Observational Data Analysis: MSE Loss Comparisons of Local Control with Parametric Modeling Approaches

Robert L. Obenchain[1], Quan Hong[2], Anthony Zagar[3] and Douglas E. Faries[3]

[1]Risk Benefit Statistics LLC, 13212 Griffin Run, Carmel, IN 46033, USA
[2]Bristol-Myers Squibb, 3553 Lawrenceville Rd, Princeton, NJ 08540, USA
[3]Eli Lilly and Company, Lilly Corporate Center, Indianapolis, IN 46285, USA

Abstract: Some readers may find our results somewhat surprising: nonparametric methods for analysis of large, observational datasets have genuine potential to be more accurate than traditional regression-like models, with or without propensity score adjustment. Whenever (1) the true heterogeneity in treatment effect-sizes exceeds the purely random noise in patient-level outcome measures and (2) the hypothesized models are wrong (too simple and smooth), nonparametric methods can be more accurate because the assumptions made are fewer, weaker and yet quite realistic, locally. Since our findings are based upon simulation studies, we start by outlining how sources of bias -- such as variation in treatment selection fractions, patient heterogeneous response and unmeasured confounders -- can be mimicked when generating pseudo-observational data. We then show that Local Control (LC) counterfactual difference estimates, from simple, post-hoc blocking / matching of patients on pre-treatment covariates, can achieve lower root MSE loss than five traditional approaches based upon global, parametric models. [153 words; 1,106 characters including spaces.]

Keywords: observational data, patient subgroups, local control, multivariable modeling, propensity for treatment, bias, confounding, comparative effectiveness research.

1. INTRODUCTION

In what Thompson [1] depicts as the current “Age of Risk Management,” warnings from vaguely specified new studies about serious side effects of medications have become steady features on the nightly news. While health care professionals have criticized the analytical tactics and decision rules commonly employed in such studies [2-9], the popular press has even suggested that statistical inference procedures, if not the scientific method itself, may be flawed [10-12]. In reality, the very same statistical methods that are universally accepted for use in clinical trials - where highly relevant data are carefully collected and reviewed, treatment assignment is randomized, and both patient evaluations and data analyses are initially blinded – can easily be misused or deliberately abused in typical observational study contexts. Especially when health care datasets are massive but prone to errors, missing covariates and coding biases, there is a real need for new analytical approaches that are, intrinsically, more realistic and objective [13,14]. The LC approach does this by making fewer and weaker assumptions than traditional parametric models as well as by reducing reliance on highly-subjective opinions and publication biases involving p-values (observed significance levels.) After all, p-values for global effects always tend to become quite small in truly large samples …i.e. they become meaningless!

To better inform individualized medicine, new and different analytical approaches must specifically address patient differential response to treatment [8,9,15,16]. When the patient population represented in a database encompasses multiple, diverse classes of patients, the (overall) main-effect of treatment is clearly no longer a sufficient statistic worthy of primary focus. Comparative Effectiveness Research (CER) needs to focus on providing information about how patients with different pre-treatment ***x***-characteristics are likely to experience different ***y***-outcomes under alternative treatment choices.

Post-hoc formation and analysis of specific patient subgroups is widely, if not universally, discouraged in clinical trial settings, and there is little consensus on the ultimate validity of such results [17]. In sharp contrast, conditional statistical inference using patient subgroups has been a primary focus of observational study research ever since the seminal work of Cochran [18,19].

Treatment comparisons made within small, relatively well-matched subgroups of patients are intrinsically fair (self-adjusted for treatment selection bias and confounding), and observing the variability of these "local" findings across many subgroups generates a full ***distribution of treatment effect-size estimates*** based only on sample information. In other words, the fundamental design-of-experiments concept of "blocking" becomes an efficient, nonparametric observational data analysis tool (nested ANOVA, treatment with subgroups). Starting out by forming many patient subgroups becomes a valid "divide and conquer" analysis strategy whenever the ultimate objective is to put all of the resulting disparate, local pieces of information back together again!

Using the pre-treatment ***x***-characteristics of individual patients, there are many, diverse ways to form subgroups – visualized here as mutually exclusive and exhaustive subsets of all patients with data available for analysis. When formed informally or via unsupervised learning techniques [20-24] that ignore treatment choice, such subgroups are typically called "strata" or "clusters." Otherwise, subgroups can be formed using information on observed treatment choices to emphasize local imbalances in treatment fractions (channeling) and possibly also to avoid "uninformative" subgroups (containing only treated or only control patients). The resulting subgroups are then called (i) "subclasses" or propensity score "bins" when formed by rank ordering estimated treatment selection probabilities from a discrete choice model [25-27], (ii) "leaf nodes" when formed via recursive partitioning (classification tree models) [28], or even simply (iii) "matched sets" [29-32].

Strategies for analysis of an existing observational dataset that insist on matching treated and control patients in some fixed ratio (such as 1:1) tend to be inefficient (i.e. to ignore the observed outcomes from many relevant patients.) This strategy can even represent "trust me" science …by providing a cover-up for deliberate bias introduced via faulty or serendipitous matching. Needless to say perhaps, but methods of forming subgroups that also use information from patient ***y***-outcome variables should also almost always be avoided.

2. REALISTIC SIMULATION OF OBSERVATIONAL DATA

To realistically compare observational data analysis methodologies, the datasets simulated here need to possess all basic features typically found in observational data. The starting point for our simulation was a master analytical file of actual observational data collected on 40K patients who had sought treatment for Major Depressive Disorder (MDD). Thus, a relatively rich variety of datasets of size 25K each can be formed by resampling patients, with replacement, from this master dataset, while generating new, synthetic treatment choices (***t*** = 1 or ***t*** = 0) and corresponding ***y***-outcomes for each selected patient. A full discussion of our observational data simulations, including datasets and R-code, is provided in companion technical documentation

[33,34] available from the authors. Only the main features that define our simulation strategy / tactics are summarized here in Subsections §2.1 - §2.5.

