Local Control Strategy: Simple Analyses of Air Pollution Data can reveal Heterogeneity in Longevity Outcomes

Robert L. Obenchain, Risk Benefit Statistics, wizbob@att.net
S. Stanley Young, CGStat, genetree@bellsouth.net

Abstract

Claims from observational studies that use traditional model specification searches often fail to replicate, partially because the available data tend to be biased. There is an urgent need for an alternative statistical analysis strategy that is not only simple and easily understood but also is more likely to give reliable insights when the available data have not been designed and balanced. The alternative strategy known as Local Control first generates local, nonparametric effect-size estimates (fair treatment comparisons) and only then asks whether the observed variation in these local estimates can be predicted from potential confounding factors. Here, we illustrate application of Local Control to a historical air pollution dataset describing a "natural experiment" initiated by the federal Clean Air Act Amendments of 1970. Our reanalysis reveals subgroup heterogeneity in the effects of air quality regulation on elderly longevity (one size does not fit all), and we show that this heterogeneity is largely explained by socio-economic and environmental confounders other than air quality.

KEY WORDS: Local Control, Observational Study, Heterogeneous Treatment Effects, Researcher Incentives

1. Introduction

There is extensive literature on the question: Does air quality have health effects?[1,2] The vast majority of published papers find overall associations between air quality and health effects (death). A few papers use relatively sophisticated analyses to make the case that, when potential biases are carefully taken into account, little association between air quality and longevity remains[3-7]. Even when the association between improvements in air quality and increased longevity is positive on overall average, this association can still be negative in some identifiable subgroups of locations within the US.[8] Logically, only one such true negative is required to invalidate any causal claim that implies uniformly positive associations.

The statistical analysis strategy illustrated here is known as *Local Control* (LC)[9-14]. Basic LC strategy is quite simple and easily appreciated, even by nontechnical audiences. It focuses upon generation of visual displays of distributions of local effect-sizes that allow all stakeholders to literally see and

evaluate the variability and uncertainty in local treatment effect-sizes. Each of the statistical "tactics" used in LC is a well-established method (clustering of experimental units to form blocks, nested ANOVA for treatment within blocks, permutations for resampling without replacement, multivariable model fitting, etc.) As is illustrated within this tutorial, LC merely combines traditional statistical tactics into a coherent strategy for improved analysis of data, where the treatment cohorts to be compared have different distributions of confounding factors.

LC strategy looks beyond treatment "main effects" for two main reasons. Clustering of experimental units on their pre-treatment confounding characteristics reveals local interaction effects that can be seen as variation in treatment effect-sizes across clusters. Visual displays of realistic information about observed effect-size variation can be much more relevant to stakeholders than traditional point estimates for means and variances or their p-values. Secondly, modeling of local effect-size estimates yields inferences about their statistical distribution, classifying observed variation as either mostly *random (unexplained)* or else satisfactorily *predictable from observed confounding factors (truly heterogeneous).*

As typically practiced today[15,16], observational research is quite unlike (randomized) clinical research, where all hypotheses and statistical models are required to be pre-specified before any supporting data are collected. Instead, after many alternative models for observational data have been explored, a single model (say, smallest p-value for the main-effect of treatment) is all-too-frequently presented almost as if it had been pre-specified. In other words, no accounting for potential treatment selection bias[17] is typically attempted. Although common modeling "omissions" have been noted[18-20], they continue to be widely practiced.

Another typical problem[21] is that the only models considered are actually "wrong" in the sense of being deliberately over-simplified approximations. For example, treatment-confounder interaction terms are absent or else only interactions of the simple "multiplicative" (***t*×*x***) form are explored. In summary, we fear that over-fitting of wrong models is a primary source of many published, "statistically significant" observational claims that are likely to fail-to-replicate if retested using the same analytical approach on different/new data.

2. CAAA Data used to Illustrate LC Analysis Strategy

In this tutorial, we reanalyze some historical air quality and longevity data[3] graciously provided to us

by Carlos Dobkin. The Clean Air Act Amendments (CAAA) of 1970 designated US counties with annual, average Total Suspended Particulates (TSPs) concentrations exceeding a federally determined threshold (typically, geometric mean of TSP > 75 $\mu g/m^3$) as nonattainment locations. This legislation created a natural experiment[22] because polluters in these nonattainment locations faced stricter regulations (starting in 1972) than polluters in the attainment locations that comprise the remainder of the dataset. Here we focus on a subset of the full data[3] that covers 560 US counties (identified by their 4 or 5 digit FIPS codes) for six consecutive years, 1969 through 1974. However, to keep our analyses as simple as possible, we consider only averages over specified time-periods here. In other words, our simple LC analyses will be essentially cross-sectional rather than truly longitudinal.

Because annualized county-level measurements representing consecutive years tend to be positively correlated, less precision is gained by forming 3-year *averages* (1969-1971 for the CAAA pre-period or 1972-1974 for the CAAA post-period) than when averaging uncorrelated measures. On the other hand, taking the *difference* between pre- and post-CAAA averages then tends to increase (rather than decrease) precision in the resulting *change* estimates, again due to remaining positive correlation between pre- and post-CAAA averages.

Like the original researchers[3], we too will examine potential effects of observed changes in life expectancy due to county-level differences in CAAA compliance attainment and/or ***X***-confounding factors. Since LC focuses upon treatment effect-sizes that quantify differences in ***Y***-outcomes between nonattainment (treatment) and attainment (control) locations, this focus on "differences between correlated measures of change between consecutive 3-year periods" increases precision as well as simplifies our analyses.

As factors possibly confounding observed relationships between longevity and air pollution, Chay, Dobkin and Greenstone[3] considered 15 ***X***-measures from the Regional Economic Information System (REIS), US Bureau of Economic Analysis (BEA.) Our preliminary analyses confirmed that 4 of these 15 measures appear to be much more relevant than the other 11. All 15 REIS variables are included within the data sharing archive[14] we have created. In analyses presented here, all ***X***-confounder measures are expressed as 6-year averages (1969-74.)

