Publication - Statistical Analysis and Data Mining 2014; 7: 376-384.

Fair Treatment Comparisons in Observational Research

Kenneth K. Lopiano, Robert L. Obenchain, S. Stanley Young

Abstract

The proliferation of electronic health records, driven by advances in technology and legislative measures such as the Affordable Care Act, is leading to an explosion of passively collected administrative and medical data. The observational data collected in electronic health records presents exciting challenges and opportunities for researchers interested in comparing the effectiveness of different treatment regimes and, as personalized medicine requires, estimating how effectiveness varies from one subgroup to another. In this paper, we propose an initial simple step in a broader analysis framework using observational data to cluster patients in pretreatment covariate space and carry out comparisons of treatments within the clusters to obtain estimates of heterogeneous treatment effects. We consider such comparisons fair as they are made only among highly similar patients. A simple example of Simpson's Paradox is used to show an overall average treatment effect, that marginalizes over covariate space, can be misleading. In contrast, we present an alternative definition using a single, shared marginal distribution defining a fair overall treatment comparison. We also argue that overall treatment comparisons should no longer be the focus of comparative effectiveness research because treatment effectiveness can vary across patient sub-populations. In the spirit of the now ubiquitous concept of personalized medicine, estimating heterogeneous treatment effects in clinically relevant subgroups will allow for, within the limits of the available data, fair treatment comparisons that are more relevant to individual patients.
Keywords:observational data, fair conditional comparisons, common support, re-marginalization, local control, Simpsons Paradox

1 Introduction

Randomized clinical trials and other carefully designed experiments to assess the effectiveness of a treatment or intervention typically report an estimated average treatment effect. In practice, however, clinicians observe that treatment effectiveness will vary from patient-to-patient. Some patients have strong positive responses to a given treatment while others have negligible or even negative responses. Are these

differences purely random, or are they due to observed (or unknown) pretreatment differences among patients? The proliferation of electronic health records, driven by advances in information technology and legislative measures such as the Affordable Care Act, is leading to an explosion of interest in using "real-world" data to improve health care. The observational data collected in electronic health records presents both challenges and opportunities for researchers interested in comparing the effectiveness of different treatment regimes and estimating how effectiveness varies across diverse patient sub-populations. Unlike the case with well-designed experiments, we argue that average treatment effects are difficult, if not impossible, to reliably estimate using observational data. Moreover, in medical practice, the average treatment effect may not be of interest or reflective of the outcome a particular patient might expect. In light of this, we join others in arguing that an analysis using observational data should provide estimates of heterogeneous treatment effects that are defined conditional on pretreatment patient characteristics.

The advantage of carefully designed experiments, such as randomized clinical trials, is that the data have been generated in a way that, in theory, dictates how the data should be analyzed to obtain fair treatment comparisons. As electronic health records (EHRs) are adopted over the coming years, however, observational studies will need to appropriately adjust for undesirable features that are endemic with observational data. These features include, but are not limited to, treatment selection bias, confounding among pretreatment covariates, unmeasured covariates, and unbalanced (and/or incomplete) blocks or strata.

Despite the limitations, doctors and patients are interested in two goals: (i) identifying treatment effects in clinically relevant subgroups and (ii) predicting whether an individual might benefit from a treatment [1]. Both the Agency for Healthcare Research and Quality (AHRQ) and the Patient Centered Outcomes Research Institute (PCORI) have emphasized the need for systematic analysis strategies as part of a pipeline of methods used to identify clinically relevant subgroups [1, 2]. To address this fundamental need in observational comparative effectiveness research (OCER) and patient centered outcomes research (PCOR), *we propose an unsupervised learning approach where patients are clustered on pretreatment covariates and treatment comparisons are made within the identified patient subpopulations.* Such an approach is consistent with the idea that observational studies would benefit from analyses where responses are hidden from the analyst [3].