2.1. Confounding among Explanatory Variables

The mean values, variation and correlations between all eight of the baseline patient characteristics in our simulated data are realistic because they represent actual, resampled ***X***-vectors. For example, roughly 70% of patients seeking treatment for MDD are female. Of the 7 other baseline characteristics of patients that we used, two key ***X***-predictors of total health care cost for the current year are (1) total health care cost for the full year prior to baseline, Windsorized to be ≥ $100 and ≤ $50K, and (2) total number of psychological-treatment visits received during that prior year (measuring intensity of prior, non-pharmaceutical treatment for MDD).

2.2. Mixture of Two Signal Types plus Noise

The ***y***-response to hypothetical treatment for MDD that we generated is the simulated total health care cost for the full year following baseline, again Windsorized to be ≥ $100 and ≤ $50K. These simulated numerical values consist of a ***signal*** component plus an additive ***noise*** component. The noise components represent measurement error and consist of independent and identically distributed pseudo-random normal deviates with mean $0 and standard deviation either $1K or $5K.

The signal component represents the true conditional expected value of the ***y***-outcome given that patient's ***t***-treatment choice, baseline ***X***-characteristics and hidden ***Z***-confounders. As explained in the next three subsections, this signal (i) is a mixture of a relatively smooth component predictable from the given ***X***-covariates with an un-smooth and un-predictable ***Z***-component, (ii) depends upon ***t***-treatment choice via a multiplicative factor that varies with ***X***, and (iii) includes two different levels for predictability of the local propensity for treatment (simulation scenarios A or B.) This combination of diverse simulation mechanisms assures that no traditional, parsimonious modeling approach can provide a strictly correct model. This is quite typical of observational studies where model lack-of-fit includes the true effects of unmeasured confounders (Subsection §2.4) as well as purely random noise.

2.3. Patient Heterogeneity in Response to Treatment

The predictable signal component for control (***t*** = 0) patients comes from a multivariable regression model that is factorial-to-degree-two in all eight patient baseline ***X***-characteristics. This global, parametric model – which was fitted to actual post-baseline ***y***-outcomes from 40K patients receiving any of a variety of current treatments for MDD – has 36 degrees-of-freedom (8 main effects, 28 two-way interactions and no squared terms). These predictable signal components range from $258 to $42,309 and have a highly skewed distribution with mean $9,861.

The full signal component (predictable plus unpredictable) for treated (***t*** = 1) patients is defined to be a specified, scalar multiple – ranging from 0.79 (a 21% decrease in true cost) to 1.09 (a 9% increase) – of the corresponding full control signal for a patient with the same baseline ***X***-characteristics and hidden ***Z***-confounders. As outlined in Subsection §2.5, this multiplicative factor, ***trtmfrac***, depends only upon each individual patient's true propensity for treatment [24,25].

Together, these two sources of outcome heterogeneity assure not only that expected response to treatment received but also the true treatment effect difference (***t*** = 1 minus ***t*** = 0) varies from patient to patient.

2.4. Hidden / Unmeasured Confounders

Another feature of our simulation is that the eight baseline ***X***-characteristics were also used to hierarchically cluster the initial 40K patients into three hundred clusters ranging in size from 16 to 548 patients. The 40K actual post-baseline ***y***-outcomes used to fit the regression model described in Subsection §2.3 were then replaced by their within-cluster average values and a factorial-to-degree-two model was again fit. This second time, only the unpredictable fitted residuals were retained, multiplied by 2 and shifted to have mean $20K. These signal ***Z***-components range from $2,082 to $46,808 and have a fairly symmetric distribution.

The full signal component for ***t*** = 0 patients is then a mixture of the form [pmix × ***X***-predictable cost component] + [(1 – pmix) × unpredictable cost ***Z***-component], where pmix = 0.1, 0.5 or 0.9.

Three additional hidden simulation ***Z***-effects are described next.

2.5. Treatment Selection Bias

Treatment selection bias (channeling) exists whenever the true propensity for treatment choice ***t*** = 1 somehow varies within a dataset. The 300 hidden clusters described above were again used to define true ***Z***-propensities that vary from 0.25 to 0.75 but are constant for all patients within the same cluster. Furthermore, the cost multipliers described in subsection §2.3 are actually defined by a hidden linear ***Z***-relationship: the scalar multiplier is ***trtmfrac*** = 1.24 – 0.6 × true local-propensity.

In our six type "A" simulation scenarios, the true ***Z***-propensities are 300 pseudo-random samples from the distribution uniform on 0.25 to 0.75. As a direct result, predicting either propensity or true variation in treatment effects from the given patient ***X***-characteristics is difficult in all six A scenarios.

In our six type "B" simulation scenarios, the 300 true ***Z***-propensity values (and corresponding multipliers) are re-assigned to the 300 hidden clusters in such a way that the correlation between the true ***Z***-propensity and the predictable signal component for ***t*** = 0 patients becomes +0.720; this correlation is only +0.081 in all type A scenarios. Predicting propensity and/or true variation in treatment effects thus becomes much easier in type B scenarios, at least when pmix is 0.5 or 0.9. In fact, each type B scenario represents a situation where treatment ***t*** = 1 saves money (0.79 ≤ multiplier < 1) for patients with higher expected yearly cost and who are relatively likely to choose ***t*** = 1 (propensity 0.40 to 0.75) but increases cost (1 < multiplier ≤ 1.09 due to, say, failure to recover a higher acquisition charge for ***t*** = 1) for patients with lower expected yearly cost (say, less severe depression) and who are relatively unlikely to choose ***t*** = 1 (propensity 0.25 to 0.40).

Finally, each simulated dataset of size approximately 25K patients is formed by stratified re-sampling (with replacement) from the original 40K patients. The final **Z**-effect is that these strata are the 300 hidden clusters, and the 300 sub-sample sizes are (except for truncation to integers) proportional to the hidden cluster sizes. Due to this truncation, the number of patients in each of our simulation replications turned out to be 24,987.

