreviations: APM, antipsychotic medication; OLS, ordinary least squares; IV, instrumental variable; PCP, primary care physician. Source: Rassen and colleagues.[20]

A final key assumption is the assumption of no direct effect of the instrument on the outcome. Rassen and colleagues argue that this is a reasonable assumption here, although it could be violated if, for example, PPP is associated with higher or lower quality of care in general (e.g., if physicians who prescribe a particular type of APM also tend to provide lower or higher quality of care in general).

Thus, even done well, IV is fraught with difficulty in interpretation, due to the large standard errors. Based on these analyses, Rassen and colleagues conclude that PPP was at least a reasonably valid instrument in this setting, and implicitly that the IV estimates are superior to the OLS estimates. This type of careful analysis of whether a particular instrument is truly exogenous, or whether any of the instruments are "good enough" to yield reasonably precise estimates is not common, and without it one cannot be sure that the results are valid.

Propensity Score Methods

Propensity score methods aim to equate treatment and comparison groups on a single variable, the probability of treatment, which is modeled from a set of observed characteristics and estimated on the pool of treatment group members and potential comparison group cases. The key to this is the propensity score, p, which is defined as the predicted probability of receiving the treatment given the observed covariates.[24,25]

There are five basic steps involved in using propensity score methods:

1. Estimate p, typically estimated using logistic regression, although non-parametric approaches such as random forests[26] have been shown to potentially work better.
2. Use the propensity score to equate groups through matching, weighting, or sub-

classification. Matching involves finding one or more comparison cases for each treated case that have similar values of p. Propensity scores can also be used to create weights $p/(1-p)$, which gives less weight to comparison subjects that look less like the treated group. Sub-classification forms subgroups of individuals with similar propensity scores (for example, 10 subclasses, defined by propensity score deciles).

3. Check how well the equating worked to create balance in observed covariates. Since the goal is to reduce bias by forming groups that look similar on the observed covariates, we can see how well the matching worked by comparing the distributions of the covariates in the equated treatment and comparison groups.
4. Estimate the treatment effect by comparing outcomes in equated groups. With matching, this involves comparing outcomes in the matched groups (some weighting will be required if the number of matches selected for some treatment group observations differs from the number selected for other treatment group cases). Alternatively use the weights described in step 2 to calculate the average treatment effect. With sub-classification, effects are estimated separately within each subclass and then aggregated. (Note that these approaches are the same ones used to calculate the balance measures in Step #3).
5. Conduct a sensitivity analysis to unobserved confounding. This can be done in a number of ways, for instance by positing an unobserved confounder and obtaining adjusted impact estimates if that confounder existed, given its assumed characteristics.

Schneeweiss and colleagues[27] demonstrate how this method can be used with health care claims data to study the safety and effectiveness of medications. Using a multi-step algorithm to implement a high-dimensional propensity score adjustment with claims data, the authors demonstrate improved effect estimates compared with adjustment limited to predefined covariates, when benchmarked against results expected from randomized trials. Other researchers, however, have found propensity score methods to yield very different estimates than those from a RCT using the same treatment group cases.[28]

Example of Propensity Score Methods: Medicare Part D Prescription Drug Program

To illustrate the use of propensity score analysis, Stuart and colleagues[29] use an analysis of the effect of the Medicare Part D prescription drug program, which became available in 2006, on individuals with serious mental illness. In particular, the study was focused on individuals eligible for both Medicaid and Medicare, known as "dual-eligibles," who transitioned from state Medicaid coverage to commercial Medicare coverage and asked about the impact on medication continuity and outcomes such as inpatient admissions, mortality, and the cost of care. The study was based on Medicare and Medicaid billing data and the population was Maryland residents with schizophrenia, bipolar, or depressive disorders with dual eligibility on January 1, 2006.