Although the motivation for such exploratory analyses may be obvious, we illustrate, using an example of Simpson's paradox, the value of preprocessing the data into subgroups using only covariate information. Although the covariate space may be high dimensional, it is likely there are low dimensional representations of "neighborhoods" where patients are relatively homogeneous. Treatment effects in the subgroups are based on well matched individuals. The variability of treatment effects among the different regions of covariate space provides an empirical basis for personalized medicine.

The format of the paper is as follows. In Section 2, we define general notation and fundamental concepts. In Section 3, we explore the simple case with a univariate response and a single covariate that are both binary. In Section 4 we introduce a general functional form for fair overall treatment comparisons that can be characterized by combining local comparisons. In Section 5 we discuss HTEs and unsupervised learning in exploratory subgroup analyses. Finally, we end the paper with a discussion Section 6.

2 Basic Concepts and Notation

Let $f(y, t, \mathbf{x})$ denote the joint probability density function for a patient's outcome variable y, a binary treatment choice indicator t, $t = 1$ or $t = 0$, and a vector of pretreatment covariates $\mathbf{x}$. If treatment assignment is random with the same fixed probability of choosing $t = 1$ for every $\mathbf{x}$, then the distribution of $\mathbf{x}$ within two treatment cohorts is the same. That is,

$$f(\mathbf{x}|t = 1) = f(\mathbf{x}|t = 0). \tag{2.1}$$

That is, the distribution of $\mathbf{x}$ does not depend on t because balance between the choices of $t = 1$ and $t = 0$ is the same for every $\mathbf{x}$.

Fair Comparisons of Type I: Unbiased treatment effect estimates computed from data where treatment choice balance is uniform throughout $\mathbf{x}$-space, as in equation (2.1), yield fair comparisons. Caution: Traditional parametric models for an expected y-outcome as both $\mathbf{x}$ and t-choice vary, typically make strong assumptions that can be wrong. Their estimates are then biased and do not yield fair comparisons even when the data are uniformly balanced.

In observational studies, the $\mathbf{x}$ values observed within the two treatment cohorts are often quite different from what would be obtained from two random samples from the same population. The recently finalized AHRQ guidelines [2] for implementation of OCER describe this situation as potentially limited overlap in covariate space between treatment cohorts. The problem is also referred to as limited common support [4]. In light of the imbalance of covariates, we assume in OCER studies data typically have distributions such that

$$f(\mathbf{x}|t = 1) \neq f(\mathbf{x}|t = 0), \tag{2.2}$$

for at least some $\mathbf{x}$ values. Because of the unknown nature of the potential imbalance in covariates between the two treatment cohorts, we recommend systematic analysis strategy that starts by using unsupervised learning to first cluster patients in covariate space. The motivation for such an approach comes from the following derivation of a

local, potentially heterogeneous effect of treatment, as a function of $\mathbf{x}$:

$$\begin{aligned} HTE(\mathbf{x}) &= \mathrm{E}[(y|t=1) - (y|t=0)|x] \\ &= \int\int (y_1 - y_0) f(y_1, y_0|\mathbf{x}) d(y_1, y_o) \\ &= \int_{y_1} y_1 \int_{y_0} f(y_1, y_0|\mathbf{x}) dy_0 dy_1 - \int_{y_0} y_0 \int_{y_0} f(y_1, y_0|\mathbf{x}) dy_1 dy_0 \\ &= \int_{y_1} y_1 f(y_1|\mathbf{x}) dy_1 - \int_{y_0} y_0 f(y_0|\mathbf{x}) dy_0 \\ &= \mathrm{E}(y|t=1, \mathbf{x}) - \mathrm{E}(y|t=0, \mathbf{x}) \text{ ...conditioned on } t \text{ and } \mathbf{x}. \end{aligned} \tag{2.3}$$

The expressions here illustrate the well-known fact that researchers can ignore the details of the joint distribution of $y|t=1$ and $y|t=0$ and instead use the two marginal distributions, conditional on $\mathbf{x}$, when computing the expected difference of interest. The $HTE(\mathbf{x})$ is an inherently fair comparison because it conditions on the same value of $\mathbf{x}$. That is, an estimand like $\mathrm{E}(y|t=1, \mathbf{x}_a) - \mathrm{E}(y|t=0, \mathbf{x}_b)$ is potentially unfair whenever $x_a \neq x_b$. By investigating the $\mathbf{x}$ space for underlying structure and identifying subregions where individuals are similar, a fair HTE can be estimated based on a cluster of individuals that are well matched in covariate space. This leads to a second type of fair treatment comparison. Namely, each HTE effect estimate computed from patients well-matched on $\mathbf{x}$ is unbiased and potentially heterogeneous.

