EDITORIAL

From Medical Records to Clinical Science

Mikel Aickin, PhD; Charles Elder, MD, MPH, FACP

Perm J 2012 Fall;16(4):67-74

http://dx.doi.org/10.7812/TPP/12-047

Abstract

Medical records contain an abundance of information, very little of which is extracted and put to clinical use. Increasing the flow of information from medical records to clinical practice requires methods of analysis that are appropriate for large nonintervention studies. The purpose of this article is to explain in nontechnical language what these methods are, how they differ from conventional statistical analyses, and why the latter are generally inappropriate. This is important because of the current volume of nonintervention study analyses that either use incorrect methods or misuse correct methods. A set of guidelines is suggested for use in nonintervention clinical research.

Introduction

Data are entered into electronic medical record (EMR) systems at an enormous rate, yet the return in clinical research information is meager. There are several reasons for this, at least one of which can be addressed using current knowledge: the relatively limited awareness of the methods that are available for analyzing data in nonintervention settings. The purpose of this article is to explain in nontechnical language the issues in nonintervention clinical research and the methods that might address them. The accompanying technical supplement (available online at: www.thepermanentejournal.org/issues/2012/fall/4911-medical-records.html) explains the same issues in the more complex language of probability.

It is fair to say that the dominant view among medical researchers is that the only way to obtain reliable information about therapeutic effectiveness is through the highly developed technology of the randomized clinical trial (RCT). How this situation came about will be outlined below. The important point here is that this mentality rules out EMR-based research on philosophical grounds, with no attention to the methods that have been developed over the past 30 years for obtaining, in nonintervention situations, nearly the same information that would have been obtained if there had been an intervention. The issues in nonintervention research are indeed complicated, and the solutions are neither simple nor foolproof. Exactly the same could be said of the RCT, but unlike EMR-based research, the development of RCT methods represent a substantial cultural investment that would be put at risk if we were to acknowledge its problems openly. The best way to support the expansion of EMR-based research is to address its problems directly.

This article will develop two important themes. One is that EMRs generally have much larger samples than could ever be obtained in clinical trials. It does not seem to be generally recognized that this fact alone opens the door to methods of analysis that are difficult or impossible to apply in the RCT setting. The second theme is that because the data in EMRs have already been collected, there is ample opportunity to apply multiple methods of analysis to the same data. This is strongly discouraged in RCTs, where the plan of analysis set out before data collection must be followed. Both of these have had a common result: the conventional analysis methods that are used in RCTs have been developed to deal with inadequate sample sizes and the narrow latitude they allow for analysis. The characteristics of the analyses presented here should become more appealing when one accepts that EMR-based research is free of both these limitations.

Why Electronic Medical Record-Based Research Is Difficult

The notion that data collected from clinical practice might inform medical science goes back at least to the 1830s. Pierre-Charles-Alexandre Louis was one of the first to argue for the organized, routine collection of data on patients and treatments, with the evident intent of overthrowing the ossified opinions of the great men of medicine in his day.[1] Louis did nothing to explain how the resulting data should be viewed, nor how they might specifically alter medical practice. Inspired by Louis, Jules Gavarret wrote the first book on biostatistics in 1840. He drew on the "calculus of probability" that had been recently developed by figures such as Laplace and Poisson. The method employed by Gavarret would be difficult to distinguish from the modern technique of statistical confidence intervals. It is a measure of Gavarret's failure in the 19th century that the British founders of modern statistics in the early 20th century were evidently unaware that they were rediscovering his work.

One of the themes in the development of modern biomedical research was the belief that medicine would become more scientific to the extent that it followed the principles of experimental sciences, such as physics and chemistry. Although there were multiple attempts to realize this goal, the first self-conscious use of the methods that we now associate with RCTs was the trial of streptomycin for tuberculosis, designed and carried out by Sir Austin Bradford Hill shortly after the Second World War. It soon became apparent to the pharmaceutical industry that the RCT was admirably suited to generating the information required for regulatory approval, and virtually all subsequent methodological development of the RCT was dominated by this application. As the

Mikel Aickin, PhD, is a Professor in the Department of Family and Community Medicine at the University of Arizona in Tucson. E-mail: maickin@comcast.net.
Charles Elder, MD, MPH, FACP, is the Physician Lead for Integrative Medicine at Kaiser Permanente Northwest, and Affiliate Investigator at the Center for Health Research in Portland, OR. E-mail: charles.elder@kpchr.org.

RCT became more rigidly defined and easier to carry out in a routine fashion, the idea became fixed that the RCT was the method of medical science, and that all other approaches were flawed.