As the propensity score, Stuart and colleagues[30] estimated the probability of being a dual eligible. They fit a logistic regression relating eligibility to baseline measures of the key outcomes of interest, as well as diagnoses, demographics, and other covariates. The predicted value from this equation became each individual's propensity score, p. They then analyzed the data using a difference-in-differences design, comparing pre-post change in utilization between dual-eligibles and comparable Medicaid only patients. Stuart and colleagues tried two approaches to equating groups. The first was a 1:1 match based on nearest propensity score. The second used the propensity scores in a "weighting

by the odds" approach (with weights equal to p/(1-p)). The weighting approach allowed Stuart and colleagues to retain the full sample in the analysis; all duals are included as the "treatment" group, and all non-duals are included but weighted relative to their similarity to the duals.

Figure 7 displays the standardized differences between the experimental groups (duals vs. non-duals) on a range of covariates before (hollow circles) and after (solid circles) propensity score weighting. The figure shows that the weighting reduced nearly all of the standardized differences, and after weighting all standardized biases were less than 0.2, a threshold used to indicate adequate balance. It is particularly reassuring that the baseline measures of some of the key outcomes (e.g. unique day counts [UDC] of six types of prescription drugs) are very well balanced after weighting.

Analyzing Observational Data Like Randomized Experiments

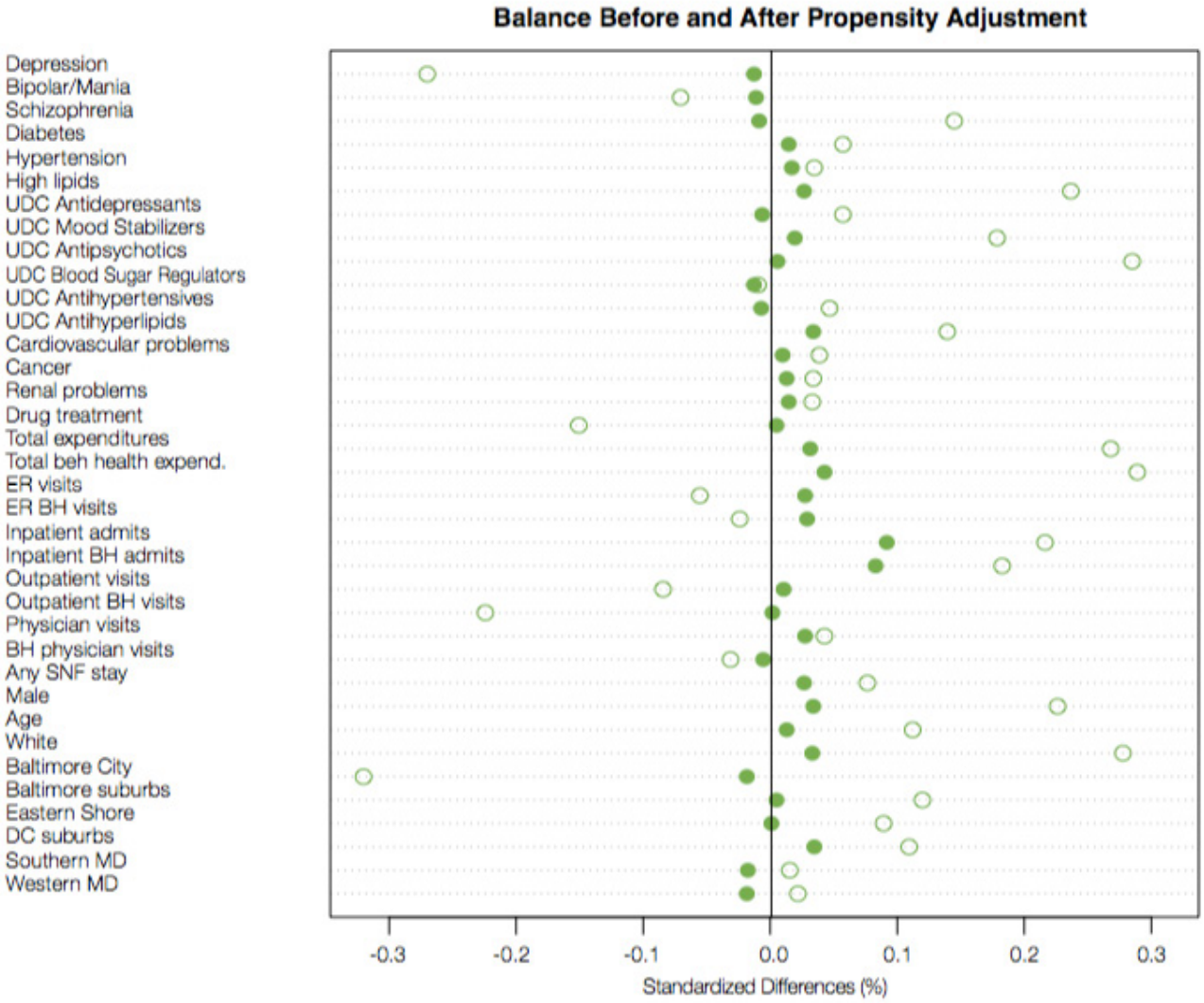
Although not directly related to assessing cause and effect relationships, Forrest and colleagues[31] have shown how methods developed for the analysis of clinical trials can address a number of challenges in analyzing observational data. Adapting the controlled trial simulation by methods of Hernan and colleagues[32] can lead to robust estimates of comparative effectiveness.

