Advancing Statistical Thinking in Health Care Research

R. L. Obenchain, *Risk Benefit Statistics LLC,*
Carmel, IN 46033, U.S.A.
Email: wizbob@att.net

S. S. Young, *National Institute of Statistical Sciences*
Research Triangle Park, NC 27709, U.S.A.
Email: young@niss.org

Abstract

Observational medical studies are becoming more prominent due to interest in using data from current medical practice to help guide treatment selection. We start by reviewing why current analysis strategies for observational data tend to produce findings that fail to replicate. We then argue that there is a real need for simple and objective statistical strategies that yield findings more likely to be dependable. Our proposed way forward is to focus on the empirical distribution of Local Treatment Differences, LTDs, which reveal heterogeneity in effect-sizes. The LTD is the difference in mean outcomes for treated and control patients within a cluster of patients relatively well-matched on their observed pre-treatment characteristics. Because we focus on only one question (comparing two alternative treatment choices for a given disease or condition) and clustering is non-parametric, our proposed approach is more simple and objective than commonly used statistical analysis strategies. By studying graphical displays of information from an LTD distribution, a doctor and patient can not only see a full picture of current treatment outcomes but also make a truly well-informed, individualized treatment choice.

Key Words: Head-to-Head Observational Studies, Nonparametric Preprocessing, Patient Subgroups, Treatment Effect Sizes, Heterogeneous Response, Distribution of Local Treatment Differences, Informative Clusters, Treatment Selection Bias, Unmeasured Confounders

Ultimate Publication: ***Journal of Statistical Theory and Practice*** 2013; 7(2): 456-469.

1. Introduction

Massive data sets are being accumulated from health care systems. There are billing records and there are also patient electronic medical records, EMRs. Within these raw data sets, is there salvageable information about how treatment choice affected resulting outcomes? The real strength of observational data is that they reflect current, actual medical practice; they are also potentially plentiful, timely and relatively inexpensive to accumulate. For diverse patients, it is unlikely that a single "best" (one-size-fits-all treatment) for a named disease exists. Patients differ genetically and are in different current health states, so they are unlikely to all respond the same way to the same treatment. Besides, one named disease may encompass multiple etiologies. For example, no one anti-depressant works for all patients. Bacteria are different, and no one anti-bacterial is effective against all bacterial infections. We should start from the position that heterogeneous patients will respond heterogeneously.

The world has changed for statistics. Because a revolution is truly needed, van der Laan and Rose (2010), we think a revolution in statistical methods to deal with massive data sets is inevitable. Some data sets are expansive in the number of variables, e.g. the "omic" data sets. Other data sets are expanding in the number of records, e.g. the health care data sets. In both cases, the largely hidden problems concern mixtures and unrecognized variation. We need separate diagnostics for subtypes of diseases. With medical record data sets, we need statistical procedures that help recognize all sorts of patient heterogeneity.

This paper is concerned with data sets where there are a large number of records. As the standard error of a mean decreases with the square root of the sample size, every treatment mean will probably be declared different from every other treatment mean. But any overall ranking of treatments, especially on just means, may be quite misleading. Again, patient heterogeneity is likely to be the norm, and easily identified patient subgroups may have diverse expectations.

What is the current state-of-the-art for analysis of medical observational studies? How are we doing? Published literature in outcomes research and pharmacoepidemiology sports a notoriously poor track record with serious lack-of-reproducibility of published findings; see Feinstein (1988), Taubes and Mann (1995), Ioannidis (2005), Kaplan et al. (2010), Young and Karr (2011), to name a few. Even the popular press is taking notice of the problems; Taubes (2007) and Hughes (2007) are two examples. Observational study thought leaders are reacting to these problems; see Feinstein (1988) and Pocock et al. (2004)**.** More ominously, it is possible that there is actual misuse and/or even deliberate abuse of model fitting methods; see Glaeser (2006) and Young and Karr (2011). New, alternative methods that are more difficult to misuse and abuse are thus badly needed, van der Laan and Rose (2010). For example, research sponsored by the Observational Medical Outcomes Partnership (OMOP) showed that seemingly trivial changes in data staging can move answers around profoundly, Ryan (2011). There was a strong hint of model selection problems back in 2002 when Norman Breslow noted in his Biometrics Presidential address that teams of students with the same training and exactly the same data set produced statistical models with vastly different findings, Breslow (2003). It has also been known for some time that different researchers often get different answers for the same question, Feinstein (1988). OMOP research also pointed out that, using the same data set, two groups of researchers found that a treatment both caused, Cardwell et al. (2010), and did not cause, Green et al. (2010), an increase in the same response variable, cancer of the esophagus.

Various sources of bias in observational data are well-known problems. For example, it is well-known that doctors channel certain types of patients to specific drugs. Because sicker patients typically have worse expectations, the drugs they choose may end up being naively indicted as causes of their ultimate poor outcomes and/or their side-effects. Because doctors apparently channeled HIV patients at higher risk for cardiovascular disease to two new drugs in 2001-2003, outcomes researchers ultimately sounded warnings because those patients exhibited a higher incidence of heart attacks, Young and Karr (2011). As another example, high risk patients undergoing heart surgery were routinely channeled to aprotinin and again, no surprise, a misleading high-profile paper resulted, Mangano et al. (2006). In both cases, when analyses were

adjusted for level of risk, differences melted away, Young and Karr(2011) for HIV drugs and Pagano et al. (2008) for aprotinin. Even small biases can lead to "statistically significant differences" because, again, as the sample size increases the standard error of the mean decreases. There can be unmeasured variables that create imbalances and, again, these may lead to statistical significance indicting a treatment when there is no real difference.

Major stakeholders in the current rebate on comparative effectiveness appear to embrace at least seven distinct and potentially conflicting perspectives. These perspectives are those of (i) patients, (ii) health care providers, (iii) health care payers (observational data owners), (iv) government funding agencies, (v) health care regulators / policy makers, (vi) academics and consultants seeking income and/or professional recognition and (vii) the pharmaceutical and device manufacturing industry. While all stakeholders support exchange of scientific (objective) information, each may sponsor only analyses tailored to their unique perspective. When the corresponding (de-identified) analytical files created from observational data are not also released, the magnitude of any induced bias remains unknown. Although not a focus of this paper, public access to data is important so that different interest groups can commission analyses and thereby provide effective oversight.