From the standpoint of finding out whether a therapy works, the key feature of the RCT is that the treatments are assigned by the researchers. This is to be done in an understandable way that can be appropriately accounted for when the data are analyzed. This feature of the RCT stands in stark contrast to the practice of medicine. In a research study, treatments are assigned to learn which ones are better than which others, not to maximize the benefits to the participants. When researchers speak of intervention in this setting, they are talking about research interventions, not medical or clinical interventions (see Sidebar: Randomized Clinical Trial versus Electronic Medical Records Research).

The data in EMRs are collected during everyday medical practice, in which treatments are prescribed with the intent of producing benefits. From the research perspective, this is nonintervention research, meaning that there are no research interventions. The fundamental problem this creates is that it is not clear what kinds of data analysis are appropriate. The advantage of routine use of certain statistical methods that have become automatic in RCTs is now lost. Unfortunately, many researchers analyzing data from EMRs use precisely the same approaches that are appropriate to RCTs. Depending on the circumstances, these researchers may confuse the therapeutic situations they study, and at worst they can produce misleading results.

Measurement. To understand the primary problem of analysis in nonintervention studies, it is useful to have some special terminology. We will use capital letters to emphasize that we attach specific meanings to certain kinds of measurements. The first of these is Treatment, which refers to two or more different actions that could be taken with therapeutic intent. We also have Outcomes, which consist of one or more ways of measuring therapeutic benefits. The basic purpose of clinical science is to learn about the causal effects of Treatment on Outcome. We will have something to say about causation here, because this issue cannot be avoided (Figure 1).

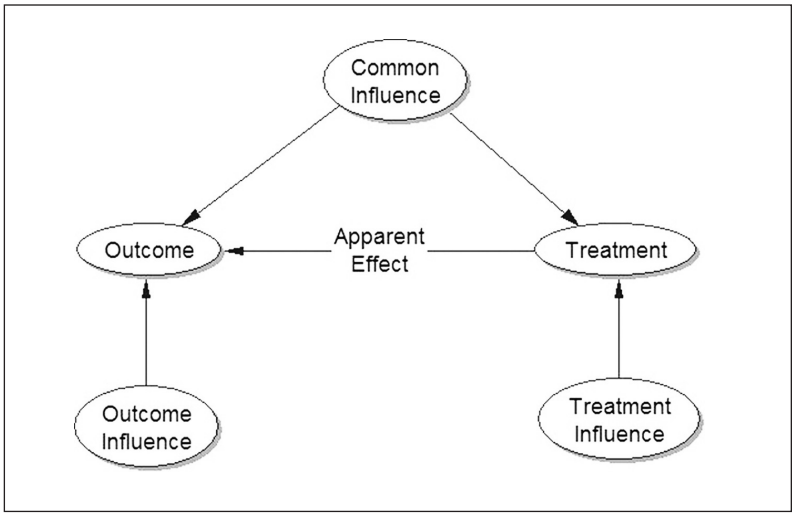

Figure 1. Roles of variables that are important in the analysis of nonintervention studies.

Common Influences. One might think that it is only necessary to observe Treatments and Outcomes in practice, to see how they are related, and then proceed accordingly. This was more or less Louis's approach. The chief threat to the validity of this path is Common Influences. A Common Influence is a factor that influences both the Treatment selection and the Outcome. A *factor* is some underlying condition, action, or state that can be measured. The idea is that as we move through a clinical population we will see the Common Influences changing from one patient to the next, and associated with these changes are changes in both the Treatments assigned to the patients and their subsequent Outcomes. If the Common Influences effectively select patients who are likely to have good clinical courses subsequent to specific Treatments, then a simplistic analysis of Treatments and Outcomes will ascribe the beneficial results to the Treatments, rather than to the action of the Common Influences. This is the primary problem with research in nonintervention situations (see Sidebar: Clinical Perspective: Common, Treatment, and Outcome Influences).

Analysis in the Absence of Intervention. It will come as no surprise that almost all of the methods for analyzing nonintervention data come from the social sciences. The opportunities for interventions in the social sciences, unlike medicine, are few and far between. Even when interventions can be mounted, it is still exceedingly difficult to eliminate Common Influences. It is, however, often possible to identify what some of the Common Influences are, and we will see that this is crucial in the clinical science setting as well.

Randomized Clinical Trial versus Electronic Medical Records Research

Randomized Clinical Trial

Question: Which treatment, A or B, is better for all qualifying patients?

Method: Emulate a scientific experiment (clear definitions of patient population, therapeutic maneuvers, intervention rules, outcomes, and efficacy analysis).

Electronic Medical Records Research

Question: How can the better treatment, A or B, be matched to an individual patient?

Method: Intensive investigation of naturally occurring clinical data (inclusion of all patients, multiple definitions of treatment and outcome, attention to patient subgroups, effectiveness analyses designed to reduce confounding).