For example, consider the case of biologic therapy for Crohn's disease. Agents targeted to reduce TNFα-mediated inflammation are rational therapeutic choices; their efficacy has been demonstrated in adults by an RCT, and the REACH study[33] and others have evaluated single group efficacy of biologics in children. However, there is currently no direct assessment of biologic agents in the pediatric population. And since some patients with moderate to severe Crohn's disease will get better regardless of treatment, a comparison group of non-biologic treated patients is needed to estimate treatment effects. ImproveCareNow (ICN) is a collaborative network of more than 50 pediatric GI practices established in 2007 to improve care for children with inflammatory bowel disease. Its dataset includes a large population of children with CD, with longitudinal follow-up, so provides an opportunity to fill in the gap in evidence about the effect of biologic therapy for Crohn's disease in children.[34]

Forrest and colleagues[35] describe an observational study using ICN data designed to contrast the 6-month outcome experience (disease activity) for patients with new biologic therapy with those not receiving biologic therapy but only usual care. Their results are based on 198 children initiating treatment with new biologics and 1157 trials (a "trial" refers to one child initiating treatment followed for a 6-month observation window) in 680 children in two control groups (biologic users pre-initiation and biologic non-users). The treatment and control groups differ in a number of respects; most importantly, the biologic initiators had a higher probability of colonic involvement and less concurrent medication use at baseline. The groups did not differ significantly in terms of the Pediatric Crohn's Disease Activity Index (PCDAI), an 11-element composite index of disease activity that includes symptoms, exam findings, and lab results.

The primary outcome variables are clinical and steroid-free remission. Forrest and colleagues[36] used multiple analytic methods to seek accurate estimates of the treatment effect. For instance, they calculate the post-test difference in the probability of recurrence during the trial, and prepare empirical cumulative probability plots. Kaplan-Meier plots and a Cox-proportional hazards model that controlled for patient characteristics both indicate a significant difference in the cumulative probability of achieving either clinical response or remission between the

Figure 7. Standardized Differences Between the Experimental Groups on Covariates Before (Hollow Dots) and After (Solid Dots) Propensity Score Weighting



Source: Stuart *et al.*[23]

biologic and control groups. As summarized in Table 2, Forrest and colleagues[37] also conducted a variety of pre- and post-test analyses, both unadjusted and adjusted (difference in difference GEE estimates) for demographic characteristics and medication use at baseline.

Based on these analyses, Forrest and colleagues[38] conclude that multiple analytic methods converge on consistent estimates of effect size. Substantively they find that biologic agents are modestly more effective than other therapies for moderate/severe disease, and that their estimates are consistent with the limited existing prospective data. The number needed to treat (NNT) to avoid one clinical remission is approximately 8. For steroid-free remission the NNT is 5.2, and indeed the reduced steroid use may be a significant benefit of biologic treatment.