The nonparametric preprocessing approach we discuss here is ideal for understanding very large, complex datasets. It is fundamentally different from both decision-theoretic methods that rely on expert opinions and predictive models based upon "supervised learning." In fact, our approach is deliberately objective and unsupervised. After all, one does not need to know in advance which treatment was chosen by individual patients or which outcome-of-interest will be evaluated to simply form subgroups of patients who are relatively well-matched on their baseline characteristics! We suggest clustering patients into many subgroups and only then look for local treatment differences, LTDs, within informative clusters, i.e. clusters containing at least one patient who chose each of the two treatments being compared head-to-head.

Improved ways of statistical thinking, van der Laan and Rose (2010), are needed for medical observational data sets used in Comparative Effectiveness Research (CER). Graphical displays should be used prominently in conveying analysis results: Are the claimed effects large enough to literally "see" and to be widely recognized as important? And, of course, the full variability and uncertainty in the data and in predictions should be revealed.

We are confident that health outcomes researchers could greatly benefit from the improved forms of statistical thinking that we advocate here. Our nonparametric preprocessing is simple to understand and appears to produce sound insights that supplement and may, ultimately, supplant traditional analytical approaches.

2. Simulated Observational Data for Numerical Examples

We would have liked to use actual observational data here to illustrate our new statistical thinking concepts. Unfortunately, today's reality is that "interesting" observational datasets from published studies are rarely shared with journal reviewers, regulatory agencies or other outsiders

…let alone made publically available. On the other hand, using simulated data does afford us a "luxury" here …that of knowing when our statistics are either on-track or off!

Obenchain, Hong, Zagar and Faries (2011) used relatively sophisticated simulation techniques to generate synthetic observational data and to compare the root MSE characteristics of alternative methods of analysis. The starting point of their simulation was actual data from 40K patients who had been diagnosed with Major Depressive Disorder (MDD) and had a full year of data both pre- and post-baseline. However, to assure maximum diversity in input datasets, their MSE comparisons used stratified random sampling with replacement from these 40K patients to generate a sequence of 25K pseudo-patient datasets with simulated treatment choices, simulated outcomes and additive noise. The same simulation techniques and initial data were also used to generate a dataset of 250K pseudo-patients for an "Observational Data Analysis Competition" held at the Midwest Biopharmaceutical Statistics Workshop (MBSW) in May 2011, Obenchain (2011) and Young (2011). Although quite large, this dataset proved to be rather easy to analyze. After all, each of the original 40K patient proto-types was then expected to have been re-sampled more than six times …assuring some very good x-space matches for comparing treatment outcomes.

Here, we wish to use a single synthetic observational dataset that is as realistic as possible. Thus we generated a dataset that uses each of the original 40K patient proto-types exactly once. As described in Table 1, the first 11 columns of this dataset contain a sequential patient ID number; a single observed ***y***-outcome cost variable, ***wyrcost***; a binary treatment choice variable, ***trtm*** = 1 or 0, for each patient; and eight patient baseline x-characteristics. No missing values are present in these data.

Table 1: Observed Variables within the "mddsim" dataset. Methods for predicting variable 2 and/or heterogeneous treatment effects are to be based only upon variables 3 through 11.

No.	Var. Name	Description of Variable
1	***patid***	Patient sequential ID number (1 to 40K.)
2	***wyrcost***	Simulated value for the Windsorized (≤ $50K) total health care cost incurred in the current year (rounded to nearest full dollar.)
3	***trtm***	Binary choice between two hypothetical treatments. ***trtm***=1 represents a new pharmacotherapy for Major Depressive Disorder (MDD), while ***trtm***=0 represents the "control" pharmacotherapy for MDD.
4	***age***	Age in years (18 to 64.)
5	***gender***	Binary indicator (1 => female, 0 => male.)

6	***pain***	Ordinal measure of pain = 0, 1 or 2 (lower back and/or neuropathic.)
7	***hoscount***	Number of hospitalizations in year before baseline.
8	***ercount***	Number of Emergency Room visits in year before baseline.
9	***offcount***	Number of office visits in the year before baseline.
10	***psycpcnt***	Percentile on number of PSYC-related visits in the year previous to baseline; this percentile ranges from 19% (0 visits) to 99% (> 57 visits.)
11	***wprevcost***	Windsorized ($50K) total health care cost incurred in the year previous to baseline.

Our interest here will focus on the question: What do these data suggest are the true effect(s) of ***trtm*** choice on ***wyrcost***? In other words, does ***trtm*** choice have no effect, the same effect for all patients, or different true effects for different patients? This final possibility is called *heterogeneous patient response* to treatment, Kaplan et al. (2010), and is discussed here in Section §3; see Figure 1.

As synthetic observational data, the essential feature of our "mddsim" dataset is that the treated cohort (17,973 patients) and control cohort (22,027 patients) represent a pair of populations that tend to be systematically different in terms of patient baseline x-characteristics. In observational (e.g. patient registry and claims database) studies, this is a common situation known as *treatment selection bias*, *confounding by indication* or *patient channeling*.

Furthermore, *unmeasured (hidden) confounders* are also present in these data. In other words, any smooth, global parametric model in the given x-space is a "wrong" model with considerable *lack-of-fit*! Finally, pure *measurement error* is also present. The simulated ***wyrcost*** outcome is a true expected cost, ***twyrcost***, plus an i.i.d. Normal error, with mean = $0 and standard deviation = $200.

Table 2: The "hidden" variables listed here are used in the simulation and are included within the "mddsim" dataset simply for completeness. This information can be helpful if you are interested in simulation details and decide to download the archive of data and R-code posted at http://www.niss.org/xxxx.

12	***cluster***	Number of the 300 Clusters formed in x-space to define variables ***pslocal*** and ***trtmfrac***, below.
13	***pslocal***	Local value of true ***propensity score*** for treatment ***trtm***=1. These values are uniformly distributed within [0.25, 0.75] but are arranged so as to have correlation +0.805 with ***prwyrcost***. Note that variation in ***pslocal*** makes it an *unmeasured (hidden) confounder* that is causing *treatment selection bias* in the simulation.