Outcome and Treatment Influences

There are two other categories of measurements that need to be included in this discussion. The first is Outcome Influences. These are factors that influence the patient's medical results, independently of the Common Influences. Recall that a Common Influence affects the outcome, but it must influence the Treatment selection as well. An Outcome Influence is distinguished by the fact that it does not influence Treatment. The second category is Treatment Influences. These are factors that influence the selection of Treatment, independently of the Common Influences. Again, like the Common Influences they influence Treatment; they do not, however, influence the Outcome.

It is convenient for the purposes of exposition to consider the special case in which there are only two Treatment options to be compared, say A and B. It is natural to imagine that these Treatments occur with certain probabilities in the clinical population, and that we could estimate those probabilities on the basis of an EMR sample. It is also reasonable to imagine that we could refine this procedure by estimating the probability of A (or B) for patients with specified characteristics. The mathematical way to do this is to express these probabilities as functions of patient measurements, using customary methods of statistical modeling. Some of these measurements might be factors that do not change (like male or female gender), whereas others (like blood pressure) could depend on the time of measurement. The point is that we can generally develop formulas for predicting Treatment, based on any selected class of measurements. Any such formula that correctly expresses the probability of Treatment given a battery of predictors is a *propensity score*. It measures the propensity for treatment, expressed as the probability of Treatment A given the values of the predictor variables.

Now consider the following sequence of analytical actions. First, develop a propensity score for A based on some set of prediction variables. Since the probability of B is just 1 minus the probability of A, only one propensity score is necessary. (This is not true when there are more than two Treatment options.) Second, stratify the patient population into groups having the same propensity score. (Assume there are no problems here; if one were to compute the propensity scores to the last decimal, then presumably no two patients would share a score, and we want to rule out this triviality.) Third, in each of these propensity groups (patients with equal propensity scores) investigate the relationship between Treatment and the battery of prediction variables that were used to determine propensity score. You will find that within each propensity-score group, Treatment is independent of all these prediction variables. This important finding was discovered by Donald Rubin and Paul Rosenbaum.[2]

Here is how this helps with the problem of nonintervention. Suppose that we could identify and measure all Common Influences. Suppose further that we use precisely this set of variables to construct propensity scores. According to the Rubin-Rosenbaum result, within each propensity-score group, Treatment is independent of the Common Influences. This means that in the small sample of a propensity group, the Common Influences have ceased to act as Common Influences, because they no longer influence Treatment.

It is often possible to form many groups of patients, including perhaps some small groups—*matched comparison groups*—within which the patients are very similar if not identical with respect to the battery of prediction factors.

This approach for doing away with Common Influences is so attractive that propensity scores have been endowed by various authors with properties they do not have. Indeed, there is considerable confusion in the biomedical literature about what propensity scores do, and how to use them, with the level of confusion depending on the credulity of each author. These mistaken views are 1) that the purpose is to accurately predict treatment, 2) that logistic regression is the only relevant statistical model, 3) that Treatment Influences should be included, and 4) that controlling for a propensity score is the same as controlling for all of its component variables. In order not to interrupt the flow of ideas at this point, these issues are developed further in the Sidebar: Misapprehension About Propensity Score Analysis.

Rubin and Rosenbaum developed the propensity score for the modest sample sizes that are typical of RCTs. One of the advertised benefits of propensity scores is that in the analysis one only needs to condition on a single value (the propensity score), irrespective of how many variables went into it. The situation is completely

Clinical Perspective: Common, Treatment, and Outcome Influences

Several categories of measurement must be considered when analyzing data from nonintervention studies.

Consider a study assessing the impact of a lifestyle modification class on hypertension. Assume that it is possible to discern from the electronic medical record whether a given patient attended the class. The research question is, does attending the class improve blood pressure?

A *Common Influence* is a factor that might influence both the Treatment selection (whether or not the patient signed up for and attended the class) and the Outcome (whether the patient's blood pressure improved). A Common Influence in this case might be patient motivation. More motivated patients may be more likely to sign up for and to attend the lifestyle class, and they may also be more likely to try dietary, exercise, and stress reduction recommendations even if they didn't attend the class, resulting in improved blood pressure. Here, either some indicator of motivation would have to be found in the EMR, or motivation would need to be ascertained through patient contact.

An *Outcome Influence* is a factor that influences the outcome but not the Treatment selection. An example of an Outcome Influence in this case might be a cause of secondary hypertension, such as renal artery stenosis. Such a factor could influence changes in blood pressure independently of whether patients attended the lifestyle class, but it would not influence the choice to attend the class or not.

A *Treatment Influence* is a factor that influences Treatment selection but not Outcome. An example could be neighborhood of residence. Individuals living farther away from the class location may be less likely to sign up and attend, but residence would have no direct influence on future blood pressure.