3. THE SIX ALTERNATIVE APPROACHES COMPARED HERE

The patient-level quantities of primary interest here are their true Counterfactual Differences in Expected Outcome:

$$\Delta(\boldsymbol{X}, \boldsymbol{Z}) \;=\; \mathrm{E}(\, \boldsymbol{y} \mid \boldsymbol{t} = 1, \boldsymbol{X}, \boldsymbol{Z}\,) - \mathrm{E}(\, \boldsymbol{y} \mid \boldsymbol{t} = 0, \boldsymbol{X}, \boldsymbol{Z}\,).$$

These quantities are called counterfactuals because only one ***y***-outcome (signal plus noise) is assumed to be actually observed for each patient …either that for choice $\boldsymbol{t}$ =1 or else that for the control choice, $\boldsymbol{t} = 0$. As a result, good estimators of these many, individual Δ parameters typically do not exist. The LC approach attempts to estimate these Δ parameters by looking for *approximate matches* on only the given ***X***-covariates. As a result, some of the resulting LC estimates will usually be missing values and, in fact, none may be truly unbiased.

When a method of estimation yields individual Δ estimates that (except for any missing values) are all equal numerically, that method will be said to be homogeneous. Otherwise, the method of estimation will be called heterogeneous.

The Main Effect of Treatment for each simulated dataset of roughly 25K patients is the expected value of $\Delta(\boldsymbol{X}, \boldsymbol{Z})$ when the true, joint distribution of $(\boldsymbol{X}, \boldsymbol{Z})$ corresponds to its sample distribution. Since true ***Z***-propensity for treatment is within [0.25,0.75] and thus not near either 0 or 1 in our simulation, all of the individual $\Delta(\boldsymbol{X}, \boldsymbol{Z})$ effects are well defined. The alternative estimates of Main Effects to be compared here are simply the corresponding sample means of all non-missing individual Δ estimates.

The six alternative methods for estimating the true Δs compared here are described within subsections §3.1 to §3.4.

3.1. Direct Comparison of Treatment Cohort Means (Unadjusted)

The overall treatment difference [Average of all observed ***y***-outcomes for treated ($\boldsymbol{t} = 1$) patients] minus [Average of all observed ***y***-outcomes for control ($\boldsymbol{t} = 0$) patients] is a natural estimator of $\mathrm{E}(\, \boldsymbol{y} \mid \boldsymbol{t} = 1) - \mathrm{E}(\, \boldsymbol{y} \mid \boldsymbol{t} = 0)$ that is unadjusted for ***X*** (as well as for ***Z***). This approach yields a single numerical value viewed here as providing homogeneous estimates for the individual Δs.

3.2. Covariate Adjustment via Multivariable Regression

Consider the simple linear model $\mathrm{E}(\, \boldsymbol{y} \mid \boldsymbol{t}, \boldsymbol{X}\,) = \mu\, \boldsymbol{1} + \alpha\, \boldsymbol{t} + \boldsymbol{X}\, \beta$, where $\boldsymbol{1}$ is a dummy vector of all ones, $\boldsymbol{t}$ is a dummy vector of zeros and ones, and ***X*** represents the matrix with eight columns of baseline patient characteristics. Fitting this model yields an estimate of α that provides a homogeneous estimator of the individual Δs.

Quite simple generalizations of the above model may (i) include interactions between ***t*** and the columns of ***X*** or else (ii) drop the "α ***t***" term but allow the β vector to be different for treated and control patients. Models fitted in this second way can produce two ***y***-predictions for each patient: one as if treatment (***t*** = 1) were chosen, and the other for the control choice (***t*** = 0). Both of these generalizations would then provide heterogeneous estimates of individual Δs. Such generalizations (and the strategies needed to fit them) are typically considered inappropriate for use in clinical trial settings. Although severe model selection bias can be injected into analyses of observational data in this way, published accounts rarely admit that models (and inferential statistics such as p-values) other than the one reported were also considered.

3.3. Estimation of Propensity Scores via Logistic Models

Logistic regression models are commonly used to predict propensity scores (PS), which are the conditional probabilities, ***p***, of treatment choice ***t*** = 1 given ***X.*** These models assume that the vector of conditional log odds, log[p/(1–p)], is linear of the form μ ***1*** + ***X*** β, where ***1*** is again a dummy vector of all ones, and ***X*** represents the matrix with eight columns of baseline patient characteristics. Propensities estimated in this way are typically used in analyses of observational data in one of the three ways described in subsections §3.3.1-3.

A fourth possibility is to use PS estimates to match treated and control patients in some fixed ratio, usually 1:1. This strategy is not evaluated here because data from many patients would thereby be (randomly?) "dropped" from each such analysis. After all, in our simulations, true propensity is constant within 300 hidden X-space sub-regions but varies somewhat uniformly from 0.25 (1:3) to 0.75 (3:1) across these sub-regions. As a result, roughly 25% of patients could typically be assigned a missing value as their Δ–estimate when doing 1:1 matching.

3.3.1 Using an Estimated Propensity in Covariate Adjustment

Propensities estimated via a logistic model are non-linear functions of patient baseline **X**-characteristics. Thus, they can typically be included in a regression model for predicting treatment **y**-outcomes, along with the given **X**-variables, without necessarily causing serious ill-conditioning (multicolinearity). This simple twist on traditional covariate adjustment (§3.2) apparently constitutes the most frequent use of propensity scores in published analyses of observational data [35,36] – quite possibly simply because such analyses are quite easy to do using current statistical software. These recent reviews point out that the findings in most such studies tend to change rather little due to augmenting **X** with **p** …quite possibly because all such fitted models typically demonstrate considerable lack-of-fit.

Here, we evaluate the rMSE loss resulting from replacing the **X** matrix by only this **p** vector. The regression model then becomes E(**y** | **t, p**) = μ **1** + α **t** + β **p**, where β is a scalar. After all, this is the usage originally suggested by propensity theory [24,25], and our simulation will verify that little or nothing is lost by making this substitution.

3.3.2. Weighting by the Inverse Probability of Treatment Received

The unbiasedness condition developed by Bang and Robins [37] suggests that y-outcomes from treated and untreated patients need to be "weighted" differently. Specifically, let δ denote a hypothetical binary random variable independent of the y-outcome and such that $\delta = 1$ when a patient chooses treatment ($t = 1$) and $\delta = 0$ when that patient chooses control ($t = 0$). Under the restriction that the probability, p, that this patient chooses treatment is strictly > 0 and < 1, it follows that $\delta \times y / p$ and $(1-\delta) \times y / (1-p)$ are unbiased estimates of the conditional expected outcomes given the treatment and control choices, respectively.