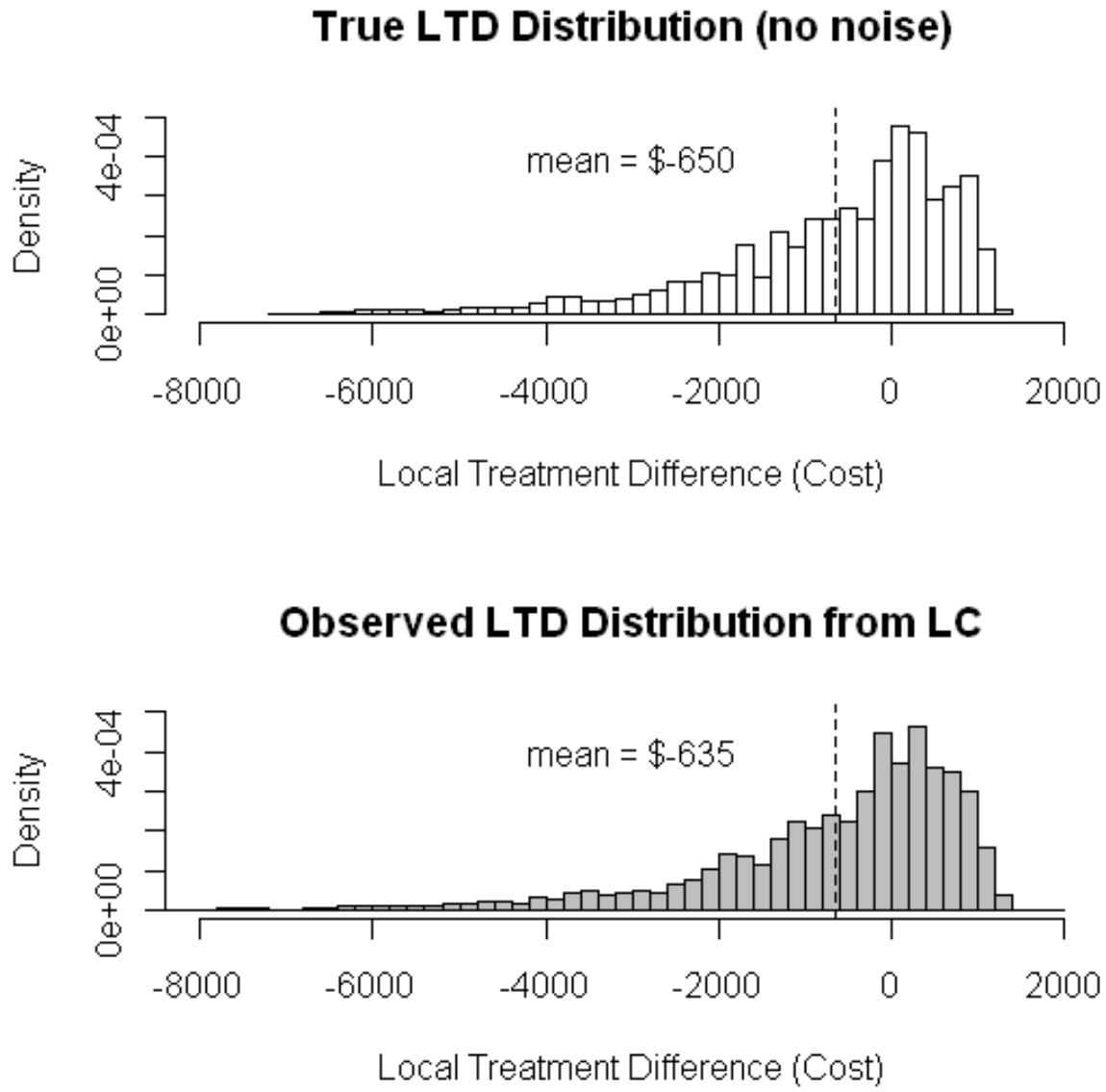
14	***trtmfrac***	Fraction of ***trtm***=0 cost that occurs when one's choice is ***trtm***=1; defined as ***trtmfrac*** = 1.24 □ 0.6 * ***pslocal***. Note that ***trtmfrac*** varies from 0.79 (a 21% reduction in cost) to 1.09 (a 9% increase in cost) for treatment choice ***trtm***=1. Thus ***trtmfrac*** creates *heterogeneous patient response* to treatment.
15	***prwyrcost***	Predictable cost component (parametric model in the eight given x-variables, 4 through 11 of Table 1) for ***trtm***=0 patients.
16	***clrescost***	Within-cluster **residual** cost = **unpredictable** (unmeasured) component orthogonal to the given x-space.
17	***untwycost***	50/50 Mixture of predictable and unpredictable costs for ***trtm***=0 patients.
18	***trtwycost***	= ***untwycost*** * ***trtmfrac*** = 50/50 cost mixture for ***trtm***=1 patients.
19	***twyrcost***	Either ***untwycost*** when ***trtm***=0 or else ***trtwycost*** when ***trtm***=1; true expected value of variable ***wyrcost.***
20	***ltdcost***	Local Treatment Difference (***ltd***) = True ***counter factual difference*** in Windsorized yearly cost (***trtm***=1 minus ***trtm***=0) rounded to nearest dollar.

3. The Essence of the Local Control Approach

The essence of the Local Control (LC) approach, Obenchain (2006, 2009, 2010), for analysis of large, observational datasets is actually rather easy to explain, even to non-technical audiences. LC assumes that health care outcomes for two alternative treatments for the same disease or condition are to be compared head-to-head. LC starts by dividing all patients, without regard for their status as either treated or control, up into many subgroups (or "blocks".) Any way that assures that, within each subgroup, patients are relatively well-matched on (only) their observed baseline x-characteristics can be used; our example will use hierarchical clustering. A Local Treatment Difference, LTD = (Avg. Outcome on Treatment) □ (Avg. Outcome on Control), is then computed within each subgroup that contains both treated and control patients, and the resulting "LTD distribution" is displayed – usually as a histogram, as shown at the bottom of Figure 1. Any cluster that contains only treated or only control patients is considered non-informative.

LC is a form of "Nonparametric Preprocessing" (NPP) for observational data that starts out much like the proposal of Ho et al. (2007) yet goes much further. Specifically, from any observed y-outcome variable, LC creates a new, corresponding LTD variable that estimates the unknown, local, true counter-factual difference in outcome due to treatment choice.

Figure 1. These histograms display distributions of treatment effect-sizes. The top histogram displays the true (hidden) LTD values for 40K individual patients from our simulation of total yearly cost incurred by patients receiving pharmaceutical treatment for MDD. The bottom histogram displays LTD estimates from LC using 2,000 subgroups of patients relatively well-matched in X-space. Clearly, these two distributions are quite similar.