Fair Comparisons of Type II: A collection of unbiased $HTE(\mathbf{x})$ estimates as $\mathbf{x}$ varies also make fair comparisons, even when treatment choice balance varies with $\mathbf{x}$. Note that each $HTE(\mathbf{x})$ estimate conditions upon $\mathbf{x}$ as well as upon t-choice, leaving no pretreatment patient characteristics to vary. The unbiased $HTE(\mathbf{x})$ estimate that computes the difference between the average y-outcome over all available patients with given $\mathbf{x}$-characteristics who chose $t=1$ minus the corresponding average over patients who chose $t=0$ is then most precise.

A special case of (2.3) occurs when the y-outcomes for different patients are statistically independent. It is frequently reasonable to make this assumption even when datasets are observational. In that case, the joint $f(y_1, y_0|\mathbf{x})$ density is simply the product of two possibly distinct conditional marginal densities: $f_1(y|\mathbf{x})$ times $f_0(y|\mathbf{x})$. Because (2.3) conditions upon a single $\mathbf{x}$ realization, no assumptions are being made about the form of the conditional $\mathbf{x}$ distribution given either treatment choice. The HTE($\mathbf{x}$) values remain well defined and valid even when the local fraction choosing treatment $t=1$ varies considerably across $\mathbf{x}$-space.

In contrast with the HTEs, the average treatment effect (ATE) marginalizes over

the distribution of $\mathbf{x}$.

$$\begin{aligned} ATE &= \mathrm{E}[(y|t=1) - (y|t=0)] \\ &= \int\int \left[\int_{\mathbf{x}} (y_1 - y_0) f(y_1, y_0, x) d\mathbf{x} \right] d(y_1, y_0) \\ &= \mathrm{E}(y|t=1) - \mathrm{E}(y|t=0) \text{ ...conditioned only on } t. \end{aligned} \tag{2.4}$$

As in (2.3), this ATE extimand also separates into a difference between terms that can be computed separately. No problems are caused by this as long as the data come from a well-designed and/or randomized study where treatment choice balance is uniform across $\mathbf{x}$-space. Unbiased ATE estimates then make fair comparisons of Type I.

On the other hand, treatment choice balance cannot be expected to be uniform across $\mathbf{x}$-space in observational study contexts. While the ATE estimand remains well defined in these unbalanced situations, where (2.2) holds, the property that the ATE estimand separates into terms that correspond to expectations over different marginal $\mathbf{x}$-distributions means that a distinctly unfair overall comparison can then result. We illustrate this using an example of Simpson's paradox in the next section.

3 Why HTEs? An Example using Simpson's Paradox

Consider the case of a binary response, two treatments, and a single binary covariate. Using t, y, and x to denote these three binary indicators, the available data can be displayed in a 2×2 table like Table 1. Here y_{ij} denotes the sum of the individual y outcomes for the n_{ij} patients with $t = i$ and $x = j$, for $i, j = 0, 1$. The marginals are defined accordingly. [Page 13]

The conditional treatment differences given x are estimable when all four values of n_{ij} are greater than or equal to one. The corresponding maximum likelihood estimates of the two HTEs can be written as

$$\widehat{HTE}(x=1) = \widehat{\mathrm{E}}[(y|t=1) - (y|t=0)|x=1] = \frac{y_{11}}{n_{11}} - \frac{y_{01}}{n_{01}} \tag{3.1}$$

and

$$\widehat{HTE}(x=0) = \widehat{\mathrm{E}}[(y|t=1) - (y|t=0)|x=0] = \frac{y_{10}}{n_{10}} - \frac{y_{00}}{n_{00}}. \tag{3.2}$$