Misapprehensions About Propensity Score Analysis

To Predict Treatment. The first misapprehension about propensity scores is that their purpose is to predict Treatment as accurately as possible. It is easy to see that this cannot be true. If one were able to predict Treatment perfectly using a battery of prediction variables, then each propensity-score group would have either every patient on Treatment A or every patient on Treatment B. There would be no basis for comparing outcomes within a propensity-score group, and so rather than eliminating the Common Influence problem, the analysis would have created a far greater problem.

The fact is that the variables that should determine the propensity score are the Common Influences, neither more nor less. If one were to mistakenly include an Outcome Influence in the propensity score, this would probably do no harm, although it is difficult to say in general what the consequences would be. On the other hand, including a Treatment Influence can be a serious mistake. The reason is that if an analysis is conditional on a propensity score, then it is also partially conditional on the Treatment Influence that was included in the propensity score. This means that the analysis might build in an association between Outcome and Treatment Influence, at least to some degree. But this amounts to converting a Treatment Influence into a Common Influence, precisely the opposite of what is intended.

Another type of variable that must be excluded from a propensity score is a Mediator. This is a variable that explains the mechanism by which the Treatment produces its effect. The idea is that the Treatment first alters the value of the Mediator, and then the change in the Mediator directly influences the Outcome. It is widely recognized that including Mediators in statistical models can reduce, obliterate, or even reverse a valid treatment effect, and presumably the same warning applies to their inclusion in propensity scores.

The Role of Logistic Regression. The second misapprehension is that since one only needs to predict Treatment A, a logistic regression is indicated. This is because logistic regression is the most frequently used method for assessing a yes/no outcome (eg, Treatment A versus Treatment B). The Rubin-Rosenbaum result only applies to the correct formula for predicting A in terms of the prediction variables. It does not apply to a prediction formula that differs from the correct one. If the correct propensity score is not of logistic form, then the Rubin-Rosenbaum result may not hold for a logistic regression approximation to the true form. There are two reasons this is concerning. Practical research results suggest that using the wrong form for the propensity score can have serious consequences for the analysis.[1] Secondly, one will typically try different batteries of prediction variables, but then the fitted logistic regressions are inconsistent with each other. (If you drop a variable from one logistic regression, then this should lead to a propensity score averaged over the dropped variable and conditional on all the retained variables, but this will never be of logistic form.) This inconsistency cannot occur if the correct form for propensity is used.

Pooled Analysis. The third misapprehension is that a pooled analysis is appropriate. The theory says that the beneficial effects of propensity-score matching happen within propensity groups, so group-level results should be used for the analysis. That is, the group becomes the unit of analysis. Many investigators obtain the propensity groups by matching patients from the two treatment groups. They then pool all the patient data and perform an analysis that would be appropriate if a randomized trial had been done. One problem with this strategy is that the theory does not imply that Common Influences will be independent of Treatment in the pooled sample. In fact, it can be shown that the only way this can happen is if the propensity score is independent of Treatment, which is precisely what would not happen if it were correctly defined. (Judging from the literature, this is widely unknown, so a demonstration is provided in the accompanying technical supplement available online at: www.thepermanentejournal.org/issues/2012/fall/4911-medical-records.html.) The second problem with pooling is that by matching patients on the basis of their propensity scores, a dependency is created between the members of each pair. The validity of the pooled analysis depends on the patients being independent of each other. This condition is violated by the induced correlation between matched patients, raising an important issue about whether the results of the pooled analysis are correct. These observations are relevant because the pooling fallacy is exhibited throughout the published applications of propensity scores.

Controlling for All Variables. The fourth misapprehension is that controlling for a propensity score is the same as controlling for all of the variables that went into it. In the common logistic case, conditioning on the propensity score is exactly the same as conditioning on a certain linear combination of the battery of prediction variables. It is obvious that fixing a linear combination is not the same as fixing each individual variable. Researchers misguided in this regard seldom check their propensity groups to see whether they are truly homogeneous with regard to Common Influences. If the logistic propensity score contains many variables, it is likely that many of them will not be matched across patients in a propensity group, which may undermine the analysis.

Reference

1. Guo S, Fraser MW. Propensity Score Analysis. Thousand Oaks, CA: Sage Publications; 2010.

different in EMR-based research, which typically has participant-rich sample sizes. It is often possible to form many groups of patients, including perhaps some small groups, within which the patients are very similar if not identical with respect to the battery of prediction factors. We will call these *matched comparison groups* (MCGs), because all of their members are very much alike regarding the matching variables. One can then easily estimate the probability of Treatment A in each small MCG. The nonparametric propensity score is then computed just by identifying which group a patient falls into and assigning the corresponding probability of A.