The LC approach thus provides an intriguing, alternative mode of "statistical thinking" for health care research. LC graphical displays allow health outcomes researchers to focus on visualizing effect-size distributions, thereby providing information essential for support of individualized medicine. These displays can also reveal the shortcomings of traditional p-value inferences when effects are small and samples are very large.

Highly rated teachers of technical material have a real knack for simplification in their presentations. The basic concepts and graphical displays employed in LC analyses can be presented in quite simple ways. However, our discussion here will include sufficient extra detail to allow us to address foundational issues in treatment effect-size estimation and to anticipate some possible objections from skeptics.

3.1 Heterogeneous Patient Response

Traditional statistical methods focus, almost exclusively, on the so-called "main-effect" of treatment. Because this measure is an average across potentially diverse patient types, main-effects can have quite limited clinical relevance, especially in individualized medicine. For example, see Kent and Hayward (2007) and Ruberg, Chen and Wang (2010) and the references they cite. In the LC approach, the "main effect" estimate is simply the mean of the observed LTD distribution, but one actually sees a full distribution of effect-sizes.

LC uses information from displays like Figure 1 to address the question of primary interest here: Does pharmaceutical treatment for MDD have no effect on total health care cost, the same effect for all patients, or different true effects for different patients?

Because the data being analyzed here are simulated, the top histogram in Figure 1 provides the definitive answer to our primary question: *heterogeneous patient response* is present! In fact, the observed LTD distribution computed within 2K clusters of relatively well-matched patients is remarkably close to the true distribution of 40K unobserved counter-factual effects. Unfortunately, in actual practice, the top histogram in Figure 1 is always missing (unobservable)!

Thus, let us now focus only on the bottom "observed" LTD histogram in Figure 1. What can one immediately "see" here? First, note the position of LTD = 0 along the horizontal axis. Patients with an LTD near zero incur essentially the same total yearly health care cost regardless of their choice of MDD treatment.

On the other hand, some patients have LTD estimates that are either strongly negative or positive. For example, 30.6% of patients have a negative LTD of negative \$1K or less. Choice ***trtm***=1 would save at least \$1K per year for these 12,131 patients with (high) average ***wyrcost*** = \$16,533, but 4,927 of these patients chose ***trtm***=0. On the other end of the LTD scale, 4.0% of patients have a positive LTD of \$1K or more. Choice ***trtm***=0 would save at least \$1K per year for these 1,601 patients with (low) average ***wyrcost*** = \$13,415, but 426 of these patients chose ***trtm***=1. LTD distributions allow patients and their health care providers to literally "see" a more complete representation of reality in health care treatment outcomes. Based on the patient's demographic values, the doctor and patient can make a well-informed treatment decision.

3.2 Problems with Traditional Visualizations of Effect Sizes

In randomized clinical trials (RCTs), the baseline characteristics of patients in the treated and control cohorts (arms) are expected to be identical, but may vary when the randomization proves

to be unlucky. The statistical analysis plan (SAP) for an RCT typically calls for an initial t-test comparing treatment outcome means and/or a rank test comparing the outcomes between arms, but some sort of statistical model is always pre-specified for "covariate adjustment" just in case.

Figure 2 displays the basic information used by the initial (unadjusted) t-test, which we will argue is quite misleading in our simulated MDD example. After all, our binary treatment choices are random, but propensity (***pslocal*** = treatment choice probability) deliberately varies in our simulation with patient baseline x-characteristics. Specifically, propensity for ***trtm***=1 increases as the conditional expected value of ***wyrcost*** given ***trtm***=0 increases. In other words, our simulation is mimicking the *treatment selection bias* in actual observational studies.

The fundamental short-coming of Figure 2 is that the displayed data have not been adjusted (for *treatment selection bias*), which we know is present here. Specifically, the implied main-effect of treatment appears to be \$14,578 □ \$14,371 = +\$207 in Figure 2. Because sample sizes are so very large here, the conventional t-test tags this main-effect estimate as "highly significant" ($p < 0.0001$.) Unfortunately, we already saw in Figure 1 that the true main effect of treatment is actually negative \$650 and that the LC estimate from 2K patient subgroups is negative \$635. In other words, the conventional (unadjusted) main-effect estimate does not even have the correct numerical sign!

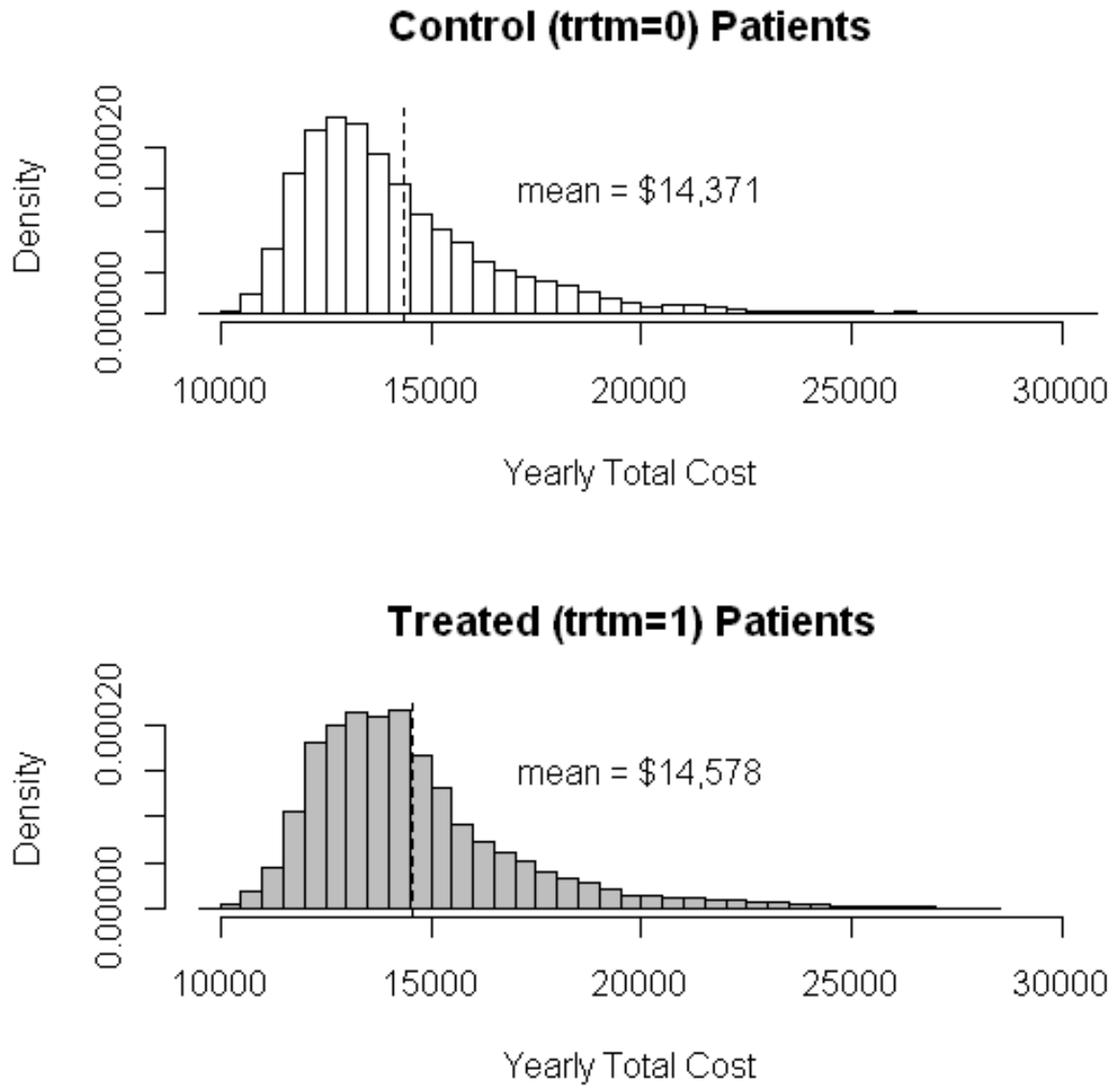
Multivariable regression modeling provides the traditionally accepted approach to "covariate adjustment." Even the very simple regression model that is linear in ***trtm*** and all 8 given x-variables (i.e. no interactions) provides here what is considered to be an "exceptional" fit for observational data; the R-squared statistic for this model is a whopping 61.3%. On the other hand, this simple model does display significant lack-of-fit ($p < 0.0001$) and has a root Mean Square for Error of \$1,702. Since we know that the true noise standard deviation in our simulation is only \$200, this model is clearly mis-specified (i.e. wrong.) However, the treatment main-effect estimate for this model (negative \$579) does at least have the correct sign.