Conclusions

When the question is whether an intervention improves outcomes of interest, the second paper in this series[39] illustrates how study design methods can help researchers identify valid results that better balance internal and external validity than RCTs. The current paper complements this by describing how analytical methods for individual-level EHD, including regression approaches, interrupted time series (ITS) analyses, instrumental variables, and propensity score methods, can also be used to address the question of whether the intervention "works."

The two major potential sources of bias in non-experimental studies of health care interventions are that the treatment groups compared do not have the same probability of treatment or exposure

Table 2. Percentage of Trials Achieving Remission and Corticosteroid-free Remission During 26- and 52-Week Follow-up Periods

OUTCOME	DURATION OF FOLLOW-UP	INITIATOR TRIALS		NON-INITIATOR TRIALS	
		% ACHIEVING OUTCOME (95% CI)			
UNADJUSTED					
Clinical remission	26 weeks	54.4	(47.7–61.1)	41.2	(38.2–44.2)
	52 weeks	66.6	(60.3–72.8)	56.2	(53.2–59.3)
Corticosteroid-free remission	26 weeks	47.3	(40.6–53.9)	31.2	(28.4–34.0)
	52 weeks	60.1	(53.7–66.5)	47.5	(44.5–50.5)
ADJUSTED					
Clinical remission	26 weeks	54.8	(47.2–62.4)	40.7	(36.5–45.0)
	52 weeks	67.3	(60.1–74.4)	55.6	(51.1–60.1)
Corticosteroid-free remission	26 weeks	45.6	(38.1–53.1)	30.8	(26.8–34.7)
	52 weeks	58.8	(51.5–66.2)	47.0	(42.5–51.5)

Note: Proportions were adjusted for patient age, gender, and race, disease location, duration, and phenotype, and concurrent medications, all measured at baseline of the trial. Adapted from Forrest and colleagues.[44]

and the potential for confounding by unmeasured covariates. This paper described a range of analytical methods for the analysis of individual data – deriving primarily from statistics and econometrics – that may help to address these problems. These methods include statistical methods such as regression approaches, propensity score methods, instrumental variables, and clinical trial methods.

Although these approaches are very different, they all are based on assumptions about data, causal relationships, and biases. Regression approaches assume that the actual causal relationship between the treatment, outcome, and other variables is properly specified, all of the variables are available for analysis (i.e., no unobserved confounders) and measured without error, and that the error term is independent and identically distributed. The instrumental variables approach requires identifying an instrument that is related to the assignment of treatment but otherwise has no direct on the outcome. Propensity score methods approaches, on the other hand, assume that there are no unobserved confounders. The epidemiological designs discussed also make assumptions, for instance that individuals can serve as their own control.

There are, however, three things that can be done. First, analysts should conduct sensitivity analyses within the assumptions of each method to assess the potential impact of what cannot be observed. It is standard practice in econometrics, for instance, to assess omitted variable bias, and similar analysis would be useful for all of the analytical methods described in this section. The second solution is to analyze the same data with different analytical approaches that make alternative assumptions, and to apply the same methods to different data sets (as Rassen and colleagues[40] did in the PPP study and Yih and colleagues[41] did in their study of intussusception risk after rotavirus vaccination. Finally, different analytical methods, each subject to different biases, can be used in combination and together with different designs, to limit the potential for bias in the final results.

Finally using any of these methods effectively to obtain unbiased estimates knowledge about the setting, the behavior, and the population being studied. In their study of the effect of Medicare Part D, for instance, Stuart and colleagues[42] limited their sample to Maryland residents, assuming that this would help to control for variation in state-level factors and policies. They also focused on patients with diagnoses of schizophrenia, bipolar, or depressive disorders, assuming that the impact of the new drug benefit would be similar for these groups and because the impact on individuals with other health conditions could be quite different.