Table 2.1 displays very simple (unadjusted) statistical comparisons between CAAA compliance cohorts for 21 variables. The available data cover 286 nonattainment counties and 274 attainment counties, a

Table 2.1 Significance of Differences between Nonattainment and Attainment Cohorts

Changes Between 3-year Averages: (1972-74) minus (1969-71).	Nonattainment Mean ± Std Err	Attainment Mean ± Std Err	p-Value of t-test
Longevity, Adults over 50: Deaths per 10K	-6.92 ± 0.992	-4.45 ± 1.265	0.1261
Longevity, Elderly 65-84: Deaths per 10K	-20.70 ± 1.686	-16.48 ± 2.457	0.1574
Arithmetic Means of TSP in $\mu g/m^3$	-18.19 ± 3.836	2.64 ± 1.442	<.0001
Geometric Means of TSP in $\mu g/m^3$	-15.22 ± 2.686	2.26 ± 0.984	<.0001

Primary REIS Confounding Factors: 6-year Averages (1969-74)	Nonattainment Mean ± Std Err	Attainment Mean ± Std Err	p-Value of t-test
Yearly Earnings, $	8168.26 ± 99.412	7432.69 ± 100.196	<.0001
Employment Fraction in Manufacturing	0.10 ± 0.003	0.08 ± 0.003	<.0001
Medicare Payments, $	91.43 ± 1.925	90.09 ± 2.155	0.6441
Unemployment Insurance Payments, $	49.59 ± 1.990	53.58 ± 2.142	0.1728

Secondary REIS Confounding Factors: 6-year Averages (1969-74)	Nonattainment Mean ± Std Err	Attainment Mean ± Std Err	p-Value of t-test
Fraction of Population Employed (EPOP)	0.46 ± 0.008	0.43 ± 0.008	0.0025
Family Assistance Payments, $	56.98 ± 2.607	46.00 ± 2.090	0.0011
Food Stamp Payments, $	18.21 ± 0.992	18.64 ± 1.337	0.7971
Income Maintenance Payments, $	118.80 ± 4.423	112.22 ± 4.780	0.3131
Military Medical Benefits, $	4.72 ± 0.338	6.20 ± 0.520	0.0170
Other Income Benefits, $	10.18 ± 0.735	9.50 ± 0.711	0.5035
Public Medical Assistance, $	69.69 ± 3.370	64.30 ± 3.276	0.2519
Retirement Benefit Payments, $	847.75 ± 11.200	864.14 ± 13.689	0.3548
Social Security Payments, $	35.86 ± 2.111	42.23 ± 2.792	0.0694
Total Medical Payments, $	165.21 ± 4.675	158.44 ± 4.649	0.3056
Transfer Payments, $	1016.12 ± 15.022	1029.91 ± 17.156	0.5456

Air Quality Confounding Factors: 6-year Averages (1969-74)	Nonattainment Mean ± Std Err	Attainment Mean ± Std Err	p-Value of t-test
Overall Arithmetic Mean of TSP in $\mu g/m^3$	92.81 ± 1.927	57.28 ± 0.974	<.0001
Overall Geometric Mean of TSP in $\mu g/m^3$	79.71 ± 1.381	49.72 ± 0.818	<.0001

rather small imbalance in compliance cohort sizes. More importantly, these two cohorts differ significantly, even in mean value alone, on 4 of the 8 outcome, treatment and confounder variables we will focus on in this tutorial. For example, while differences between 3-year-average changes in longevity are not significant, note that the corresponding changes in air quality clearly are significant:

Relative to attainment counties, the CAAA legislation of 1970 dramatically decreased TSP pollution in nonattainment counties between consecutive 3-year periods.

3. LC Phase One: Nonparametric Preprocessing of Observational Data

LC strategy dictates that the first phase of analysis be confined to statistical methods that are primarily nonparametric (make realistic but minimal assumptions) and yet are highly informative and descriptive about treatment cohort imbalances. The statistical tactics used in this initial "Nonparametric Preprocessing" phase[23-25] of LC Strategy are outlined in subsections §3.1 to §3.4 below.

3.1 Restriction to "Fair" Treatment Comparisons

A Fair Treatment Comparison (FTC) is defined[12] to be a comparison of expected ***Y***-outcomes between a given pair of treatment ***T***-cohorts at a given vector of ***X***-confounder characteristics of the general form:

$$\textbf{FTC Estimand at } X \;=\; E[\,(\,Y \mid T{=}1\,) - (\,Y \mid T{=}0\,) \mid X\,]. \qquad (3.1)$$

Note that the ***Y***-outcome is the only quantity varying randomly in expression (3.1), and the expectation is taken with respect to the conditional distribution of ***Y*** given both possible ***T*** choices and a single ***X***-covariate vector. Our motivation for calling this a "fair" comparison (apples-to-apples) is that this estimand defines the true difference in ***Y***-outcomes between the two possible treatment ***T***-choices that correspond to exactly the *same* ***X****-vector*. A similar comparison using $X = X_1$ for T=1 and $X = X_0$ for T=0 would be an apples-to-oranges comparison whenever X_1 and X_0 represent a poorly matched pair of confounder vectors (here, county REIS characteristics.)