The above observation intuitively "suggests" use of weighted least squares to estimate outcome effects, where weights are taken to be inversely proportional to each patient's estimated propensity for choice between treatment and control. While this sort of weighting is indeed used in both "doubly robust" estimation [37-39] and in LC, it remains unclear whether use of weighted least squares can be "fully justified" by a result that concerns unbiased (rather than minimum variance) estimation. In fact, the simulation results of Freedman and Berk [40] as well as the results presented here are highly unfavorable to use of traditional PS estimates in defining least squares weights.

In all fairness, problems associated with "stabilizing" regression weights are well know, and improved methods have been proposed [37, 41-43]. In other words, the rMSE loss associated with only a relatively naïve way to re-weight patients is being evaluated here.

3.3.3. Propensity Score Stratification (Binning)

Rosenbaum and Rubin [26] proposed using propensity score estimates to rank-order patients and to thereby form as few as five "bins" of adjacent patients, the number of subclasses suggested by Cochran [19]. When one's statistical model uses a linear functional, $\boldsymbol{x'\beta}$, the (three or more) resulting "interior" bins then correspond to all patients on and/or between a pair of parallel, linear hyperplanes. Patient subclasses formed without using explicit X-space boundaries are typically simply called "strata."

Given the capabilities of today's statistical computing environments, analysts can easily use more than five patient subclasses and/or relatively sophisticated within-and-across-subclass data summarization tools, such as bootstrapping. The bootstrap stratification approach whose rMSE loss is evaluated here [44,45] uses estimated propensity score deciles from logistic regression for predicting treatment choice from all baseline ***x***-variables.

3.4. Local Control using Patient Clustering to form many Subgroups

The essence of the LC approach is easy to explain and can be fully appreciated by non-technical audiences. LC starts by dividing all patients (treated or control) up into many subgroups, most typically using some form of traditional clustering [20, 21] on some subset of all available ***x***-variables.[1] Within each resulting subgroup that contains both treated and control patients, a Local

[1] Technically, any way to form subgroups can be used in LC that (i) assures that, within each subgroup, patients are relatively well-matched on their baseline X-characteristics and (ii) avoids deliberate creation

Treatment Difference, LTD = (Avg. Outcome on Treatment) – (Avg. Outcome on Control), is computed, providing a homogeneous estimate of the true Δs for those particular patients. However, LTD estimates ultimately become heterogeneous when merged across all subgroups, yielding a full "LTD distribution" that can be displayed as a histogram or as a Cumulative Distribution Function (CDF.)

If patient subgroups could be formed using hidden ***Z***-information as well as observed baseline ***x***-variables, it would be intuitively clear from first principles that the resulting LTDs would then be unbiased estimates of the corresponding local Δ(***X***,***Z***) main-effects. As alluded to in §3.3.2, LTDs correspond to local "doubly robust" estimates when the statistical model is nested ANOVA (treatment within subgroups formed using only ***x***-variables.) In other words, the local (within subgroup) estimates of p and (1–p) are then the local fractions of patients choosing treatment and control, respectively.

While some relatively sophisticated strategies and tactics for LC analyses have been proposed [46-49], one very simple "special case" of LC will be the primary focus of our rMSE loss comparisons. Since each replication in our simulation will use pseudo-observational data on roughly 25K patients, the LC specification we will evaluate uses hierarchical, complete-linkage clustering on all baseline ***x***-variables (in their standardized form) to form 1K patient subgroups. In other words, theaverage number of patients per cluster will be approximately 25, which is slightly larger than the LC recommended minimum average values of 8 to 10 [49].

On the other hand, as part of a secondary objective to illustrate the variance-bias trade-offs that can result from reducing the number of subgroups formed in an LC analysis, we will also display the rMSE losses corresponding to forming 200, 300, 400 or 500 clusters from the same hierarchical tree used to form 1K clusters within each simulation replication.

4. POTENTIAL FOR VARIANCE-BIAS TRADE-OFFS

Of the six alternative analysis approaches compared here, only the LC approach focuses upon estimation of heterogeneous treatment effects.[2] On the other hand, LC estimates are typically homogeneous within patient subgroups, so the degree of potential heterogeneity in LC estimates is thus determined by the number of subgroups actually used in a LC analysis.

4.1. Uniform Shrinkage towards the Main-Effect of Treatment

A relatively simple way to illustrate possible MSE trade-offs is to consider so-called "shrinkage"

of "pure" subgroups (containing only treated or only control patients) in those sub-regions of X-space that lie within the common support of both treatment cohorts. Approaches that either prohibit formation of subgroups with fewer than, say, 10 or 25 patients or which penalize formation of small subgroups can be helpful.

[2] Again, as discussed in section §3.2, multivariable regression models can produce heterogeneous estimates but are not traditionally used for this purpose.

estimators, Obenchain [50]. In direct analogy with the heterogeneous $\Delta(\boldsymbol{X},\boldsymbol{Z})$ effects discussed here in Section §3, this particular shrinkage reference discusses the estimation of true potencies that vary from batch-to-batch within an industrial production process. The basic idea is to use a constant, multiplicative factor, δ, which is strictly positive ($0 < \delta$) but less than one ($\delta < 1$), to literally shrink assay results from individual production batches towards their long range average. Which numerical value of δ then minimizes MSE risk?

A well-defined and intuitive answer is provided by the following random effects formulation. Suppose that the true heterogeneity of effects has "between" subgroup variance $\boldsymbol{\sigma^2_B}$ while the local "within" subgroup variance of estimation is $\boldsymbol{\sigma^2_W}$. The MSE optimal value for δ is then $\boldsymbol{\sigma^2_B / (\sigma^2_B + \sigma^2_W)}$.

In particular, suppose that the local "within" subgroup variance, $\boldsymbol{\sigma^2_W}$, is considerably larger than the corresponding "between" subgroup heterogeneity variance, $\boldsymbol{\sigma^2_B}$. Considerable shrinkage of estimates of heterogeneous $\Delta(\boldsymbol{X},\boldsymbol{Z})$ effect estimates towards their "main effect" is MSE optimal in this case. This happens whenever patient subgroups are numerous and thus are relatively small. Using many subgroups tends to increase LC estimate heterogeneity but also tends to increase $\boldsymbol{\sigma^2_W}$. Unfortunately, from a MSE reduction perspective, increasing $\boldsymbol{\sigma^2_W}$ encourages shrinkage that would then reduce potential for heterogeneity in LC estimates.