Because this nonparametric approach involves many fewer assumptions than the logistic (or any other parametric) approach, it has an exceedingly valuable advantage to be exploited in EMR-based research. But once one has formed MCGs that are alike with respect to the Common Influences, the reason for computing and using a propensity score vanishes. The suggestion is, then, that one should record the Treatment effect within each homogeneous MCG, study how the Treatment effects vary by characteristics of MCGs, and then combine them or model them, as indicated by the findings. In other words, the excessive amount of modeling that goes on in an RCT (or any other patient-poor situation) is usually because of the relatively small number of patients in the study. This is a false economy in a patient-rich environment (see Sidebar: Clinical Perspective: Propensity Scoring and Matched Comparison Groups).

The Problem of Causation

David Hume so thoroughly scared scientists about the idea of causality that the natural tendency of most researchers is to run whenever they hear the word. We believe that there is a very simple way around this in biomedicine, and in fact an appropriate consideration of causality clarifies rather than obscures the analysis of nonintervention data.

We have been careful to use the word Influences in this discussion rather than Causes, because we do not want to make any assumptions about the nature of these influences. We are willing to say that Common Influences, Treatment Influences, and Outcome Influences are associated

with Treatment and/or Outcome without having to identify the nature of the association. But we definitely want to know what Treatment causes what Outcome.

Prediction

One place to start this discussion is in clinical practice. Suppose that, following Louis, we carefully collect Treatment and Outcome data in a given clinic for one year. If the clinic does not change its method of operation and continues seeing patients from the same population, then we would have a good basis for saying what the relationship between Treatment and Outcome would be the next year. It is because the clinic will work the same way and on the same type of patients, and because the Treatment → Outcome relationship is causal, that our prediction is justified.

In practice, the way we apply the idea of causation is to assert that some process that happens under one set of circumstances will continue to happen under some other set of circumstances. This means that the phenomenon we observe is *portable.* When engineers test the strength of a building material in a laboratory (one set of circumstances), they fully expect that the material will have the same properties when it is used to build something (a separate set of circumstances). In the clinical example, the reason for observing the Treatment → Outcome relationship is to modify it if there is an indication to do so. If Treatment A seems to work better than Treatment B, then next year we should use A more than B, and things will improve. This will only happen, however, if the Treatment → Outcome relationship is causal, meaning that it can be ported from the first year to the next.

Causation in Randomized Clinical Trials

Randomized drug trials apply the principle of portable causation in the following way. The patient pool is randomized into two groups, one of which receives A, the other B. Group outcomes are summarized and compared at the end of the trial. The winning treatment is then recommended for everyone. The causal assumption is that the pattern seen in the trial is portable to clinical practice. The objections that are often raised here are 1) RCTs rarely sample the patient population in any meaningfully representative sense, and 2) treatment administered in a trial may differ in important ways from the "same" treatment administered in an ordinary clinical setting. Both of these arguments challenge portability, suggesting that causation is in fact a necessary assumption for an RCT, not a magical consequence of randomization, as is frequently claimed. The counterargument is that random selection of Treatment eliminates all potential Common Influences, whether they are known or not. This means 1) the search for Common Influences for a propensity score is unnecessary, and 2) Treatment is independent of potential Common Influences in the entire sample, not just in narrow propensity groups or MCGs.

Causation in Electronic Medical Records Research

EMR-based studies generally preclude facile causation arguments. If Treatment → Outcome relationships are observed, but they are partially produced by Common Influences, then the same pattern may not result if the method for selecting Treatments is changed. Thus, if identifying winning treatments is viewed as the purpose of medical research, we must consider the fact that the superior Outcomes observed with Treatment A may not be repeated when the Common Influences change. This is the core argument for removing Common Influences in EMR-based research. Researchers want to identify a Treatment → Outcome causal relationship that can be ported to a similar but different set of circumstances, but it is difficult to argue that this has been achieved unless Common Influences have been addressed. This is the fundamental reason why researchers using RCT methods of analysis in an EMR-based setting may be creating problems instead of solving them (see Sidebar: Clinical Perspective: Causation and Adjustment).

Why Not Simply Adjust?

The argument for propensity scores or MCGs may seem excessively complicated, raising the issue, why can we not use simple, conventional methods of analysis? Aren't these methods a far easier means of accomplishing the same thing?

Clinical Perspective: Propensity Scoring and Matched Comparison Groups

On the basis of electronic medical record (EMR) data, it is possible to predict whether a patient will be assigned a certain treatment. For example, consider an EMR-based study assessing outcomes for men with clinically localized prostate cancer undergoing either radical prostatectomy or radiation therapy. Retrospective analysis of EMR data might suggest that certain characteristics, such as age, level of education, income, and tumor staging, predict which treatment will be assigned. Then, given the values for each of these factors (four in this case) for an individual patient, we can calculate the probability of the patient's being assigned one treatment or the other. The formula for this probability is known as a *propensity score*.

Participants receiving different treatments can then be matched according to propensity score. For example, for each patient in the prostatectomy group, a radiation therapy patient with the same propensity score would be included.