The concept of a FTC at ***X*** is essential for improved analyses of observational studies, especially those that address conditional/individualized treatment choices rather than one-size-fits-all recommendations. This is easily demonstrated by an obvious (but usually ignored) shortcoming in the traditional definition of the treatment "main" effect.

$$\textbf{Average Treatment Effect (ATE) Estimand} = E[\,(\,Y \mid T{=}1\,) - (\,Y \mid T{=}0\,)\,] \qquad (3.2)$$

$$= \; E(\,Y \mid T{=}1\,) - E(\,Y \mid T{=}0\,). \qquad (3.3)$$

A well-known result from elementary statistics and probability theory is that the right-hand-side of equation (3.2) is the difference in *conditional* <u>*marginal*</u> *expectations* of the outcome given alternative

treatment choices, (3.3). Specifically, an ATE estimand can be misleading because it ignores potentially important details of the *joint probability distribution* of ***Y***, ***T*** and ***X***. In other words, the ATE estimand makes an "unfair" (apples-to-oranges) overall comparison whenever the two conditional distributions of ***Y*** given only ***T*** actually have substantially different marginal ***X***-distributions[12], typically due to low *common support*[17].

In summary, estimation of FTC effects as ***X*** varies is key to improved analyses of observational studies because bias from severe imbalance on confounding factors between treatment cohorts is commonplace. Individual fair treatment comparisons at a single, given ***X***-vector are inherently local; global fair comparisons can then be formed as (weighted) averages of these local effects as ***X*** varies[12].

3.2 Microaggregation of Experimental Units

Micro-aggregation[26] of data using Ward clustering is an established method of statistical disclosure control and personal privacy protection. LC uses this tactic in its initial nonparametric, unsupervised learning[27] phase to (a) form many, small subgroups of experimental units with highly similar ***X***-confounding vectors and, thus, to (b) provide local effect-size estimates that approximate a collection of FTC estimands of form (3.1).

Unlike methods that use matching of experimental units simply as a "getting-started" tactic[23,24], LC never discards any of the available data simply to balance treatment cohorts. Instead, LC strategy utilizes established statistical tactics that preserve power while providing sound and simple research guidance. Specifically, unbiased estimation of local effect-sizes does not require local balance; see equation (3.4) below.

In our example application of LC, we use Ward clustering to divide the 560 available counties on the four primary REIS confounding variables listed in Table 2.1 into 25 mutually exclusive and exhaustive subgroups of counties. Counties within the same subgroup are not necessarily close together geographically ...but are *relatively* close together in primary REIS-confounder ***X***-space.

Importantly, clustering counties on only their ***X***-vectors ignores all available data on ***Y***-outcomes and treatment cohort membership. Again, clustering is a form of unsupervised learning[27], and important observational study "design" advantages[28] result from this restriction. Because traditional covariate adjustment methods of model fitting are forms of supervised learning, the statistical validity of their

estimates depends critically upon strong assumptions; *their parametric model must be specified correctly*[21].

Because TSP thresholds were used to determine CAAA compliance in 1971, it is thus essential to avoid bias that could result from also using TSP measurements as continuous ***X***-confounding factors in our initial phase (unsupervised) LC analyses. Since polluters within the nonattainment counties faced increased air quality regulation, a significantly larger average decrease in TSP did indeed occur during the 1972 -74 period within the 286 county nonattainment (treatment) cohort, ***T***=1. In fact, TSP pollution actually increased on overall average during this period within the 274 county attainment (control) cohort, ***T***=0; see Table 2.1. Thus TSP is highly likely to partially predict CAAA compliance status and, thereby, function as a (supervised) propensity score estimate[29]. On the other hand, we will definitely consider including TSP level as a potential, continuous ***X***-confounder in the final (supervised) phase of LC strategy described and discussed here in sections §4 and §5.

Some Clustering Details:

- Twenty six counties had to be excluded from clustering because they had missing values for one or more of their 4 primary REIS confounders. Of these 26 counties, 21 were in Virginia, and 17 were in the attainment cohort.
- The county containing New York city has unique confounding characteristics; it falls into an uninformative cluster of size one.
- An informative cluster of size 2 consists of just the two Alaskan counties; Anchorage (attainment) is paired with Fairbanks North Star (nonattainment.)
- The 9 smallest informative clusters each contained 2 to 16 counties; the 9 largest informative clusters each contained 26 to 53 counties.
- The observed, within-cluster fraction of nonattainment counties (local propensity for nonattainment) varied from a low of 0.2353 for a cluster containing 17 counties to a high of 0.9091 for a cluster containing 22 counties.

3.3 Local Treatment Difference (LTD) Distributions of Effect-Sizes

Again, micro-aggregation forms local subgroups (clusters) of experimental units that are relatively

well-matched on their specified ***X***-confounder vectors. Since the average ***X***-vector within a cluster defines its centroid, the FTC estimand of (3.1) at this ***X***-point serves as the asymptotic target vector for a Local Treatment Difference (LTD) estimator of the form:

LTD Estimate within Cluster = (Average *Y*-outcome when *T*=1 and *X* is within Cluster) minus (Average *Y*-outcome when *T*=0 and *X* is within Cluster). **(3.4)**