To illustrate this point, let $\boldsymbol{\sigma^2_N}$ denote the variance of the additive white-noise in our simulations, which is always either $(\$1K)^2$ or $(\$5K)^2$. With n_1 and n_0 denoting the local number of treated and control patients, respectively, within a patient subgroup, it follows that $\boldsymbol{\sigma^2_W = \sigma^2_N}$ (n1+ n0) / (n1×n0). In other words, $\boldsymbol{\sigma^2_W \geq \sigma^2_N}$ unless n_1 and n_0 are both at least 2.

Again, $\boldsymbol{\sigma^2_W}$ can always be made smaller by decreasing the number (and increasing the size) of patient subgroups. This tactic tends to increase both n_1 and n_0 and thus decreases $\boldsymbol{\sigma^2_W}$ for any fixed value of $\boldsymbol{\sigma^2_N}$. While this tactic decreases heterogeneity of LC estimates, it also discourages any further shrinkage simply to reduce MSE.

Meanwhile, note that $\boldsymbol{\sigma^2_B}$ is $(\$1.7K)^2$ or less in each of our heterogeneous treatment-effect simulations. This suggests that heterogeneous LC estimates will have much less desirable rMSE loss characteristics when $\boldsymbol{\sigma^2_N}$ is $(\$5K)^2$ rather than the other, much smaller value of $(\$1K)^2$.

4.2. Shrinkage via Choice of Number of Subgroups

The variance-bias trade-offs of primary interest in LC applications result simply from choice of the total number of subgroups formed. Such trade-offs are more subtle than the above illustration using uniform shrinkage towards the single, overall "main-effect" of treatment. Especially when patient subgroups are formed via hierarchical clustering in $\boldsymbol{x}$-space, as they are in our simulations, larger subgroups are literally formed by merging smaller subgroups together. This sort of simple averaging of **y**-outcomes for treated and control patients within larger subgroups yields diffuse shrinkage towards local, heterogeneous targets.

Intuitively, using relatively many (but necessarily smaller) patient subgroups reduces the bias in

LTD estimates. However, using fewer (but larger) subgroups clearly reduces the variances of individual LTD estimates. Thus fewer LC subgroups may well be better in a rMSE loss sense simply because one is thereby *shrinking via additional averaging* of LTD estimates.

5. rMSE LOSS COMPARISONS BETWEEN HETEROGENEOUS AND HOMOGENEOUS TREATMENT EFFECT ESTIMATES

The primary focus of our root MSE loss comparisons will be on situations where the traditional, global, parametric models underlying the approaches of subsections §3.2, §3.3.1 and §3.3.2 are not completely correct models. True counterfactual differences are then also not actually constant within the patient subgroups formed in the PS stratification and LC approaches of subsections §3.3.3 and §3.4. In other words, we will focus on comparisons among alternative approaches in situations where no one approach is exact, but all are reasonable approximations.

Our simulation scenarios are of three basic types, called A, B and C. Our type A scenarios are particularly challenging for all approaches to estimation because they illustrate cases where propensity for treatment choice (t = 1 versus t = 0) is essentially independent of all observed patient baseline ***x***-characteristics …i.e. treatment choice depends at most on hidden ***Z***-factors. In our type B scenarios, the propensity for choice t = 1 tends to steadily increase as the expected ***y***-outcome (total yearly cost) predictable from observed patient baseline ***x***-characteristics increases.

Our type A and B simulations each include the six special cases listed in Table 1. Specifically, the additive white noise has two levels of variance: $(\$1\text{K})^2$ in the odd cases (1, 3 and 5) and $(\$5\text{K})^2$ in the even cases (2, 4 and 6). Meanwhile, the "pmix" factor, which determines how much of ***y***-outcome signal is predictable from patient baseline ***x***-characteristics, has three levels: the true signal is only 10% predictable in the cases 1 and 2, 50% predictable in the cases 3 and 4, but 90% predictable in the cases 5 and 6.

For completeness, our simulations also include a single type C scenario where the true effect of treatment is homogeneous for all patients, the ***y***-outcome signal is 100% predictable from patient baseline ***x***-characteristics, and the additive white noise has variance $(\$1\text{K})^2$. In other words, the traditional, global, parametric modeling approaches are exactly "correct" in our single type C scenario. Scenarios A5 and B5 use pmix = 90% and the same noise level as Scenario C.

Table 2 displays the root MSE loss estimates for 10 alternative approaches to estimation of true counterfactual differences in expected ***y***-outcome for our 13 simulation scenarios. Each scenario was replicated to assure that all rMSE loss estimates are accurate to the nearest dollar ($).

Because the five right-most columns of Table 2 describe approaches yielding only homogeneous estimates of counterfactual differences, their root MSE loss cannot be less that the true standard deviation of the unknown, true counterfactual differences in the 13 individual scenarios. These values are listed in the middle column of Table 2 under the heading "Homo Lower Bound". Note that this value is zero for the type C scenario in the bottom row of the table, but this lower bound exceeds $1K in all of the heterogeneous effect scenarios of types A and B. Note further that this lower bound tends to decrease in special cases 1 through 6 as the "pmix" proportion increases. This results because the hidden (unpredictable) heterogeneity is apparently slightly larger than is the heterogeneity predictable from patient baseline ***x***-characteristics.

Finally, the variance-bias trade-offs discussed in Section §4 predict some of the patterns of rMSE loss variation observed in Table 2. In particular, the LC approach using 1K patient clusters is not competitive in any of the even-numbered type A and B scenarios where the variance of the additive noise, $(\$5K)^2$, greatly exceeds the true variability of treatment effect heterogeneity, which ranges from $(\$1.1K)^2$ to $(\$1.7K)^2$. These sets of one thousand LTDs may well have relatively low bias as estimates of almost 25K heterogeneous counterfactual differences but they clearly also have relatively high variance. On the other hand, the LC approach using 1K patient clusters is a clear overall winner in 5 of the 6 odd-numbered type A and B scenarios where the variance of the additive noise is only $(\$1K)^2$.