The corresponding average treatment effect estimator that essentially ignores x is defined as

$$\widehat{ATE} = \widehat{\mathrm{E}}[(y|t=1) - (y|t=0)] = \frac{y_{1.}}{n_{1.}} - \frac{y_{0.}}{n_{0.}}. \tag{3.3}$$

Instances of Simpson's paradox [5] are said to occur when the two conditional esti-

mates (3.1) and (3.2) are observed to have the same numerical sign while the marginal estimate from (3.3) has the opposite sign. For concreteness, we will use the numerical example displayed in Table 2. [Page 13]

Consider patients with coronary artery disease who recently received treatment at one of two revascularization facilities that are to be compared on rates of mortality within 12-months of the initial procedure. Suppose $t = 1$ represents a renowned, cutting-edge facility while $t = 0$ represents a typical, smaller facility within the same geographical region. And suppose further that $x = 1$ denotes a severe level of disease and/or patient frailty while $x = 0$ represents patients with relatively mild conditions still requiring revascularization.

The estimated HTEs for Table 2 are -3.04% for severe patients and -2.02% for mild patients. The estimated ATE that ignores x is 0.60%.

Notice the renowned facility has a better performance in terms of mortality when treating either severe patients or mild patients, but overall the smaller facility has a lower death rate. We contend that the ATE of 0.60% is meaningless in this example. This seemingly contradictory statistic results only because the small facility tends to treat mild patients while the renowned facility treats mostly severe patients. This difference in patient mix is clearly the source of Simpson's paradox in this problem.

Again, ATE estimands make fair comparisons of Type 1 only when the available data come from a well designed and/or randomized experiment. As the example clearly illustrates, ATE estimands and estimators are potentially misleading when computed from unbalanced, observational data.

4 Point-estimates of Overall Treatment Comparisons

Although we believe it is best to report estimated HTEs and their empirical distribution, it is useful to consider ways to make an overall treatment comparison that is inherently fair. Consider the following definition:

$$\begin{aligned} FATE(\Phi) &= \int_{\mathbf{x}} HTE(\mathbf{x}) d\Phi(\mathbf{x}) \\ &= \int_{\mathbf{x}} \mathrm{E}[(y|t=1) - (y|t=0)|\mathbf{x}] d\Phi(\mathbf{x}), \end{aligned} \tag{4.1}$$

where $\mathbf{x}$ denotes a vector of patient covariates and $\Phi(\mathbf{x})$ denotes a specified cumulative distribution function for $\mathbf{x}$. By integrating over this common distribution for $\mathbf{x}$ covariates, $FATE(\Phi)$ represents a weighted estimate of fair, local treatment comparisons.

Fair Comparisons of Type III: Overall treatment effect estimates computed by averaging only fair, local treatment effect estimates across $\mathbf{x}$-space provide unbiased

estimates of their FATE estimand and, thus, also yield fair overall comparisons.

Returning to the binary $\mathbf{x}$ example in the previous section, fair overall estimators are weighted averages of the two HTE estimates:

$$\widehat{FATE}(w) = w\left[\frac{y_{11}}{n_{11}} - \frac{y_{01}}{n_{01}}\right] + (1-w)\left[\frac{y_{10}}{n_{10}} - \frac{y_{00}}{n_{00}}.\right] \tag{4.2}$$

with w denoting the scalar value of the relative weight given to patients with $x = 1$, where w is strictly between 0 and 1. For any such w, (4.2) will be between the two conditional values in (3.1) and (3.2). Thus, if these HTE values have the same sign, then (4.2), unlike the $\widehat{ATE}$ of (3.3), will also have that same sign.

An empirical choice for w in (4.2) can be defined using the within-column sample size totals denoted by $n_{\cdot 1}$ and $n_{\cdot 0}$. Using these values w is defined as $\frac{n_{\cdot 1}}{n_{\cdot 1}+n_{\cdot 0}} = \frac{711}{1296} = 0.549$. $\widehat{FATE}(w)$ is thus -2.58% which, unlike the estimated average treatment effect of 0.60%, strongly favors the renowned facility.