References

1. Stoto MA, Oakes M, Stuart EA, Stuart L, Priest E, Zurovac J. Analytical methods for a learning health system: 1. Framing the research question. eGEMs (Generating Evidence & Methods to improve patient outcomes). 2016;5(1):28.
2. Stoto MA, Oakes M, Stuart EA, Stuart L, Priest E, Zurovac J. Analytical methods for a learning health system: 1. Framing the research question. eGEMs (Generating Evidence & Methods to improve patient outcomes). 2016;5(1):28.
3. Stoto MA, Oakes M, Stuart EA, Priest E, Savitz L. Analytical methods for a learning health system: 2. Design of observational studies. eGEMs (Generating Evidence & Methods to improve patient outcomes). 2016;5(1):29.
4. Stoto MA, Parry G, Savitz L. Analytical methods for a learning health system: 4. Delivery system science. eGEMs (Generating Evidence & Methods to improve patient outcomes). 2016;5(1):31.
5. Rosenbaum PR. Choice as an alternative to control in observational studies. Statistical Science. 1999; 14(3): 259-304.
6. Rosenbaum PR. Observational studies. 2nd edition. New York: Springer-Verlag; 2002. 377 p.
7. Rubin DB. The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials. Stat Med. 2007 Jan 15; 26(1): 20-36.
8. Stoto MA, Oakes M, Stuart EA, Priest E, Savitz L. Analytical methods for a learning health system: 2. Design of observational studies. eGEMs (Generating Evidence & Methods to improve patient outcomes). 2016;5(1):29.