As further evidence for potential variance bias trade-offs, Table 2 lists the rMSE loss for our secondary interest in using fewer than 1K clusters in LC. In this regard, note that using 500 patient clusters always yields lower rMSE than 1000 clusters, and rMSE can always be further reduced here by using only 400 LC clusters! In fact, using only 200 LC clusters yields the lowest overall rMSE in 6 of the 12 type A or B scenarios.

5.1 Insights from Type A Scenarios

The most striking result from our 6 type A scenarios is that "no adjustment" yields the lowest overall rMSE loss in scenario A1 as well as in the 3 even-numbered scenarios where the variance of the additive white noise is high, $(\$5K)^2$. In other words, the fact that numerical values of PSlocal -- the local propensity for treatment (choice ***t*** = 1) -- were randomly assigned to the 300 clusters of 40K baseline ***X***-vectors in the type A scenarios makes propensity quite difficult to predict. Needless to say, perhaps, but difficulty in predicting propensity from biased patient samples can also make it difficult to accurately predict ***y***-outcomes as well as counterfactual treatment differences!

In a sense, it's not surprising that LC using 1K clusters yields the lowest overall rMSE loss in the 2 remaining odd-numbered scenarios …where the variance of the additive white noise is low, $(\$1K)^2$, and pmix is 0.5 or 0.9. After all, the LC approach does not rely in any way on model-based estimates of propensity. Once subgroups of patients who are well matched in ***x***-space have been formed, propensity estimates can then, of course, be directly observed as local fractions of patients choosing treatment. But estimation of propensity is a purely optional exercise in the LC approach. In other words, difficulty in estimation of propensity can have little direct impact on the LC approach.

5.2 Insights from Type B Scenarios

In our type B scenarios, "no adjustment" again yields the lowest overall rMSE loss in scenario B2, while "MV Reg" and "PS as Cov" win in the other 2 even-numbered scenarios. Meanwhile, LC with 1K clusters minimizes rMSE in all 3 odd-numbered scenarios …where the variance of the additive white noise is low, $(\$1K)^2$. In all 6 type B scenarios, variance-bias trade-offs could be used to further reduce rMSE loss via using only 200 or 300 clusters in LC.

5.3 Insights from the Type C Scenario

When models assume that true counterfactual differences are homogeneous and that assumption is a correct one, estimates from those models can achieve really low rMSE. The key difference between scenarios B5 and C is that the lower bound on rMSE loss of homogeneous estimates thereby drops from $1,342 to $0. As suggested by the propensity scoring work of Rosenbaum and Rubin [24,25], adjustment using a single covariate (the estimated propensity for choice $\boldsymbol{t}$ = 1) can do just as well in this case as a multivariable regression model that includes all observed patient baseline ***x***-characteristics. Even the naïve IPW estimator does better here than LC with 1K patient clusters. In fact, the really obvious loser (rMSE = $1,928) in this highly special case is failure to make any adjustment for treatment selection bias.

6. OTHER ADVANTAGES OF PATIENT SUBGROUPING APOPROACHES

This paper has focused upon demonstrating that the Local Control approach to estimation of heterogeneous counterfactual differences in treatment effects can be more accurate, in an rMSE sense, than homogeneous (main-effect only) estimators whenever the true heterogeneity of counterfactual treatment differences is larger than any purely random noise in the data. Luckily, the LC approach also offers many other advantages.

Parametric models can be a big help in situations where data are not sufficiently plentiful to be locally "dense." In these cases, models can interpolate within or extrapolate beyond the available data. Unfortunately, in traditional approaches to analysis of large, observational datasets [3,4,14], a common tendency is to ignore the lack-of-fit in models and focus, instead, upon the statistical significance of their estimates.

Neither patients nor their caregivers should really care very much about what some poorly-fitting model says their ***y***-outcome "should" be given their known ***x***-characteristics. Rather, and especially as observational data become more and more plentiful, patients should want to know what ***y***-outcomes have recently been experienced by other patients "like me" [51]. Given adequate data, oversimplifications from global, parametric models are not needed to address the health care information needs of individual patients. In LC, local estimates (LTDs) become the building blocks of a realistic, overall analysis.

The greatest strengths of the LC approach are its (nonparametric) simplicity and focus on the information truly needed by individual patients, caregivers and health policy makers. For example, suppose that the ***y***-outcome of interest is an adverse event rate, and a LC analysis yields an observed LTD distribution with a long left-hand tail of negative differences and a long right-hand tail of positive differences in adverse event rates. This means that patients in the left-hand tail experience fewer adverse events on treatment (choice $\boldsymbol{t}$ = 1), while patients in the right-hand tail experience fewer adverse events on control (choice $\boldsymbol{t}$ = 0.)

Finally, using the above information from an LC analysis, suppose that further research shows that it is possible to fairly accurately predict in which LTD tail certain individual patients will fall. Suddenly, the question of which treatment choice has the higher overall adverse event rate has become medically moot. Instead, all ethical attention should then be focused on using individualized medicine to assure that the left-tail patient types receive treatment while the righttail types receive control. Furthermore, this particular treatment choice will also be seen as relatively unimportant for some patients, specifically, those whose LTDs are near zero.

The new emphasis on Comparative Effectiveness in health care research needs to compare alternative treatments head-to-head using more meaningful and relevant criteria than simply traditional main-effects. This will require improved statistical methods [14], sound and realistic statistical thinking [52] and avoidance of deliberate underrepresentation of uncertainty in findings from observational studies [53,54].

References:

[1] K. M. Thompson, Risk In Perspective: Insight and Humor in the Age of Risk Management, Newton Centre, MA: AORM (2004).

[2] D. C. Des Jarlais, C. Lyles and N. Crepaz, Improving the reporting quality of nonrandomized evaluations of behavioral and public health interventions: the TREND statement, Am J Public Health 94 (2004), 361-366. (Transparent Reporting of Evaluations with Nonrandomized Designs, http://www.trend-statement.org)

[3] E. von Elm and M. Egger, The scandal of poor epidemiological research [Editorial], BMJ 329 (2004), 868–869.