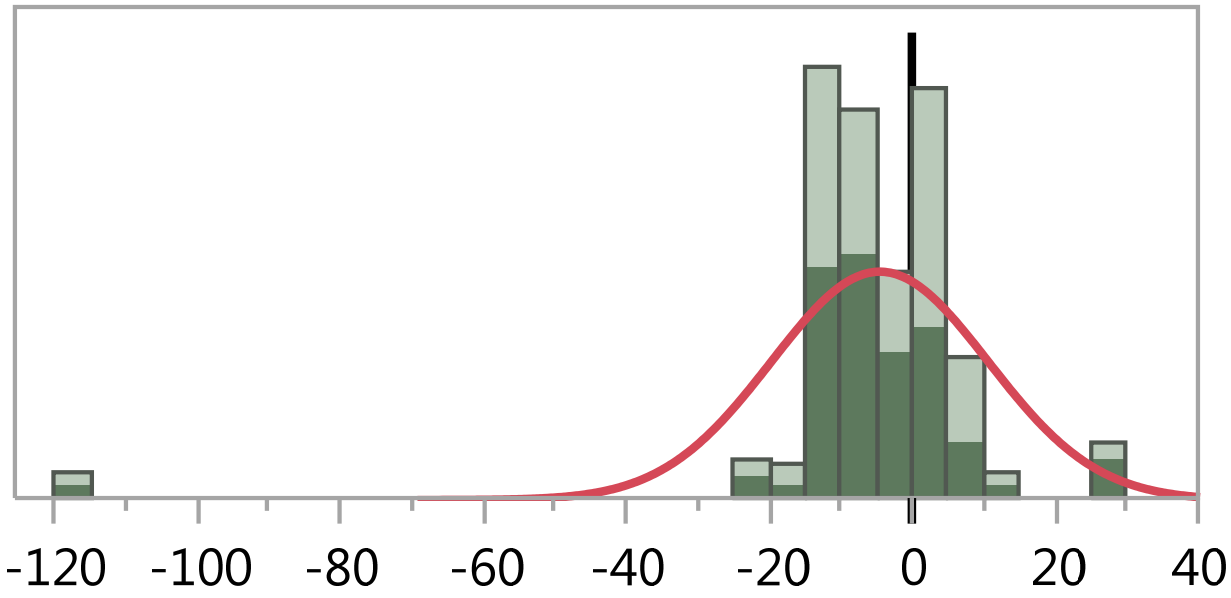
Since expression (3.4) uses the observed, local difference in *average values* of ***Y***-outcomes between treatment cohorts, it follows from first principles that (3.4) yields an asymptotically unbiased estimate even when the local sample sizes for ***T***=1 and ***T***=0 units differ (local imbalance on treatment fraction.) See the Appendix for technical details on the relevant asymptotics.

A key feature of LC analysis strategy is to view the collection of local estimates of form (3.4) across clusters as constituting a *statistical distribution* of local effect-size estimates. Since the clusters (subgroups of locations) being formed are mutually exclusive, researchers can typically view LTD estimates as being statistically independent. On the other hand, LTD estimates almost always differ in precision (have *heteroschedastic dispersion*) because cluster sizes vary and within-cluster fractions of nonattainment counties also vary. Thus, in final-phase LC modeling efforts to predict LTDs, choice of differential weighting for individual LTD estimates is usually essential.

To greatly reduce the total length of our tutorial, we will now restrict attention to LC analysis of changes in Elderly Mortality (ages 65 through 84.) While our corresponding LC reanalysis of Adult Mortality (population aged 50 and over) gave highly similar results on all of the important issues discussed below, some tangential issues (such as treatment of unpredictable outliers in modeling) tended to somewhat complicate Adult Mortality reanalyzes and their discussion.

Note that the distribution of observed LTD estimates displayed in Figure 3.1 clearly overlaps zero. The most negative LTD estimate (118.572 fewer deaths per 10K elderly population) comes from a cluster of size 6 that includes 2 nonattainment and 4 attainment counties. At the other extreme, a cluster of 16 counties that includes 10 nonattainment and 6 attainment counties yields an LTD estimate of +28.622 elderly deaths. Since negative and positive LTD estimates have starkly different interpretations, more information about the breakdown of CAAA compliance nonattainment and attainment counties into clusters with positive or negative LTD estimates is provided in Table 3.1.

Figure 3.1. Distribution of Elderly Mortality Local Treatment Difference (LTD) Estimates



Normal Fit (Mean = -4.81 Deaths per 10K Population, Std Dev = 15.17)
CAAA 1970 Compliance Classification in 1971:
Histogram Shading: Attainment Light, Nonattainment Dark

Table 3.1. Classification of 533 US counties on CAAA compliance and the numerical sign of LTD estimates.

Sign of Elderly LTD Estimate	Number of Informative Clusters	Nonattainment Counties	Attainment Counties
Positive	9	**79**	111
Negative	15	197	**146**
Total	24	276	257

Note in Table 3.1 that 79 nonattainment counties were classified into 9 clusters with positive LTD estimates from the LC nonparametric preprocessing analysis described in §3.2. Apparently, increased TSP regulation was detrimental to elderly mortality in US counties with the REIS characteristics of these 9 clusters. Specifically, relative to the 111 attainment counties classified into the same 9 clusters with positive LTD estimates, these 79 nonattainment counties had higher (more positive and undesirable) changes in elderly mortality than the 111 attainment locations they were clustered with.

On the other hand, 146 attainment locations were classified into 15 clusters with negative LTD estimates, suggesting that increased TSP regulation could have been beneficial to elderly mortality in these 146 attainment locations. Specifically, relative to the 197 nonattainment counties classified into the same 15 clusters with negative LTD estimates, these 146 attainment counties had higher (more undesirable) changes in elderly mortality than the 197 nonattainment locations they were clustered with.

In section §4, we will see that both the most negative and the most positive LTD estimates displayed in Figure 3.1 are well predicted by the ***X***-confounder characteristics of their clusters.

We recommend that the precision of LTD estimates be determined using the observed mean square for error from fitting a nested ANOVA model with effects for informative clusters and for treatment within those clusters. Technically, the 23 degrees-of-freedom for the "blocking" effects of 24 informative clusters in our example provide no information relative to treatment effects (CAAA compliance) unless all ***X***-confounders used to form clusters are *instrumental variables* (IVs)[30], i.e. variables that have no direct effects on treatment ***Y***-outcomes even though they may influence treatment assignment. Because IV assumptions are very strong and untestable, LC strategy does not use any of this information when estimating LTDs. The simplifying assumption that LC strategy typically does make is that ***Y***-outcomes have *homoschedastic dispersion,* measured by the nested ANOVA mean square for error. In our example, this mean square is 881.58 with 485 degrees-of-freedom, which corresponds to a standard deviation of 29.69 deaths on change in elderly mortality. Again, LTD estimates are *heteroschedastic* due to variation in both cluster sizes and CAAA compliance fractions within clusters in equation (3.4).