Suppose we were to then stratify the sample into groups having the same propensity score. It has been shown that, *within each propensity-score group,* treatment assignment is independent of each of the individual prediction variables. So, in our example, within each group of patients with identical propensity scores, treatment assignment will be statistically independent of age. This approach, then, allows us to effectively account for potential Common Influences in the setting of a nonintervention EMR-based study.

One drawback of this approach is that patients within a propensity-score group may differ drastically with regard to individual variables contributing to the propensity score. For instance, there could potentially be great heterogeneity of income for those assigned prostatectomy versus radiation, despite identical propensity scores. However, in EMR-based research, this pitfall can potentially be avoided by forming matched comparison groups of patients with similar or identical values for the entire battery of prediction variables. This is a unique advantage of EMR-based research because of the potential for very large samples.

By "conventional analysis," we mean fitting customary linear models. An example is linear regression, in which the Outcome is regressed on a Treatment indicator and the presumed set of Common Influences. Virtually all models used in biomedicine are variants of linear regression. For example, logistic regression is also a linear model in which the natural log of the probability of an Outcome success is linear in the explanatory effects. The same may be said of proportional hazards modeling for survival, and a variety of other, more complex models.

The essential difficulty with any kind of modeling is that models only summarize the patterns observed in the data; they do not generally have the capacity to reverse flaws in the study design. For example, if one uses a method of sampling a population that is unrepresentative (it persistently oversamples some kinds of people and undersamples others), then taking the average of an Outcome and computing a confidence interval does not overcome the poor sampling in the design; it simply portrays the consequences. In the same way (albeit with more complexity), a regression equation does not have the ability to correct unwanted features of a design. In common language, when we include Common Influences in a regression equation, it is often said that we are adjusting for them. But it is unclear what this actually means. It certainly does not mean adjustment in the usual epidemiologic sense (to summarize outcomes in strata, then pool using stratum probabilities from a standard population), in which the intent is obvious and the method transparent.

The language that is usually used to interpret regression equations contributes to the confusion. When a regression coefficient is interpreted, it is often said that it captures the effect of changing the corresponding variable by 1 unit, while holding all other variables fixed. Although this is correct in the language of mathematics, it need not be correct in the data on which the regression is based. In a typical EMR study, each patient receives one of the available Treatments. No Treatments are changed, and so one must search elsewhere for the meaning of "changing Treatment." Given how the regression equation is constructed, "changing" a Treatment means proceeding from a patient who has one Treatment to another patient who has a different Treatment. If there are Common Influences, then they will tend to change too (otherwise they would be independent of Treatment, and not be Common Influences). It is thus meaningless to talk about changing Treatment while holding Common Influences constant, because this is not possible with the data that the regression summarizes. Even more dangerous, however, is the implication of the conventional mathematical interpretation of a regression coefficient, that it captures the causal effect of changing treatment. This is because it is a short step from saying, "the effect of changing Treatment while holding Common Influences fixed" to saying "the causal effect of Treatment."

The whole point of propensity scores or MCGs is to change the basis of the analysis, to recover to the extent possible the independence of Treatment and Common Influences that would have been achieved directly by randomization. The price to be paid for this is that the freeing of Treatments from Common Influences only happens in patient subgroups, either propensity-score groups or MCGs. This beneficial effect does not spread out over the entire patient sample and does not justify the fitting of overall models, as would be appropriate in an RCT.

If one is determined to fit an overall model to the entire pooled sample, then it appears to be necessary to make some assumptions about how the treatments were selected. Heckman[3] laid the groundwork for such an approach in economics. His method has been adapted to apply to treatment groups in medical research. One of the simplest versions starts with a logistic regression of Treatment on the presumed Common Influences. This gives a propensity score p for each patient. In a regression model for the Outcome, an additional term is added: the product of the treatment indicator (t, a 0/1 variable) and a function of the propensity score. Specifically, the added term is: $(t - p)\cdot[-p\cdot\ln(p) - (1-p)\cdot\ln(1-p)]/p(1-p)$.

Somewhat surprisingly, this simple change in the analysis often results in a substantial reduction in the bias in the treatment effect estimate, at the price of a decrease in its precision. The additional term in the previous paragraph has the form of an interaction between the treatment residual (the part of the treatment not predicted by the propensity score) and a function of the propensity itself, which is of course a function of the Common Influences. The success of Heckman's procedure rather strongly suggests that performing an analysis that simply injects the propensity score directly into an explanatory linear model is not the correct approach, which is again concerning because of the frequency with which this maneuver appears in the literature.

Clinical Perspective: Causation and Adjustment

Consider the example of a randomized clinical trial (RCT) comparing drug A with placebo for management of depression. Suppose that the trial shows an improvement in depression scores for the treatment group, one that is both statistically and clinically significant. The assumption is that the relationship is *causal,* and that the treatment effect of drug A will thus be *portable* to other patient populations and clinical settings.