Both the patient matching and sub-classification (or binning) approaches to propensity score (PS) [6] estimation have target estimands of form (4.1). These PS estimates are generalizations of (4.2) to some finite number of subgroups, usually formed by first ranking PS estimates and then matching patients within specified PS "calipers" [7]. Since a logit or probit model based upon a linear function, $\mathbf{x}'\boldsymbol{\beta}$, is typically used to estimate propensity, the patient subgroups being formed correspond to infinite, parallel covariate-space "slabs" oriented so as to be orthogonal to the estimated $\widehat{\boldsymbol{\beta}}$ vector. A highly influential tutorial [8] described a third approach in which PS estimates are simply used as an additional covariate in a multivariable, covariate-adjustment model. Unfortunately, such parametric models yield treatment effect estimates which are unrelated to (4.1) or (4.2) and thus provide no assurance of making fair comparisons.

The local control approach [9, 10, 11] estimates local treatment effects of form (2.3) and displays them in an empirical distribution of effect-sizes. The mean of this distribution, when each local estimate is weighted proportional to the total number of patients in its subgroup, is another example of (4.2). Its FATE estimand is also distinct from the traditional ATE.

Again, HTE($\mathbf{x}$) estimands can vary with $\mathbf{x}$. Thus outcomes researchers should keep in mind that individual patients could be much more interested in likely treatment outcomes for patient subgroups similar to themselves than in overall ATE or FATE estimates.

In practice, statisticians will frequently be expected to produce an overall summary for an OCER study. In order for this summary statistic to be meaningful, we believe it is necessary to obtain local estimates of treatment effectiveness and combine these estimates in a meaningful way. In light of Simpsons paradox, we believe any estimate that marginalizes over the covariate space in different ways within each treatment cohort, rather than in a way common to both, yields an unfair summary.

Authors should be careful and journal editors and readers should be vigilant.

5 Heterogeneous Treatment Effects and Unsupervised Learning

Although the approach is conceptually simple, the analyst faces several choices that shape the outcome of the study. How should subgroups be identified? How many subgroups should be identified? Should only a subset of the available x-variables be use to define patient similarity? This is clearly related to problems with "irrelevant" $\mathbf{x}$-variables. The HTE is a statement about the local mean response; how should one incorporate information about the variability of outcomes under the two treatment choices given this choice of pretreatment covariates, $\mathbf{x}$? Finally, what are the best ways to visualize and synthesize the results from such a systematic examination of subgroup effects such that the results are easily understood by different stakeholders (e.g., doctors, patients, other researchers)? We suggest exploring each of these questions as areas of future research for those interested in OCER methodology [10].

In the first stage of the analysis, observations are first clustered using an unsupervised learning approach. Clustering continues to be an area of active research in the statistics and computer science literature. Advances in OCER will be associated with advances in clustering methods. In particular, how the clustering method interacts with the rest of the analysis strategy needs further study. As more methods are developed and compared, we will be able to better identify the structure in covariate space where neighborhoods of similar patients can be found.

Although in Section 2 we suggested a HTE be defined as the difference the expected outcome, it is easy to imagine cases where a difference in means is not fully informative. For example, suppose two weight loss programs are being considered. Suppose for a given patient covariate vector $\mathbf{x}$, the expected value of weight loss is 10 lbs for treatment 0 and 10 lbs for treatment 1. The HTE estimate would be equal to 0. Such a result may suggest that for this particular patient the treatments do not differ. However, suppose the distributions of weight loss under the two treatment regimes were such that the probability of weight loss greater than 10lbs given $\mathbf{x}$ and treatment 1 was greater than the probability of weight loss greater than 10lbs given $\mathbf{x}$ and treatment 0. By also reporting distributional information beyond the first moment, the patient would be able to make a more informed decision regarding which weight loss program they would like to choose. Namely, the patient would know there is a greater probability of losing more than 10lbs if the patient chose treatment 1 versus treatment 0.