3.4 Confirmation and Systematic Sensitivity Analysis of LTD Distributions

The initial, nonparametric preprocessing phase of LC strategy ideally includes two additional checks on the form and interpretation of observed LTD distributions: [i] confirmation that the ***X***-confounder characteristics used to cluster experimental units "truly matter," and [ii] systematic sensitivity analyses to reveal how numerically stable the LTD distribution is under *changes* in main LC parameter settings, such as number of clusters formed, choice of ***X***-space distance metric, and choice of clustering method.

These two additional checks can be particularly insightful, especially when observational datasets are quite large. Because the data available for the CAAA natural experiment are quite limited and contain

so few relevant ***X***-confounders, the additional checks have limited impact in our current example.

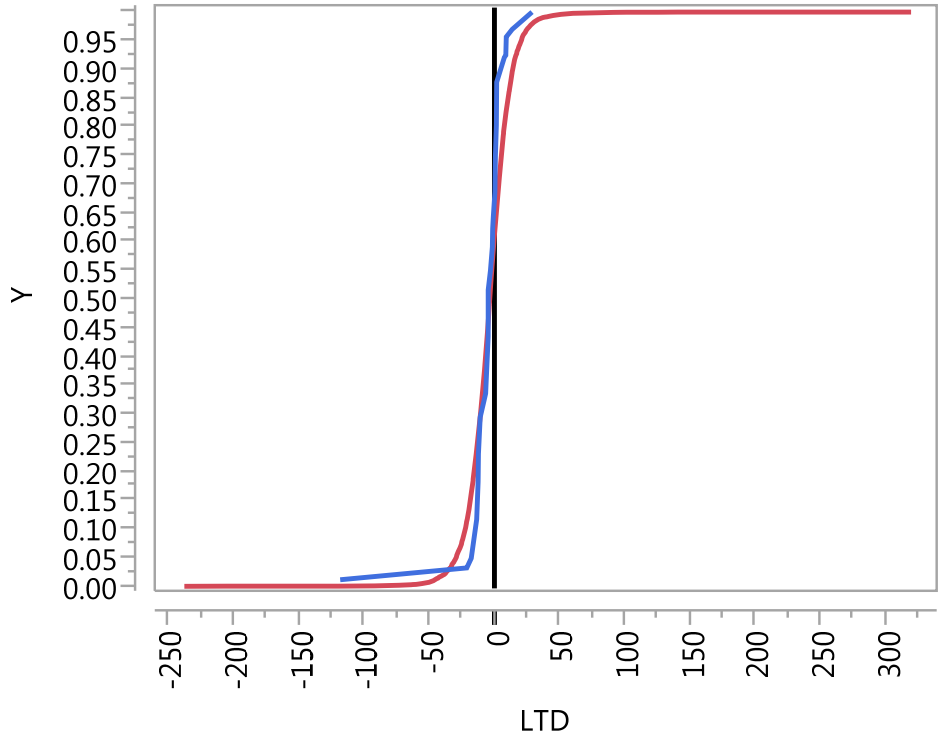
If the four primary REIS ***X***-confounders used in our CAAA example actually are not relevant to longevity outcomes, the 24 informative clusters formed using them are essentially being formed randomly. Thus, to confirm that ***X***-matching really does matter, it is essential to verify that the observed LTD distribution formed using these ***X***-confounders has a different location, spread or shape than the corresponding permutation LTD distribution formed by assigning the 533 pairs of observed changes in elderly mortality and compliance level *purely at random* to 25 clusters of the same sizes as the observed clusters. A simple way to do this is to randomly permute[31,32] the observed cluster labels on the 533 counties and compute the 24 implied LTD-like measures that result.

In fact, this permutation LTD distribution can be computed to *arbitrary precision* by accumulating multiple, independent replications of random cluster label permutations. For example, the very smooth Cumulative Distribution Function (CDF) displayed in Figure 3.2 results from 250 permutation replications and has a (essentially true) mean value of –4.211 deaths per 10K elderly population and a (essentially true) standard deviation of 16.451 deaths.

The short-tailed CDF depicted in Figure 3.2 corresponds to the observed LTD distribution for change in elderly mortality from all 24 informative clusters of counties relatively well-matched on their primary REIS characteristics. The standard deviation of this observed LTD distribution of 15.168 elderly deaths per 10K population is significantly less than that of its random permutation counterpart: the chi square statistic for testing that the true standard deviation is 16.451 deaths is 452.3 with a two-tailed p-value of $0.0106 < 0.05$. The sample mean value of the observed LTD distribution of –4.812 elderly deaths is not significantly different from its random permutation counterpart of –4.211 elderly deaths.

In summary, the full observed LTD distribution from all 24 informative clusters has significantly lower dispersion (less spread) than its corresponding LTD-like permutation distribution.

Figure 3.2 Cumulative Distribution Functions comparing the observed LTD distribution for Change in Elderly Mortality with its Random Permutation Counterpart



CDF for Permutation Distributions (250 reps) and **CDF for Observed LTD Distribution**

In one of the many sensitivity analyses[33] we tried, we found that using 17 informative "Complete Linkage" clusters yielded essentially the same sorts of LC findings and interpretations as 24 Ward clusters. We also found that we could not consider using larger numbers of clusters (than either 17 Complete or 24 Ward) without relegating many more than 27 of 560 counties to uninformative clusters. A smaller total number of clusters could be used, of course, but much interesting detail in the observed LTD distributions is then quickly lost. In other words, the historical data available for the CAAA natural experiment[22] appear to be sufficient to explore only a quite limited range of most relevant LC analyses.