9. Sauer B, VanderWeele TJ. Supplement 2. Use of directed acyclic graphs, Developing a protocol for observational comparative effectiveness research: a user's guide. AHRQ Publication No. 12(13)-EHC099 [Internet]. Rockville (MD): Agency for Healthcare Research and Quality; 2013 Jan [cited 2014 Dec 17]. Available from: http://www.effectivehealthcare.ahrq.gov/ehc/products/440/1166/User-Guide-to-Observational-CER-1-10-13.pdf
10. Dowd B, Oakes M. Causality beyond simple RCTs. Paper presented at: AcademyHealth Health Services Researcher of 2020: Summit II. Summit on the Future of HSR Data and Methods; 2014 Jun 1-2; Washington, DC.
11. AcademyHealth, Evaluating Complex Health Interventions: A Guide to Rigorous Research Designs. 2017 Jun. Retrieved from: http://www.academyhealth.org/files/AH_Evaluation_Guide_FINAL.pdf
12. Cook TD, Shadish WR, Wong VC. Three conditions under which experiments and observational studies produce comparable causal estimates: new findings from within-study comparisons. Journal of Policy Analysis and Management. 2008; 27(4): 724-750.
13. AcademyHealth, Evaluating Complex Health Interventions: A Guide to Rigorous Research Designs. 2017 Jun. Retrieved from: http://www.academyhealth.org/files/AH_Evaluation_Guide_FINAL.pdf
14. Ross-Degnan D. Observational research methods for a learning healthcare system: interrupted time series methods for natural and quasi-experiments. Presentation at: AcademyHealth Annual Research Meeting; 2013 Jun 23-25; Washington, DC.
15. Fretheim A, Oxman AD, Havelsrud K, Treweek S, Kristoffersen DT, Bjorndal A. Rational prescribing in primary care (RaPP): a cluster randomized trial of a tailored intervention. PLoS Med 2006;3:e134.
16. Ross-Degnan D. Observational research methods for a learning healthcare system: interrupted time series methods for natural and quasi-experiments. Presentation at: AcademyHealth Annual Research Meeting; 2013 Jun 23-25; Washington, DC.
17. Stoto MA, Oakes M, Stuart EA, Priest E, Savitz L. Analytical methods for a learning health system: 2. Design of observational studies. eGEMs (Generating Evidence & Methods to improve patient outcomes). 2016;5(1):29.
18. McClellan M, McNeil BJ, Newhouse JP. Does more intensive treatment of acute myocardial infarction in the elderly reduce mortality? Analysis using instrumental variables. JAMA. 1994 Sep 21; 272(11): 859-866.
19. Baiocchi M, Small DS, Lorch S, Rosenbaum PR. Building a stronger instrument in an observational study of perinatal care for premature infants, Journal of the American Statistical Association. 2010 Dec, 105(492): 1285-1296.
20. Rassen JA, Brookhart MA, Glynn RJ, Mittleman MA, Schneeweiss S. Instrumental variables II: instrumental variable application-in 25 variations, the physician prescribing preference generally was strong and reduced covariate imbalance. J Clin Epidemiol. 2009 Dec; 62(12): 1233-1241.
21. Rassen JA, Brookhart MA, Glynn RJ, Mittleman MA, Schneeweiss S. Instrumental variables II: instrumental variable application-in 25 variations, the physician prescribing preference generally was strong and reduced covariate imbalance. J Clin Epidemiol. 2009 Dec; 62(12): 1233-1241.
22. Rassen JA, Brookhart MA, Glynn RJ, Mittleman MA, Schneeweiss S. Instrumental variables II: instrumental variable application-in 25 variations, the physician prescribing preference generally was strong and reduced covariate imbalance. J Clin Epidemiol. 2009 Dec; 62(12): 1233-1241.
23. Rassen JA, Brookhart MA, Glynn RJ, Mittleman MA, Schneeweiss S. Instrumental variables II: instrumental variable application-in 25 variations, the physician prescribing preference generally was strong and reduced covariate imbalance. J Clin Epidemiol. 2009 Dec; 62(12): 1233-1241.
24. Stuart EA. Matching methods for causal inference: a review and a look forward. Stat Sci. 2010 Feb 1; 25(1): 1-21.
25. Stuart EA, DuGoff E, Abrams M, Salkever D, Steinwachs D. Estimating causal effects in observational studies using Electronic Health Data: challenges and (some) solutions. EGEMS (Wash DC). 2013; 1(3).
26. Lee BK, Lessler J, Stuart EA. Improving propensity score weighting using machine learning. Stat Med. 2010 Feb 10; 29(3): 337-346.
27. Schneeweiss S, Rassen JA, Glynn RJ, Avorn J, Mogun H, Brookhart MA. High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. Epidemiology. 2009 Jul; 20(4): 512-522.
28. Peikes DN, Moreno L, Orzol SM. Propensity score matching: a note of caution for evaluators of social programs. The American Statistician. 2008; 62(3): 222-231.
29. Stuart EA, DuGoff E, Abrams M, Salkever D, Steinwachs D. Estimating causal effects in observational studies using Electronic Health Data: challenges and (some) solutions. EGEMS (Wash DC). 2013; 1(3).
30. Stuart EA, DuGoff E, Abrams M, Salkever D, Steinwachs D. Estimating causal effects in observational studies using Electronic Health Data: challenges and (some) solutions. EGEMS (Wash DC). 2013; 1(3).
31. Forrest CB, Crandall WV, Bailey LC, Zhang P, Joffe MM, Colletti RB, Adler J, Baron HI, Berman J, del Rosario F, Grossman AB, Hoffenberg EJ, Israel EJ, Kim SC, Lightdale JR, Margolis PA, Marsolo K, Mehta DI, Milov DE, Patel AS, Tung J, MD, Kappelman MD. Effectiveness of Anti-TNFα for Crohn Disease: research in a pediatric learning health system. Pediatrics. 2014 Jun;134(1): 37-44.
32. Hernan MA, Alonso A, Logan R, Grodstein F, Michels KB, Willett WC, Manson JE, Robins JM. Observational studies analyzed like randomized experiments: an application to postmenopausal hormone therapy and coronary heart disease. Epidemiology. 2008 Nov; 19(6): 766-779.
33. Hyams J, Crandall W, Kugathasan S, et al; REACH Study Group. Induction and maintenance infliximab therapy for the treatment of moderate-to-severe Crohn's disease in children. Gastroenterology. 2007;132(3): 863–873.