In clinical practice, oftentimes treatment choices are a subjective function of potential risks and benefits. In order to allow for better informed treatment decisions, analyses should consider the distribution of outcomes under different treatment choices conditional on pretreatment covariates, not just first moments. How to best visualize,

report, and validate these distributions is an open area that will involve collaboration with doctors and patients.

6 Discussion

In this paper we review some of the challenges associated with estimating fair treatment effects from observational data. We describe an analysis framework where unsupervised learning is used to form subgroups within which local treatment effects can be estimated. Even in situations where treatment cohorts are balanced, ATEs are becoming less relevant simply because they provide only one-size-fits-all answers about treatment effectiveness. Personalized medicine is asking questions that only HTE estimates can attempt to answer satisfactorily. Such local treatment effects constitute inherently fair, potentially heterogeneous treatment effects, i.e., the effects of treatment for different patients. The fair treatment comparisons do not require 1:1 balance, or any other fixed ratio of treated and untreated patients within the subgroups. By conditioning on pretreatment covariates, the effect estimates are unbiased given available information, i.e., fair. Choice of the number of subgroups can be viewed as a variance-bias trade-off. Local treatment effect estimates will be more precise when blocks are large but less potentially biased when blocks are small.

Our goal is to have an analysis process that produces reliable results. We say that *the comparisons* are fair given the observed covariates. Despite the apparent advantages of our proposed methods, we should keep in mind that making claims from observational studies that replicate is difficult. Multiple data sets should be analyzed with these methods to get an idea of the reliability. The insights drawn from such analyses can be motivation for subsequent descriptive analyses, where data related to pre-specified subgroups are published in the hopes of being used in a subsequent meta-analysis, and confirmatory analyses, where hypothesized results are formally tested for potentially clinically relevant subgroups. That is, continuing research is needed to study strategies for avoiding over-interpretation of estimates from local estimation strategies.

Our example of Section 3 with only one $\mathbf{x}$ variable is too simple to illustrate that traditional covariate adjustment models can yield biased effect estimates. When the covariate $\mathbf{x}$ vector contains several variables, identifying the most appropriate model for traditional covariate adjustment becomes difficult and subjective. These global models are potentially biased locally. By using unsupervised learning to identify patients most similar in $\mathbf{x}$-space and then comparing outcomes within the resulting patient subgroups, the resulting local treatment effect estimates will be less biased and more relevant to new patients facing those same treatment choices. That is, doctors and patients can identify the most relevant subgroup(s) and focus on alternative treatment outcomes within the most relevant patient subpopulation.

To date, published observational studies relying on traditional models have accumulated a rather poor track-record on reproducibility [12]. The lack of reproducibility

is likely due to many factors, including multiple testing and multiple modeling, but bias due to unbalanced covariates is a likely important contributor. We note in passing that it appears that designed studies can also have severe reproducibility problems [13]. As a result of unbalanced covariates, effect estimates are often reflective of an "apples-to-oranges" comparison. Although the analysis strategy suggested here may lead to a large number of estimated HTEs, publishing these estimates and the subgroup covariate information will be more informative to both patients and doctors than average treatment effects. Our hope is that OCER analyses will eventually shift to focus public attention upon heterogeneous effect-size estimates and visualization of their distribution, providing key information about the variability and uncertainty in medical outcomes.

In practice, relevant but unknown explanatory variables will be missing from the data. Because analyses cannot condition upon these missing variables, our proposed comparisons are fair only relative to available covariates in $\mathbf{x}$. As OCER databases include more patient covariates and HTEs are recalculated in the larger covariate space, estimated HTEs will become more accurate. Until then, estimating and disseminating local treatment effects conditional on available covariates are necessary steps in providing patients with truly personalized treatment plans.