4. LC Final Phase: Detecting Heterogeneity by Successfully Predicting LTDs

The nonparametric, nested ANOVA model used in initial-phase LC to estimate local treatment effect-sizes is not the sort of model one can also use for traditional "covariate adjustment," in which variation in local effect-sizes is to be predicted from observed variation in ***X***-covariates. Instead, each initial LTD estimate is *conditional* upon the highly similar ***X***-covariate vectors of the experimental units (counties) grouped together to form each local estimate. This makes a new variable, consisting of a collection of

different initial LTD estimates, an *ideal* ***left-hand-side*** *variable* for traditional statistical modeling and prediction of how local treatment effects vary when ***X***-covariate vectors also vary.

There are several good reasons to anticipate that across-cluster variation in initial LTD effect-sizes will be relatively easy to predict; this will be a primary topic of our main discussion in §5.

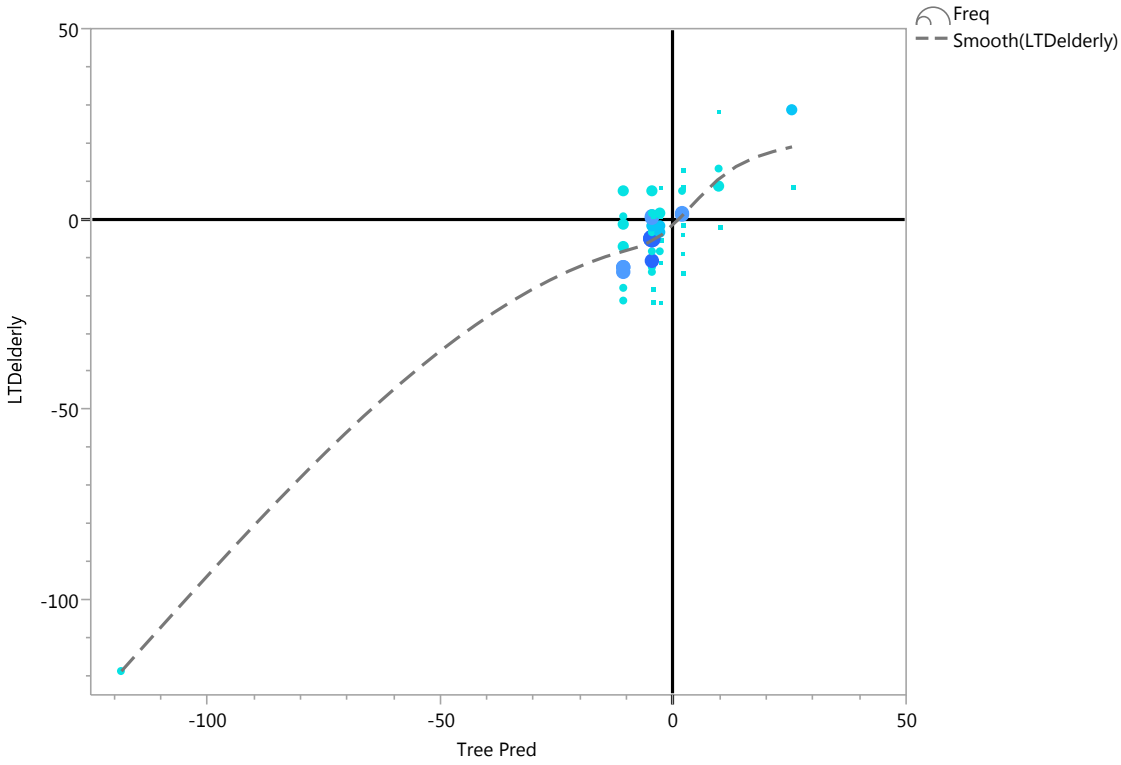
To get started with final-phase LC model fitting, a decision on how to weight individual LTD estimates must be made. Using weights inversely proportional to the sample variances of the left-hand-side variable is frequently recommended. For example, when the j^{th} quantity to be predicted is the mean of N_j independent and identically distributed observations, the "frequency" weights of $(N_1, N_2, N_3, ...)$ have this property because the values being averaged are homoscedastic. Under this assumption, the corresponding weights suggested by equation (3.4) would be proportional to $(N_{j1}+N_{j0}) \times [p_j(1-p_j)]$, where the second subscript on N denotes treatment cohort (1 or 0), and $p_j = N_{j1}/(N_{j1}+N_{j0})$ is the CAAA compliance nonattainment proportion within the j^{th} cluster. Note that the final weighting factor varies only from 0 to 0.25; it down-weights the LTDs from clusters where p_j differs most from 0.5.

Unfortunately, common assumptions like independence and constant true mean values across years seem unlikely in our CAAA compliance application. Thus we recommend use of frequency weights defined using only the initial "total cluster size" factor, $(N_{j1}+N_{j0})$. In fact, we frequently recommend using this particular choice of weighting because it essentially weights the data from each observed experimental unit equally. In other words, the observed LTD estimate from a cluster can then be used as the adjusted ***Y***-outcome for every location within that cluster. This convention also allows researchers to use the given ***X***-vector for each experimental unit (rather than the ***X***-vector of the cluster centroid) when fitting "unweighted" final-phase LTD prediction models.

4.1. Many Details of and Interpretations for Fits are relatively Unimportant

Details concerning the exact form of the predictive models used in final phase LC are much less important than the goodness-of-fit of the models to observed LTD estimates. After all, goodness-of-fit information is quite easily *communicated visually* by simply plotting observed LTD estimates versus their model predictions. Note that Figure 4.1 uses both the shading and the area of the plotting symbol to indicate the relative sizes of the 24 informative clusters identified in initial-phase LC; larger clusters generally provide more precise LTD estimates for change in elderly mortality.