[4] S. J. Pocock, T. J. Collier, K. J. Dandreo, B. L. de Stavola, M. B. Goldman, L. A. Kalish, L. E. Kasten and V. A. McCormack, Issues in the reporting of epidemiological studies: a survey of recent practice, BMJ 329 (2004), 883–888.

[5] J. P. A. Ioannidis, Contradicted and initially stronger effects in highly cited clinical research, JAMA 294 (2005), 218–229.

[6] J. P. A. Ioannidis, Why most published research findings are false, PLoS Med 2 (2005), e124. doi:10.1371/journal.pmed.0020124.

[7] E. von Elm, D. G. Altman, M. Egger, S. J. Pocock, P. C. Gøtzsche and J. P. Vandenbroucke, Strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies, BMJ 335 (2007), 806–808. (http://www.strobe-statement.org)

[8] D. M. Kent and R. A. Hayward, When averages hide individual differences in clinical trials, American Scientist 95 (2007), 60–68.

[9] D. M. Kent and R. A. Hayward, Limitations of applying summary results of clinical trials to individual patients: the need for risk stratification, JAMA 298 (2007), 1209–1212.

[10] T. Siegfried, Odds Are, It's Wrong: Science fails to face the shortcomings of statistics. Science News 177 (2010), 26-29.

[11] J. Lehrer. The Truth Wears Off: Is There Something Wrong with the Scientific Method? New Yorker December 13 (2010).

[12] S. Begley. Why Almost Everything You Hear About Medicine Is Wrong, Newsweek January 24 (2011).

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

[14] M. van der Laan and S. Rose, Statistics Ready for a Revolution: Next Generation of Statisticians Must Build Tools for Massive Data Sets, AMStat News September (2010), 38-39.

[15] S. H. Kaplan, J. Billimek, D. H. Sorkin, Q. Ngo-Metzger and S. Greenfield, Who Can Respond to Treatment? Identifying Patient Characteristics Related to Heterogeneity of Treatment Effects, Medical Care 48 (2010), S9–S16 [Comparative Effectiveness.]

[16] S. J. Ruberg, L. Chen and Y. Wang, The mean does not mean as much anymore: finding sub-groups for tailored therapeutics, Clin Trials 7 (2010), 574–583.

[17] S. J. Pocock, S. E. Assmann, L. E. Enos and L. E. Kasten, Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practice and problems, Stat Med 21 (2002), 2917–2930.

[18] W. G. Cochran, The planning of observational studies of human populations (with Discussion), J Roy Stat Soc A 128 (1965), 234–266.

[19] W. G. Cochran, The effectiveness of adjustment by subclassification in removing bias in observational studies, Biometrics 24 (1968), 295–313.

[20] H. B. Barlow, Unsupervised Learning, Neural Computation 1 (1989), 295–311.

[21] L. Kaufman and P. J. Rousseeuw, Finding Groups in Data: An Introduction to Cluster Analysis, New York: John Wiley and Sons, (1990).

[22] R. L. Obenchain, Unsupervised Propensity Scoring: NN and IV Plots, 2004 Proceedings of the American Statistical Association on CD-ROM. Alexandria, VA: American Statistical Association, (2005).

[23] R. L. Obenchain, Identifying Meaningful Patient Subgroups via Clustering - Sensitivity Graphics, 2006 JSM Proceedings on CD-ROM. Alexandria, VA: American Statistical Association, (2007).

[24] C. Fraley and A. E. Raftery, Model-based clustering, discriminant analysis, and density estimation, J Am Stat Assoc 97 (2007), 611–631.

[25] P. R. Rosenbaum and D. B. Rubin, The central role of the propensity score in observational studies for causal effects, Biometrika 70 (1983), 41–55.

[26] P. R. Rosenbaum and D. B. Rubin, Reducing bias in observational studies using subclassification on the propensity score, J Amer Stat Assoc 79 (1984), 516–524.

[27] L. Q. Yue, Statistical and regulatory issues with the application of propensity score analysis to nonrandomized medical device clinical studies, J Biopharm Stat 17 (2007), 1–13.

[28] D. Westreich, J. Lessler and M. J. Funk, Propensity score estimation: neural networks, support vector machines, decision trees (CART), and meta-classifiers as alternatives to logistic regression, Journal of Clinical Epidemiology 63 (2010) 826-833.

[29] D. B. Rubin, Matching to remove bias in observational studies, Biometrics 29 (1973), 159–184.

[30] B. B. Hansen, Full matching in an observational study of coaching for the SAT, J Amer Stat Assoc 99 (2004), 609–618.

[31] Daniel E. Ho, Kosuke Imai, Gary King and Elizabeth Stuart, Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference, Political Analysis 15 (2007) 199-236.

[32] E. A. Stuart, Matching Methods for Causal Inference: A Review and a Look Forward, Statistical Science 25 (2010), 1–21.

[33] R. L. Obenchain, Q. Hong, A. Zagar and D. E. Faries, Observational Data Simulation Scenarios for Windzorized Yearly Costs of Patients with Major Depressive Disorder, Unpublished Technical Material (2011).

[34] Q. Hong, R. L. Obenchain, A. Zagar and D. E. Faries, An Archive of R-code and Datasets for Comparison of Observational Data Analyses via calls to SAS/STAT Procedures and Macros, Unpublished Technical Materials (2011).

[35] B. R. Shah, A. Laupacis, J. E. Hux and P.C. Austin, Propensity score methods gave similar results to traditional regression modeling in observational studies: a systematic review, J Clin Epidemiol 58 (2005), 550–559.

[36] T. Stürmer, M. Joshi, R. J. Glynn, J. Avorn, K. J. Rothman and S. Schneeweiss, A review of the application of propensity score methods yielded increasing use, advantages in specific settings, but not substantially different estimates compared with conventional multivariable methods. [Review Article] J Clin Epidemiol 59 (2006), 437–447.

[37] H. Bang and J. M. Robins. Doubly robust estimation in missing data and causal inference models, Biometrics 61 (2005), 962–973.

[38] J. K. Lunceford and M. Davidian, Stratification and weighting via the propensity score in estimation of causal treatment effects: A comparative study, Stat Med 23 (2004), 2937–2960.