Acknowledgments

This material was based upon work partially supported by the National Science Foundation under Grant DMS-1127914 to the Statistical and Applied Mathematical Sciences Institute. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

The authors organized the OCER working group within the Statistical and Applied Mathematical Science's (SAMSI's) 2012-2013 program on Data-Driven Decisions in Health Care (DDDHC), and they particularly wish to thank the researchers who most actively participated in many weekly teleconferences: Kumer Pial Das, Raymond Falk, Myron Katzoff, Meena Khare, Ilya Lipkovich, Marianthi Markatou, John Olaomijo, Paramita Saha Chaudhuri and Jiayang Sun.

Bob Obenchain also benefited from many conversations with Doug Faries about taking treatment differences "locally" or "globally". Bob first used the simple numerical example of Table 2 in his workshop on "Analysis of Nonrandomized Studies" presented to clinical statisticians at Eli Lilly in April 2002.

References

[1] Ravi Varadhan, Elizabeth A Stuart, Thomas A Louis, Jodi B Segal, and Carlos O Weiss. (March 29, 2012). Review of guidance documents for selected methods in patient centered outcomes research: Standards in addressing heterogeneity of treatment effectiveness in observational and experimental patient centered outcomes research [online]. Available: www.pcori.org/assets/Standards-in-Addressing-Heterogeneity-of-Treatment-Effectiveness-in-Observational-and-Experimental-Patient-Centered-Outcomes-Research.pdf.

[2] Priscilla Velentgas, Nancy A Dreyer, Parivash Nourjah, Scott R Smith, and Marion M Torchia eds. Developing a protocol for observational comparative effectiveness research: A user's guide. AHRQ Publication No. 12(13)-ECH099. Rockville, MD: Agency for Healthcare Research and Quality, www.effectivehealthcare.ahrq.gov/Methods-OCER.cfm, 2013.

[3] Donald B Rubin. For objective causal inference, design trumps analysis. *The Annals of Applied Statistics*, 2(3):808–840, 2008.

[4] Elizabeth A Stuart. Matching methods for causal inference: A review and a look forward. *Statistical Science*, 25(1):1–21, 2010.

[5] Colin R Blyth. On simpson's paradox and the sure-thing principle. *Journal of the American Statistical Association*, 67(338):364–366, 1972.

[6] Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

[7] Peter C Austin and Muhammad M Mamdani. A comparison of propensity score methods: A case-study estimating the effectiveness of post-ami statin use. *Statistics in Medicine*, 25(12):2084–2106, 2006.

[8] Ralph B d'Agostino Jr. Tutorial in biostatistics: propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Statistics in Medicine*, 17(19):2265–2281, 1998.

[9] Robert L Obenchain. The local control approach using JMP. In Douglas E Faries, Andrew C Leon, Josep M Haro, and Robert L Obenchain, editors, *Analysis of Observational Health Care Data using SAS*. Cary, NC: SAS Press, 2010.

[10] Robert L Obenchain and S Stanley Young. Advancing statistical thinking in observational health care research. *Journal of Statistical Theory and Practice*, 7(2): 456–469, 2013.

[11] Douglas E Faries, Yi Chen, Ilya Lipkovich, Anthony Zagar, Xianchen Liu, and Robert L Obenchain. Local control for identifying subgroups of interest in observational research: Persistence of treatment for major depressive disorder. *International Journal of Methods in Psychiatric Research*, 22(3): 185–194, 2013.

[12] S Stanley Young and Alan Karr. Deming, data and observational studies. *Significance*, 8(3):116–120, 2011.

[13] C Glenn Begley and Lee M Ellis. Drug development: Raise standards for preclinical cancer research. *Nature*, 483(7391): 531–533, 2012.

Table 1: Notation for the case with binary response, treatment and covariate.

	$x = 1$	$x = 0$	
$t = 1$	y_{11} / n_{11}	y_{10} / n_{10}	$y_{1.} / n_{1.}$
$t = 0$	y_{01} / n_{01}	y_{00} / n_{00}	$y_{0.} / n_{0.}$

Table 2: Hypothetical Example

	Severe	***Mild***	
Renowned Facility	41 / 678	3 / 327	44 / 1005
Smaller Facility	3 / 33	8 / 258	11 / 291