Figure 4.1 Observed and Predicted LTDs for Change in Elderly Mortality



Correlation = +0.929, R^2 = 0.862

To help the viewer determine the degree of linear association between the observed LTDs and their predictions, a smoothing-spline is displayed as a dashed-line. To keep the display simple and relatively clean, a normal-theory correlation ellipsoid was not overlaid upon Figure 4.1. However, the correlation here of 0.929 is quite strong and positive; the adjusted R-squared (goodness-of-fit) measure of 0.862 is the square of this correlation. The data point in the lower left of Figure 4.1 is rather distinctly different from the other points. Still, if it is removed, the resulting correlation is 0.782 and the R-squared is 0.611.

A distinct possibility is that the ***X***-factors found to predict heterogeneity[(34)] in one's final phase LC model are mere surrogate measures of some unobserved and unknown vector of true, root causes, ***Z***. The fact that ***X***-variables quantified within the available observational data have been found to be highly predictive of effect-size variation necessarily implies only this: *Some true **Z** (causal agent) does exist*. Such a finding does not necessarily imply that the available ***X***-confounders are, themselves, true ***Z***-factors.

If observational data on true ***Z***-factors were available, application of LC analysis strategy could produce final phase fits that turn out to be either more predictive or else less predictive of effect-size heterogeneity. Hopefully, the new fits would at least be as much or more easily interpretable as causal models than the current ***X***-based models.

A rather simple partition regression "tree" model[35] provided the LTD predictions for change in elderly mortality displayed in Figure 4.1. The tree used a total of only 6 splits using only 2 REIS confounders: Employment Fraction in Manufacturing (MANFR) and Unemployment Insurance Payments (UI). Specifically, the overall TSP geometric mean measure was not selected for use in any of these splits.

Because this small tree has only 7 final leaf nodes, it provides only 7 distinct, numerical predictions for observed LTD estimates from 24 informative clusters and, yet, has an adjusted R-squared of 0.862 for its goodness-of-fit. Although they lack the interpretability of a single tree, predictions from a "forest" of trees[35] by model-averaging are known to be more stable. Here, predictions from a forest of 100 trees were slightly less predictive: correlation with observed LTDs = 0.881, R-squared = 0.777.

Naturally, we also tried using multivariable regression models in final-phase LC prediction, with somewhat less success than with tree models. We explored potential fits using factorial and polynomial models of degree at most two in 5 ***X***-confounders: 6-year average levels for the 4 primary REIS variables and the TSP geometric mean. The best fitting regression for predicting LTDs in Elderly mortality change used 10 degrees of freedom and had an adjusted R-square of 0.669 and overall p-value < 0.0001.

Information on 6-year Average TSP levels was *not selected* for inclusion in our "best fitting" tree or multivariable regression models for prediction of observed LTDs. All predicted heterogeneity in LTDs appears due to the four primary REIS ***X***-confounders, i.e. appears due to socio-economic and environmental factors other than air quality. As one might expect, the simple LC analyses described here generally agree with the relatively sophisticated econometric modeling of the original authors[3] and, perhaps, modestly clarify the interpretation of their findings.

4.2. Detecting Numerically Large Effects is very Important

Especially when observational datasets are quite large, model fits with relatively low R-squares (less than 0.1 or 10%) can still be highly significant statistically (say, p-values much less than 0.01.) That is

why our discussions here have focused, instead, on treatment effect-size distributions and on model goodness-of-fit measures for predicting them. Again, the statistical significance of individual effect estimates or parameters within model fits are relatively unimportant in determining the validity of final phase LC modeling in establishing treatment effect-size heterogeneity. In other words, final phase LC model fits need to be much more than merely statistically significant; they also need to imply that the *predicted heterogeneity is large enough numerically to be a practically important consideration in treatment policy making*.

5. Discussion

When analyzing observational data, applications of traditional model specification and selection methods are complicated by a very practical dilemma. Parsimonious models are usually too simple to provide new, detailed insights about the full range of possible effects of treatment on ***Y***-outcomes. But exactly which sorts of more complex models should be considered? Researchers can make very strong (but possibly wrong) assumptions[21] or simply use trial-and-error to fit models suggested in published literature. Results generated in any arbitrary or unspecified way can be highly dependent upon researcher incentives[15,16].

By using nonparametric preprocessing in the initial phase of LC to provide LTD estimates that serve as left-hand-side variables, LC strategy *greatly simplifies* the statistical modeling needed in final phase LC to determine whether effect-sizes are truly heterogeneous. After all, by focusing on only estimates of the FTC estimands of (3.1) as ***X***-confounders vary, LC strategy essentially *takes all relevant treatment effects to the left-hand-side of final-phase modeling equations*. This justifies and strongly encourages reliance on right-hand-side models that are parsimonious in ***X***-confounders; more complicated final-phase models simply are not needed in either theory or practice. As long as a final-phase LC model provides satisfactory predictions, the simpler that model, the better.

By initial clustering of experimental units on their ***X***-confounders, each within-cluster LTD estimate initially classifies all observed differences in ***Y***-outcome as conditional effects of treatment at its given, local ***X***-centroid. However, within final-phase predictive modeling, high goodness-of-fit focuses attention upon the role of ***X***-confounder variation in moderating effect-size heterogeneity.

In other words, whenever the goodness-of-fit of final-phase LC models is relatively poor (e.g. low R-squares), the observed LTD estimates represent effects primarily due to treatment, and any observed

variation in effect-sizes is primarily considered unexplained (random) variation. On the other hand, when the goodness-of-fit of final models is relatively good (e.g. R-squares above a relatively high threshold, say above 0.4 or 40%), then observed LTD estimates appear to be fixed effects predictable from variation in given ***X***-confounders. Thus, we think that an appropriate summary statement for our CAAA findings would be: *Sound evidence that local treatment effects are truly heterogeneous*[34] *has been accumulated and is not related to TSP level.*

WARNING: We have referred to R-squared measures here primarily qualitatively rather than quantitatively. We are definitely not claiming that R-squared statistics or any other specific measure of goodness-of-fit in final-phase LC modeling validly determines the quantitative percentage of effect-size variation that can be considered to be heterogeneous rather than random.