In reality, however, the sample recruited to such a clinical trial may not be broadly representative. For instance, if participants were recruited from psychiatric practices and then treated at a research clinic, all in 2008, results may not be generalizable to patients presenting to and treated in primary care settings in 2012. Thus, when interpreting RCT results, we may view portable causation as the underlying assumption, rather than the inherent consequence of randomization.

Statistical modeling cannot guarantee a portable result, and adjusting for covariates in a regression equation cannot correct shortcomings in the study design. In our example, then, primary care physicians in a community clinical practice may have difficulty applying the results of this trial to their patients.

An electronic medical records-based study of the same drug may have much larger samples, allowing for alternative statistical approaches. Treatment effects can be estimated for each homogeneous, matched comparison group, potentially providing insight into which types of patients are most likely to benefit from the intervention.

Summary: Steps for Conducting an Electronic Medical Research Study

The considerations we have laid out can be used in various ways for different types of EMR-based studies. We do not want to create a new orthodoxy for EMR research, because that is the last thing we need. But it does seem worthwhile to propose steps that should be employed unless there are reasons not to do so.

1. *Identify the Common Influences.*
   There are many ways to detect statistical associations, so there are many ways to implement this step. It would certainly be appropriate to fit regression models (of various types) in which Outcomes were explained by Treatments and candidate Common Influences, and Treatments were simultaneously explained by the same Common Influences. In the case of a new Treatment compared to usual care, it might be more appropriate to look for factors that influence Outcomes among the usual care patients, and then turn to the whole sample to see whether they also influence selection of the new Treatment. In any case, a key point here is to remove Treatment Influences from the subsequent computation of propensity scores or forming of MCGs. Because of the importance of this step, previous research or clinical opinions may make critical contributions. Because analysis is potentially sensitive to the choice of Common Influences, multiple possibilities might be retained for subsequent analyses.
2. *Form propensity groups or MCGs.*
   For each possible selection of Common Influences, create nonparametric MCGs if possible; failing this, create parametric propensity groups. A variety of methods for doing this have been developed,[4] although they have been oriented toward settings with relatively few patients. The results of Step 1 should inform this process. In particular, they should suggest an appropriate form for the propensity score. Again, because of the sensitivity of the analysis, it may be wise to use several forms for the propensity score.
3. *Estimate a Treatment effect in each homogeneous patient group.*
   Customary estimates, such as averages or fractions, will usually be appropriate. However, if the patient groups are large enough, more complex models might be used. In particular, if Outcome Influences vary in a patient group, then it seems wise to "adjust" for them in some fashion. (This raises the issue of whether Outcome Influences should be used for matching in MCGs; this question needs to be decided on a case-by-case basis.)
4. *Evaluate whether Treatment effects vary over homogeneous patient groups.*
   With one Treatment effect estimate per patient subgroup, the subgroups themselves can be regarded as the units of analysis. Each subgroup has a set of measured characteristics (the values of the Common Influences, and possibly Outcome Influences), which can be related to Treatment effects using customary statistical methods. The reason this makes sense is that the Treatment effects have presumably been freed from the effects of Common Influences, so they represent genuine within-subgroup causal effects, and the issue is to see whether these effects vary in some systematic fashion across different homogeneous patient subgroups.

For an example following most of these steps, see Newgard et al.[5] Similar approaches can be found in Stukel at al,[6] Sun et al,[7] Smeeth et al,[8] Rutten et al,[9] Polsky et al,[10] and Pollack et al.[11] Not all of the latter studies were equally careful to exclude Treatment Influences from propensity scores.

The presentation of final results depends on Step 4. If there is substantial homogeneity of Treatment effect across patient subgroups, then a very simple story can be told. This is in some ways the equivalent of a conventional RCT, without the cumbersome analysis to compensate for the absence of randomization. If Treatment effects differ in understandable ways across MCGs, then this may be important information that deserves careful presentation. Here is one case in which EMR research can clearly outdo the RCT, taking advantage of a larger number of patients to leverage more patient-specific clinical recommendations. In the worst scenario, the Treatment effects vary in ways that are neither explainable nor ignorable. In this case, one has to consider whether there are substantial problems in the source EMR data that have not been taken into account, despite the best attempts to do so.

An important theme in this summary is the employment of multiple strategies. This is in stark distinction to RCT analyses, where the conventional wisdom is that one should plan all analyses in advance and not deviate from that plan even in the face of considerable evidence that it can be improved. In nonintervention research, each specific approach to analysis has its own strengths and weaknesses, and it is only by trying several of them that one builds confidence in the final results.