[39] M. Jonsson-Funk, D. Westreich, M. Davidian and C. Weisen, Introducing a SAS® macro for doubly robust estimation, SAS Global Forum Paper 189 (2007).

[40] D. A. Freedman and R. A. Berk, Weighting Regressions by Propensity Scores, Evaluation Review 32 (2008), 392–409.

[41] Robins JM, Hernan MA, Brumback B. Marginal structural models and causal inference in epidemiology. ***Epidemiology*** 2000; 11: 550–560.

[42] G. Ridgeway, gbm R-package: Generalized Boosted Models, http://www.r-project.org (2005).

[43] G. Ridgeway, D. F. McCaffrey and A. Morral, twang R-package: Toolkit for Weighting and Analysis of Nonequivalent Groups. http://www.r-project.org (2006).

[44] R. L. Obenchain, ICEpsbbs: Incremental Cost-Effectiveness inference via Propensity Score Bin Boot Strapping (two biased samples), ~~http://members.iquest.net/~softrx~~ (2003).

[45] D. E. Faries, X. Peng and R. L. Obenchain, "Analysis of Costs and Cost-Effectiveness Data Using Propensity Score Bin Bootstrapping," Analysis of Observational Health Care Data Using SAS, D. E. Faries, A. C. Leon, J. M. Haro and R. L. Obenchain eds, Cary, NC: SAS Press (2010), 315–337.

[46] R. L. Obenchain, USPS: R-package for Unsupervised and Supervised Propensity Scoring and Instrumental Variable Adjustment, CRAN http://www.r-project.org (2006).

[47] R. L. Obenchain, JMP Scripts for Local Control and Artificial Local Treatment Difference Distributions, ~~http://members.iquest.net/~softrx~~ (2006).

[48] R. L. Obenchain, 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~~ (2009). http://localcontrolstatistics.org

[49] R. L. Obenchain, "The Local Control Approach using JMP," Analysis of Observational Health Care Data Using SAS, D. E. Faries, A. C. Leon, J. M. Haro and R. L. Obenchain eds, Cary, NC: SAS Press (2010), 151–192.

[50] R. L. Obenchain, Common industrial applications of mixed linear models, Proceedings of MWSUG `93, LeRoy Bessler ed, Fox Point, WI: MidWest SAS Users Group (1993).

[51] PCOR Institute, Working Definition of Patient-Centered Outcomes Research, http://www.pcori.org/images/PCOR_Rationale.pdf (May 2011).

[52] J. W. Tukey, Sunset Salvo, Amer Statist 40 (1986), 72–76.

[53] G. Taubes and C. C. Mann, Epidemiology faces its limits, Science 269 (1995), 164-169.
[54] N, E. Breslow, Are Statistical Contributions to Medicine Undervalued? Biometrics 59 (2003), 1–8.

Table 1. Six Simulation Special Cases

Scenario Number	pmix: Mixture Proportion	Additive Sigma Noise ($)
1	0.1	1000
2	0.1	5000
3	0.5	1000
4	0.5	5000
5	0.9	1000
6	0.9	5000

Table 2. Bias in the Estimation of the Overall Main-Effect (ME) of Treatment for Six Analytical Methods in Thirteen Observational Data Simulation Scenarios. Bias is expressed on Dollars ($).

Simulation Scenario	True ME of Trtm	Bias of LC	Bias of MVREG	Bias of PCOV	Bias of IPW	Bias of PBIN	Bias of UNADJ
A1	-1099	-151	-388	-388	-2017	-379	-377
A2	-1099	-147	-387	-387	-2011	-377	-376
A3	-878	-120	-308	-308	-1404	-277	-202
A4	-878	-114	-302	-302	-1395	-271	-196
A5	-658	-90	-229	-229	-798	-177	-27
A6	-658	-65	-196	-196	-735	-143	4
B1	-676	168	239	239	-3034	241	193
B2	-676	171	240	240	-3027	242	195
B3	-650	113	112	109	-973	219	906
B4	-650	114	113	110	-981	220	906
B5	-624	58	-15	-22	1090	195	1619
B6	-624	56	-14	-20	995	197	1592
C1	-1000	86	68	63	704	332	1928

Color Coding:

BEST (least biased) Approach
Second BEST Approach
Third BEST …but clearly better than other three approaches.

Bias in the PBIN approach is uniformly larger (and in the same direction) as that of LC.
Bias of the IPW approach is always largest or second largest …and more erratic in sign.

Table 3. Simulated root Mean Squared Error in the Estimation of Heterogeneous Treatment Effects for Six Analytical Methods in Thirteen Scenarios. rMSE values are expressed in Dollars ($).

Scenario	LC 200	LC 300	LC 400	LC 500	LC 1000	Homo Lower Bound	MV Reg	PS as Cov	PS Bin Boot	IPW	UN-Adj
A1	1666	1680	1712	1750	1887	1680	1725	1725	1723	2625	1722
A2	1888	2009	2140	2268	2808	1680	1725	1725	1723	2622	1723
A3	1264	1239	1233	1235	1257	1343	1378	1378	1372	1944	1359
A4	1545	1657	1778	1902	2435	1343	1378	1378	1372	1939	1359
A5	1047	1027	1019	1023	1042	1097	1121	1121	1111	1360	1099
A6	1363	1486	1617	1746	2291	1097	1116	1116	1108	1327	1100
B1	1286	1338	1401	1451	1617	1626	1644	1644	1644	3443	1638
B2	1580	1753	1915	2070	2651	1626	1645	1645	1645	3438	1639
B3	958	954	962	976	1033	1447	1452	1452	1464	1746	1708
B4	1332	1479	1627	1763	2338	1447	1453	1453	1465	1755	1709
B5	831	821	828	837	889	1342	1342	1342	1357	1729	2103
B6	1215	1372	1519	1657	2223	1342	1344	1344	1358	1675	2083
C	607	639	673	704	819	0	68	63	332	704	1928

Highlighting Legend:

GREEN background denotes the lowest rMSE of the six primary approaches.

GYAN background denotes lower rMSE values in the secondary comparison of LC with only 200, 300, 400 or 500 clusters to the six primary approaches.