A natural question arises. Is there a display like that of Figure 1 that shows the "corrected" answer from a traditional regression model? While the answer to this question is YES, the reality is that such a display is typically not worth making! If one specifies a "main-effect only" model, such as the above model (with only 9 degrees-of-freedom, plus an intercept), the resulting histogram would then simply represent a Dirac delta (point mass) at the resulting main-effect estimate. Regression models that include interactions between ***trtm*** and patient baseline x-characteristics could yield non-degenerate distributions. But why should any such histograms be of interest when they almost surely represent fits of "wrong" models?

A basic problem with all regression models with more than two explanatory x-variables is that their fit (or lack-of-fit) cannot easily be fully visualized in only three dimensions. In our simulation, treatment and control patients would need to be represented geometrically as some swarm of 40K points, each labeled say with an X symbol for treated or an O symbol for control, that are embedded within an 8-dimensional space. The fit for the most simple, linear model would then be a pair of 7-dimensional, parallel hyper-planes. Would this visualization (or one of

its non-linear generalizations) really be helpful? The answer here would have to be NO, at least for most patients and doctors.

Figure 2. This pair of histograms would not strike most people as being clearly different. The top distribution displays the *wyrcost* variable for 22,027 patients who chose *trtm*=0, while the corresponding distribution for 17,973 patients who chose *trtm*=1 is on the bottom. While the difference in mean *wyrcost* is a mere $207, a conventional t-test nevertheless tags this difference as "highly significant" (p < 0.0001.) Note that any differences in patient pre-treatment characteristics between treatment cohorts are being ignored here.



Thus, we maintain that traditional, global, parametric modeling approaches fail to provide realistic treatment effect-size visualization tools that represent worthy competitors to LTD

histogram displays, such as the one at the bottom of Figure 1. Again, the LC approach provides, via nonparametric preprocessing, essential information about potential treatment effect heterogeneity that is needed to form a scientific basis for individualized medicine.

3.3 Validation of LC Adjustment for Treatment Selection Bias

When one's sample size is quite large (40K here), even effects that are quite small numerically can be "significant" statistically. After all, we saw an example of this in Figure 2. When treatment heterogeneity is present, tests for only main effects can be meaningless.

Since the LC approach emphasizes visualization of treatment effect-size distributions from potentially massive datasets, it seems quite natural to also rely upon visualizations (rather than asymptotically meaninglessly small p-values) to confirm that LC adjustment for treatment selection bias has been effective. The key concept needed for this visualization comes from answering the question: What would one expect to see in an LTD distribution if all of the observed patient baseline x-characteristics are actually totally unrelated to expected y-outcomes?

Logically, if all patient subgroups are formed using only "irrelevant" observed x-variables, then the supposedly "local" comparisons being made are actually just random comparisons. Furthermore, another way to form random subgroups would be to *ignore all observed x-variables* and simply form purely "artificial" subgroups.