34. Forrest CB, Margolis PA, Seid M, Colletti RB. PEDSnet: how a prototype pediatric learning health system is being expanded into a national network. Health Affairs 2014 Jul;33(7): 1171–1177.
35. Forrest CB, Crandall WV, Bailey LC, Zhang P, Joffe MM, Colletti RB, Adler J, Baron HI, Berman J, del Rosario F, Grossman AB, Hoffenberg EJ, Israel EJ, Kim SC, Lightdale JR, Margolis PA, Marsolo K, Mehta DI, Milov DE, Patel AS, Tung J, MD, Kappelman MD. Effectiveness of Anti-TNFα for Crohn Disease: research in a pediatric learning health system. Pediatrics. 2014 Jun;134(1): 37-44.
36. Forrest CB, Crandall WV, Bailey LC, Zhang P, Joffe MM, Colletti RB, Adler J, Baron HI, Berman J, del Rosario F, Grossman AB, Hoffenberg EJ, Israel EJ, Kim SC, Lightdale JR, Margolis PA, Marsolo K, Mehta DI, Milov DE, Patel AS, Tung J, MD, Kappelman MD. Effectiveness of Anti-TNFα for Crohn Disease: research in a pediatric learning health system. Pediatrics. 2014 Jun;134(1): 37-44.
37. Forrest CB, Crandall WV, Bailey LC, Zhang P, Joffe MM, Colletti RB, Adler J, Baron HI, Berman J, del Rosario F, Grossman AB, Hoffenberg EJ, Israel EJ, Kim SC, Lightdale JR, Margolis PA, Marsolo K, Mehta DI, Milov DE, Patel AS, Tung J, MD, Kappelman MD. Effectiveness of Anti-TNFα for Crohn Disease: research in a pediatric learning health system. Pediatrics. 2014 Jun;134(1): 37-44.
38. Forrest CB, Crandall WV, Bailey LC, Zhang P, Joffe MM, Colletti RB, Adler J, Baron HI, Berman J, del Rosario F, Grossman AB, Hoffenberg EJ, Israel EJ, Kim SC, Lightdale JR, Margolis PA, Marsolo K, Mehta DI, Milov DE, Patel AS, Tung J, MD, Kappelman MD. Effectiveness of Anti-TNFα for Crohn Disease: research in a pediatric learning health system. Pediatrics. 2014 Jun;134(1): 37-44.
39. Stoto MA, Oakes M, Stuart EA, Priest E, Savitz L. Analytical methods for a learning health system: 2. Design of observational studies. eGEMs (Generating Evidence & Methods to improve patient outcomes). 2016;5(1):29.
40. Rassen JA, Brookhart MA, Glynn RJ, Mittleman MA, Schneeweiss S. Instrumental variables II: instrumental variable application-in 25 variations, the physician prescribing preference generally was strong and reduced covariate imbalance. J Clin Epidemiol. 2009 Dec; 62(12): 1233-1241.
41. Yih WK, Lieu TA, Kulldorff M, martin D, McMahill-Walraven CN, Platt R, Selvam N, Selvan M, Lee GM, Nguyen M. Intussusception risk after rotavirus vaccination in U.S. infants. N Engl J Med. 2014 Feb 6; 370(6): 503-512.
42. Stuart EA, DuGoff E, Abrams M, Salkever D, Steinwachs D. Estimating causal effects in observational studies using Electronic Health Data: challenges and (some) solutions. EGEMS (Wash DC). 2013; 1(3).
43. Rassen JA, Brookhart MA, Glynn RJ, Mittleman MA, Schneeweiss S. Instrumental variables II: instrumental variable application-in 25 variations, the physician prescribing preference generally was strong and reduced covariate imbalance. J Clin Epidemiol. 2009 Dec; 62(12): 1233-1241.
44. Forrest CB, Crandall WV, Bailey LC, Zhang P, Joffe MM, Colletti RB, Adler J, Baron HI, Berman J, del Rosario F, Grossman AB, Hoffenberg EJ, Israel EJ, Kim SC, Lightdale JR, Margolis PA, Marsolo K, Mehta DI, Milov DE, Patel AS, Tung J, MD, Kappelman MD. Effectiveness of Anti-TNFα for Crohn Disease: research in a pediatric learning health system. Pediatrics. 2014 Jun;134(1): 37-44.