Conclusions

There are methods for analyzing data from nonintervention studies that attempt to reveal what an intervention study would have found. These methods are more complicated, primarily because of the requirement that Common Influences be explicitly identified, rather than ignored, as is customary in RCTs. Ordinary multivariable and multivariate statistical methods are appropriate in the search for Common Influences, but they are generally inadequate for the purpose of final analysis, where MCGs or propensity-score groups must be used. A significant advantage of EMR-based research is the possibility of increasing the therapeutic focus of patient-specific treatment. This and other benefits of EMR research are consequences of the generally large patient samples and the fact that EMR data directly reflect clinical practice, which should be the goal of clinical science.

As suggested in the Introduction, the concepts of this essay are not new; they are simply underappreciated in EMR-based research. William Cochran, one of the statistical pioneers of methods for nonintervention studies, formulated many of these ideas, although he used different language and some approaches had not yet been developed when he did his work, in the 1950s and 1960s. His student Donald Rubin, who worked in the 1970s and 1980s, is responsible for many of those approaches.[4,12] The dominance of the RCT in biomedical research has fostered the view that its methods, rather

than a response to its limitations, reflect universal principles that should be followed in all areas of clinical science. In the case of EMR-based research, this view is neither true nor helpful. ❖

Disclosure Statement

This research was partially supported by a grant from the National Institutes of Health(RC1AT005715). The views expressed here are solely those of the author. The author(s) have no other conflicts of interest to disclose.

Acknowledgment

Leslie Parker, ELS, provided editorial assistance.

References

1. Matthews JR. Quantification and the quest for medical certainty. Princeton, NJ: Princeton University Press; 1995.
2. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika 1983 Apr;70(1):41-55. DOI: http://dx.doi.org/10.1093/biomet/70.1.41
3. Heckman J. Sample selection bias as a specification error. Econometrica 1979 Jan;47(1):153-61. DOI: http://dx.doi.org/10.2307/1912352
4. Rubin DB. Matched sampling for causal effects. Cambridge, UK: Cambridge University Press; 2008. DOI: http://dx.doi.org/10.1017/CBO9780511810725
5. Newgard CD, Hedges JR, Arthur M, Mullins RJ. Advanced statistics: the propensity score—a method for estimating treatment effect in observational research. Acad Emerg Med 2004 Sep;11(9):953-61. DOI: http://dx.doi.org/10.1197/j.aem.2004.02.530
6. Stukel TA, Fisher ES, Wennberg DE, Alter DA, Gottlieb DF, Vermeulen MJ. Analysis of observational studies in the presence of treatment selection bias: effects of invasive cardiac management on AMI survival using propensity score and instrumental variable methods. JAMA 2007 Jan 17;297(3):278-85. DOI: http://dx.doi.org/10.1001/jama.297.3.278
7. Sun P, Wang R, Jacober S. The effectiveness of insulin initiation regimens in patients with type 2 diabetes mellitus: a large national medical records review study comparing a basal insulin analogue to premixed insulin. Curr Med Res Opin 2007 Dec;23(12):3017-23. DOI: http://dx.doi.rg/10.1185/030079907X242845
8. Smeeth L, Douglas I, Hall AJ, Hubbard R, Evans S. Effect of statins on a wide range of health outcomes: a cohort study validated by comparison with randomized trials. Br J Clin Pharmacol 2009 Jan;67(1):99-109. DOI: http://dx.doi.org/10.1111/j.1365-2125.2008.03308.x
9. Rutten FH, Zuithoff NP, Hak E, Grobbee DE, Hoes AW. Beta-blockers may reduce mortality and risk of exacerbations in patients with chronic obstructive pulmonary disease. Arch Intern Med 2010 May 24;170(10):880-7. DOI: http://dx.doi.org/10.1001/archinternmed.2010.112
10. Polsky D, Eremina D, Hess G, et al. The importance of clinical variables in comparative analyses using propensity-score matching: the case of ESA costs for the treatment of chemotherapy-induced anemia. Pharmacoeconomics 2009;27(9):755-65. DOI: http://dx.doi.org/10.2165/11313860-000000000-00000
11. Pollack M, Seal B, Joish VN, Cziraky MJ. Insomnia-related comorbidities and economic costs among a commercially insured population in the United States. Curr Med Res Opin 2009 Aug;25(8):1901-11. DOI: http://dx.doi.org/10.1185/03007990903035505
12. Rubin DB. Estimating causal effects from large data sets using propensity scores. Ann Intern Med 1997 Oct 15;127(8 Pt 2):757-63. DOI: http://dx.doi.org/10.1017/CBO9780511810725.035

Counting

Some seem to have been misled by the term numerical system, which has been said to be that of M Louis. They seem to have thought that his peculiarity consists in this merely, that he counts … . We call some experienced, scientific. Is it not by comparing individual cases, by adding what they have observed in one to what they have observed in another, by counting, that they have become so?

— *Researches on the Yellow River,* George Cheyne Shattuck Jr, MD, 1813-1893, American physician