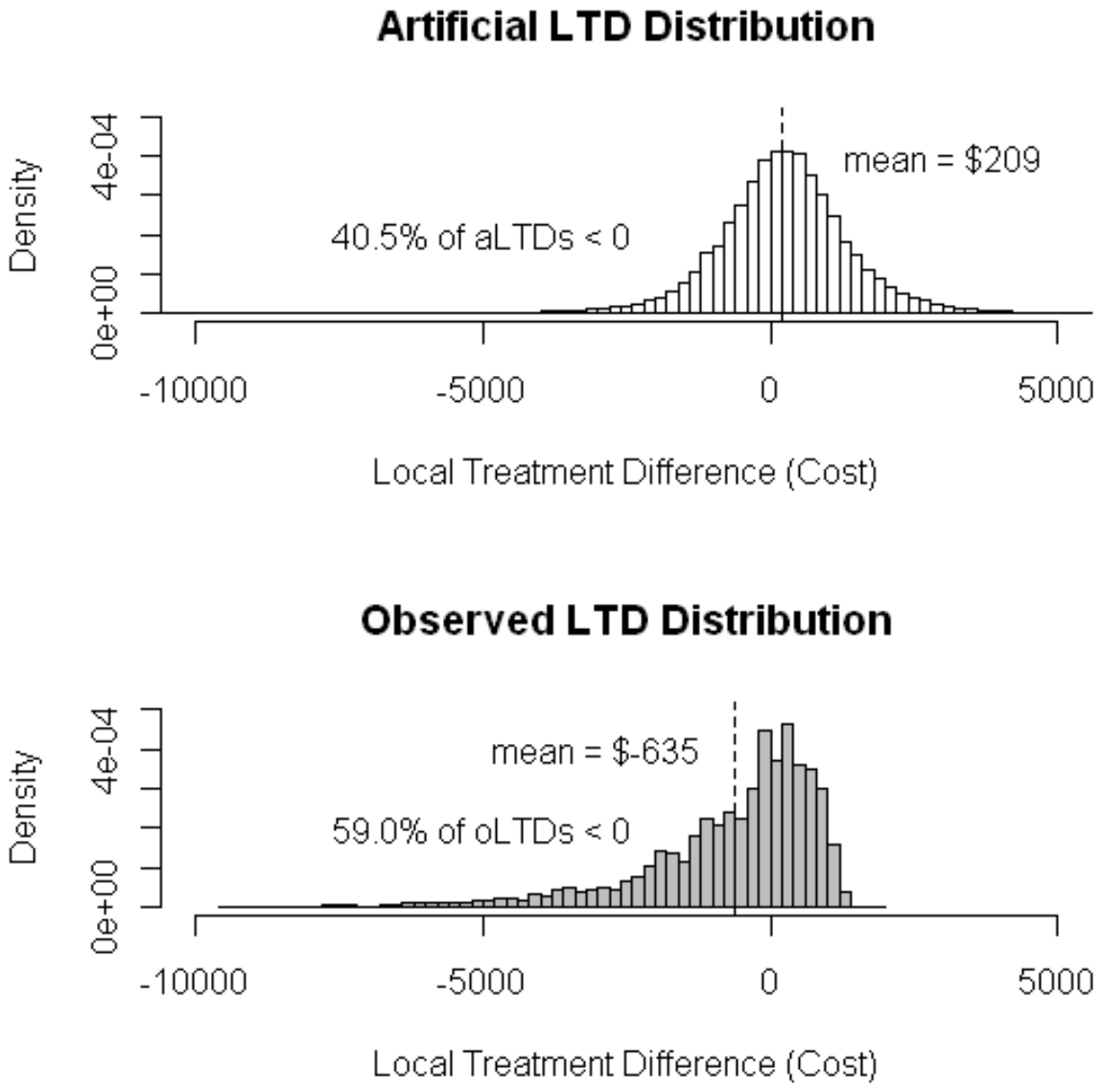
To eliminate any effects of the choice of the number of subgroups being formed, the sizes of these subgroups, or even the fractions of treated patients within these subgroups, the "artificial" subgroup formation process can exactly mimic all of these characteristics from an "observed" LTD distribution that does use patient baseline x-characteristics to form its subgroups.

For example, corresponding to the "observed" LTD distribution at the bottom of Figure 1, an "artificial" LTD distribution could be formed by randomly assigning the 40K patients to 2K subgroups of the same sizes and with the same within-subgroup treatment fractions. In fact, to greatly increase the precision of the resulting "artificial" LTD distribution, independent replications of these random assignments could be made. In each such replication, 99 subgroups (containing a total of 415 patients) would be uninformative (contain only ***trtm***=1 or only ***trtm***=0 patients) because those were the characteristics of the "observed" subgroups.

Figure 3 displays a pair of LTD distributions like Figure 1. But the top histogram in Figure 3 depicts an "artificial" LTD distribution corresponding to 10 independent replications. Each replication formed 2K random subgroups that mimic the subgroups that define the "observed" LTD distribution shown at the bottom of Figures 1 and 3.

LC adjustment has been effective (and *treatment selection bias* has been detected) if and only if an observed LTD distribution is clearly different from its corresponding artificial LTD distribution. This is quite clearly the case in Figure 3. The artificial LTD distribution is centered near the positive, unadjusted main-effect of treatment and looks symmetrical like stereotypical noise. The observed LTD distribution has a negative, adjusted mean and is highly skewed. The observed x-covariates are showing their effects on LTDs.

Figure 3. The observation that the two distributions depicted here are clearly different provides strong evidence that the Local Control approach has made an important adjustment and has revealed treatment selection bias.



Dividing patients randomly into subgroups really should not have any systematic, long-term effects on treatment comparisons. Thus, as expected, the mean of the artificial LTD distribution at the top of Figure 3, which is $209, is quite close to the unadjusted main-effect estimate from Figure 2, which is $207. Furthermore, the large spread in the artificial LTD distribution represents increased noise caused by randomly making some *rather inappropriate* direct comparisons. For example, most *randomly formed* patient subgroups could contain not only an

elderly female taking ***trtm***=1 who used many health care services the previous year but also a young male taking ***trtm***=0 who used almost no health care services the previous year, etc, etc.

In sharp contrast, an observed LTD distribution results from making only the direct (within subgroup) comparisons that *appear to be most appropriate* …in terms of given patient baseline x-characteristics. If information on unmeasured (hidden) confounders could also be used in forming subgroups, the resulting LTD distribution would probably be even more different from its corresponding artificial LTD distribution. Unfortunately, that's usually just not possible. Thus LC concentrates on what is possible and asks: Are the given x-variables related to any observed variation in treatment outcomes? Displays like Figure 3 answer this question visually. (Use of the given x-variables to formally predict LTDs is discusses in the next section, §3.4.)

Note that observed LTD distributions can be interpreted much like Bayesian posterior distributions, but they are entirely objective. Traditional "subjective" Bayesian inferences use *informative priors*, and thus represent a compromise between objective, sample information derived from the current data and prior information representing, say, expert opinion or results from older and/or possibly less relevant studies. Objectivity is widely considered an essential feature of scientific credibility.

3.4 Predicting Treatment Effects for Individual Patients

The LC approach is clearly a "divide and conquer" analysis strategy, executing the analysis strategy in simple steps. Besides being firmly based on "blocking" concepts (forming many subgroups of relatively well-matched patients and relying upon the resulting local estimates), LC also divides analyses of large, complicated datasets into distinct phases. Traditional parametric modeling approaches may appear to do "everything" at the same time …but the (frequently hidden) reality is that many preliminary analyses were probably considered and incrementally "refined." Of course the reader, unless informed, will never know how much model selection has gone on.

As illustrated in Figures 1 and 3, LC can quite successfully use Nonparametric Preprocessing techniques to form LTD estimates and display effect-size distributions. Unfortunately, the key remaining question is: "Can the given x-variables be used to adequately predict the observed LTDs?" This final phase of LC is clearly the most uncertain, subjective and potentially frustrating analysis phase (due to unmeasured confounders.) With observed LTDs as the new "left-hand-side variables" created via LC, "supervised learning" methods are now needed for prediction. With the nonparametric "gloves" potentially off, what would you do now?

We will continue to favor approaches that emphasize visualizations that make predictions which are easy to understand. Recursive Partitioning (RP) answers the need for a relatively non-parametric prediction method. For example, Figure 4 presents a RP "tree" model for our simulated MDD example. Here, we used the conditional approach of Hothorn, Hornik and Zeileis (2006), and Figure 4 is a default plot from their "party" R-package with Strobl (2010). Since the only baseline x-characteristic variables used in this tree are ***wprevcost*** (Windsorized

total health care cost in the 12 months prior to baseline) and ***offcount*** (number of office visits in that prior period), this model is actually quite easy to explain. Note that the "predictions" from this model are graphical I-plots (box and whisker diagrams) of observed LTDs within its six terminal Nodes or final "leaves."

The tree of Figure 4 makes it quite clear that patients with the highest ***wprevcost*** values (> \$17.6K and especially > \$31.4K in the two right-most Nodes, #10 and #11) should take ***trtm***=1 for MDD. The implied savings in ***wyrcost*** (negative LTDs) tend to be considerable for these particular patients.

In fact, the left-most Node (#4) in Figure 4 (where ***wprevcost*** < \$6.4K and ***offcount*** is at most 9) is the only final leaf in which choice of cost-effective ***trtm*** for MDD remains in any doubt. This Node contains the lowest percentage (40.2%) of negative LTDs, but its inter-quartile range, [negative \$258, positive \$601], indicated by the wide box in the I-plot of Node #4, straddles zero dollars. Further splitting of this rather large Node (21K patients) is possible but, unfortunately, not particularly helpful. In other words, choice of ***trtm*** (new vs control) for MDD is fairly unimportant for slightly more than half of the patients in our 40K pseudo-observational sample.

On the other hand, within Node #5, 70.0% of the 7,443 patients have negative LTD estimates. And this percentage increases until it reaches 99.6% in the right-most terminal Node, #11. Thus ***trtm***=1 is rather clearly favored in all final leaves to the right of Node #4 in Figure 4, containing 47.5% of the patients in our sample.

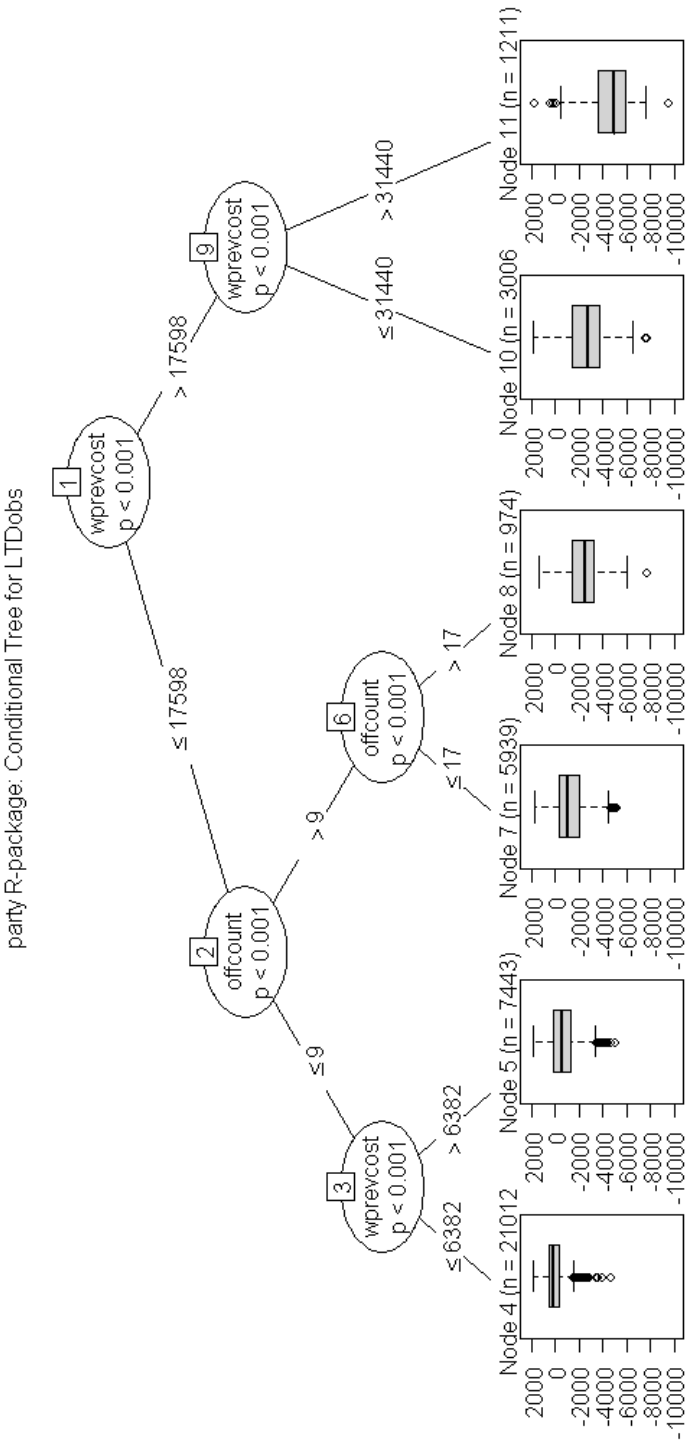
4. Summary

The statistical model underlying LC is a particularly simple one, that of nested ANOVA (treatment within subgroup.) Nothing about "how" subgroups were formed using patient baseline x-characteristics is actually used in computing a LTD distribution. Here we have simply used (hierarchical) clustering techniques, as in Obenchain (2010, 2011), to form subgroups.

The foundational issue in LC is usually: How much of the variation in an observed LTD distribution is due to true variation in treatment effect size (i.e. resulting from variation in patient baseline x-characteristics) and how much is due simply to noise in y-outcome measurements? This question usually cannot be resolved definitively. However, as we saw in §3.4, efforts at predicting LTDs from patient baseline x-characteristics can definitely shed some light on this and related questions. When the best model one can find for predicting LTDs has considerable lack-of-fit, much of the remaining variation could easily be due to unmeasured confounders rather than to pure noise.

Statistical inferences from global, parametric models have potential to be misleading because they are based upon strong assumptions that can be dead wrong. Because LC methods make fewer and much weaker assumptions, LC produces LTD estimates that depend, essentially, only on the available data.

Figure 4. A "tree model" for prediction of observed Local Treatment Differences.



The conceptual "simplicity" of the LC approach does, however, place great emphasis on the computational power needed to cluster large numbers patients (unsupervised learning), perform

sensitivity analyses and generate detailed graphical visualizations. As illustrated in the examples cited by Efron (1979), this can be a very welcome trade-off. LC is a prime example of the new statistical thinking demanded by van der Laan and Rose (2011) and Stuart (2010).

The essence of the LC approach is to provide basic computational and graphical tools that allow observational health care data analyses to be more objective and therefore help in development of consensus views. Specifically, LC focuses on LTD estimates, derived via nonparametric preprocessing, that reveal effect-size distributions. These distributions provide estimates of counterfactual differences and thus can quantify heterogeneous patient response to treatment, thereby providing a scientific basis for individualized medicine.

References:

Breslow, N.E., 2003. Are statistical contributions to medicine undervalued? *Biometrics* 59, 1□8.

Cardwell C.R., Abnet C.C., Cantwell M.M., Murry L.J. 2010. Exposure to oral bisphosphonates and risk of exophageal cancer. *Journal of the American Medical Association* 304, 657-663.

Efron, B., 1979. Computers and the Theory of Statistics: Thinking the Unthinkable, *SIAM Review* 21, 460-480.

Feinstein A.R. 1988. Scientific standards in epidemiologic studies of the menace of daily life. *Science* 242**,** 1257-1263.

Glaeser E.L. 2006. Researcher incentives and empirical methods. www.economics.harvard.edu/pub/ hier/2006/HIER2122.pdf

Green J., Czanner G., Reeves G., Watson J., Wise L., Beral V. 2010. Oral bisphosphonates and risk of cancer of oesophagus, stomach, colerectum: case-control analysis with a UK primary care cohort. *British Medical Journal* 341:c4444.

Ho, D.E., Imai, K., King, G., Stuart, E.A., 2007. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference, *Political Analysis* 15, 199-236.

Hong, Q., Obenchain, R.L., Zagar A. and Faries, D.E., 2011. An archive of R-code and datasets for comparison of observational data analyses via calls to SAS/STAT Procedures and Macros, Unpublished Technical Materials. http://members.iquest.net/~softrx

Hothorn T., Hornik K., Zeileis A. 2006. Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics* 15(3), 651-674.

Hothorn T., Hornik K., Strobl C., Zeileis A., 2010. Party: A laboratory for recursive partitioning. http://cran.r-project.org/web/packages/party

Hughes S., 2007. *New York Times Magazine* focuses on pitfalls of epidemiological trials. September 18, 2007. http://www.theheart.org/article/813719.do

Ioannidis, J. P. A., 2005. Contradicted and initially stronger effects in highly cited clinical research, *Journal of the American Medical Association* 294, 218–229.

Kaplan, S.H., Billimek, J., Sorkin, D.H., Ngo-Metzger, Q., Greenfield, S., 2010. Who Can Respond to Treatment? Identifying Patient Characteristics Related to Heterogeneity of Treatment Effects, *Medical Care* 48, S9–S16.

Mangano D.T., Tudor I.C., Dietzel C. 2006. The risk associated with aprotinin in cardiac surgery. *New England Journal of Medicine* 354, 353-365.

Obenchain, R.L., 2009. SAS Macros for local control (Phases One and Two), Observational Medical Outcomes Partnership (OMOP), Foundation for the National Institutes of Health (Apache 2.0 License), http://members.iquest.net/~softrx

Obenchain, R.L. 2010. The local control approach using JMP, *Analysis of Observational Health Care Data Using SAS*, Faries, D.E., Leon, A.C., Haro, J.M., Obenchain, R.L., eds, SAS Press, Cary, NC, 151–192.

Obenchain, R.L., 2011. Observational Data Analysis Competition: Heterogeneous Response Challenge. http://www.mbswonline.com/presentationyear.php?year=2011

Obenchain, R.L., Hong, Q., Zagar, A., Faries, D.E., 2011. Observational Data Simulation Scenarios for Windzorized Yearly Costs of Patients with Major Depressive Disorder, Unpublished Technical Materials. http://members.iquest.net/~softrx

Obenchain, R.L., Hong, Q., Zagar, A., Faries, D.E., 2012. Observational data analysis: MSE loss comparisons of local control versus parametric models. Submitted.

Pagano D., Howell N.J., Freemantle N., et al. 2008. Bleeding in cardiac surgery: The use of aprotinin does not affect survival. *J Thorac Cardiovasc Surg* 135, 495-502.

Pocock S.J., Collier T.J., Dandreo K.J., *et al*. 2004. Issues in the reporting of epidemiological studies: a survey of recent practice. *British Medical Journal* 329, 883-888.

Ryan P. 2011. Impact of observational analysis design: Lessons from the Observational Medical Outcomes Partnership. http://www.niss.org/sites/default/files/OMOP_Ryan_NISS_16Jun2011.pdf

Ruberg, S.J., Chen L., Wang Y. 2010. The mean does not mean as much anymore: finding sub-groups for tailored therapeutics, *Clinical Trials* 7, 574–583.

Stuart, E.A. 2010. Matching methods for causal inference: A review and a look forward, *Statistical Science* 25, 1–21.

Taubes, G., Mann, C.C. 1995. Epidemiology faces its limits. *Science* 269, 164-169.

Taubes G. 2007. Do we really know what makes us healthy? *New York Times Magazine*, September 16. http://www.nytimes.com/2007/09/16/magazine/16epidemiology-t.html?pagewanted=print

van der Laan, M., Rose, S. 2010. Statistics ready for a revolution: Next generation of statisticians must build tools for massive data sets. *AMStat News*, September, 38-39.

Young, S.S., 2011. Identification and Use of Patient Heterogeneity, http://www.mbswonline.com/ presentationyear.php?year=2011

Young, S.S., Karr, A. 2011. Deming, data and observational studies: A process out of control and needing fixing. *Significance* September, 122-126.