Our findings in Section §4.3 also illustrate the prime importance of a wise *design choice* for LC analysis strategy. Specifically, by focusing upon *mortality changes* between three-year periods (post=1972-74 minus pre=1969-71), we were able to show that average TSP levels within counties are not useful in predicting the resulting LTDs. The primary alternative formulation for the CAAA natural experiment would have been to analyze ***Y***-outcomes from the three-year post-period using data from the three-year pre-period as ***X***-confounders. We knew that this alternative formulation was doomed to failure because good ***X***-matches would have then been both rare and potentially meaningless. Specifically, the 286 nonattainment counties would then have generally had much higher TSP levels than the corresponding 276 CAAA compliance attainment counties; this would have not only greatly reduced "common support"[17] in ***X***-confounder distributions between treatment cohorts but also, again, would have violated the requirement that treatment assignment not influence LC choice of clustering (matching) of locations.

A shortcoming of LC strategy confirmed in unpublished simulation studies is that basic clustering concepts tend to be inherently low-dimensional. When ***X***-confounder space is high dimensional and noisy, a large proportion of the available experimental units can then be relatively distant from all others. In other words, the LTD of (3.4) from such a cluster would not be a credible estimate of the FTC estimand of (3.1) at the cluster ***X***-centroid. Our recommendation is that screening of ***X***-confounders be used to limit their multicolinearity and to reduce their dimensionality to at most the 5-to-10 confounder range. If excluding a given ***X***-confounder would (i) materially reduce the separation between its observed LTD and its random permutation distribution or (ii) make the observed LTD

distribution appear to be much more numerically stable, then that exclusion would be counterproductive.

The intention of LC strategy is to reveal rather than conceal alternative analyses; although each ***X***-confounder subset explored includes at most 5-to-10 factors, several different such subsets that are only partially overlapping can be explored within a single LC analysis.

Finally, because LC strategy has the potential to be applied quite algorithmically, the credibility of LC analyses would almost surely be maximized by encapsulating the LC strategy outlined here within an "expert" software system[36] for analysis of large, observational datasets. The objectivity of the output from such a system would be unquestionable, and the reproducibility of its output would also be assured. Furthermore, highly convincing visual evidence could be generated as interactive video clips displaying variation in either the observed LTD distributions of Figure 3.1 or else the observed versus predicted LTD plots of Figure 4.1. For example, observed LTD treatment effect-size distributions could first be sorted by their respective (i) mean, (ii) variance, or (iii) skewness measures, and then be repeatedly displayed in any one of these three orders. The stability and uncertainty within these distributions from alternative, credible LC analyses would then become visually obvious. Generation of consensus views embraceable by diverse stakeholders and policy makers would be facilitated.

Acknowledgements, Data Sharing and Software Use

We are indebted to Russ Wolfinger of SAS for creating a highly efficient JMP Add-In[13] for Local Control. This implementation greatly extends our macros[9] and scripts[10] and provides a menu interface for creating LC visualizations like Figure 3.1 and, especially, the smoothed LTD permutations distribution of Figure 3.2. We created Figure 4.1 using the JMP "graph builder" interface and extensively used JMP "bubble plots" in our preliminary analyses of longitudinal variation in CAAA yearly summary statistics. Furthermore, we fit both tree and forest predictions of LTDs using the "Partition" option on the JMP main menu: Analyze > Modeling.

Again, we thank Carlos Dobkin for providing historical air quality, longevity and confounder data[3] quantifying possible effects of CAAA 1970. Besides this staged dataset in CSV format, freely downloadable materials[14] include CSV files containing all data and documentation needed to recreate our analyses of change in either Elderly or Adult Mortality between the 1969-71 and 1972-74 periods.

TECHNICAL APPENDIX:

In this appendix, we outline a pair of informal, heuristic arguments that use simplified notation[29] where measure theoretic details, such as distinctions between continuous and discrete variables, are minimized.

Part I: Cluster Membership is an asymptotic "Balancing Score."

Let $\boldsymbol{x}$ denote a vector of confounder values, let $\boldsymbol{t}$ denote a binary valued (0 or 1) treatment assignment indicator, and let $\boldsymbol{C}$ denote a cluster of confounder vectors that includes the given $\boldsymbol{x}$ vector. Then, with Pr(|) denoting conditional probability, we write

$$\Pr(\boldsymbol{x}, \boldsymbol{t} \mid \boldsymbol{C}) \equiv \Pr(\boldsymbol{x} \mid \boldsymbol{C}) \Pr(\boldsymbol{t} \mid \boldsymbol{x}, \boldsymbol{C}) \qquad \text{(A.1)}$$

$$= \Pr(\boldsymbol{x} \mid \boldsymbol{C}) \Pr(\boldsymbol{t} \mid \boldsymbol{x}) \quad \text{because } \boldsymbol{x} \text{ is within } \boldsymbol{C} \text{ and } \boldsymbol{C} \text{ does not depend upon } \boldsymbol{t} \qquad \text{(A.2)}$$

$$\rightarrow \Pr(\boldsymbol{x} \mid \boldsymbol{C}) \Pr(\boldsymbol{t} \mid \boldsymbol{C}) \quad \text{as } \boldsymbol{C} \text{ shrinks to } \boldsymbol{x}. \qquad \text{(A.3)}$$

Note that:

- Relationship (A.1) follows from the basic definition of conditional probability.
- Whenever cluster formation depends only upon the available $\boldsymbol{x}$-vectors and *thus does not depend in any way upon treatment assignment*, $\boldsymbol{t}$, the right-hand side of expression (A.1) can then be rewritten as (A.2).
- In the limit as the $\boldsymbol{x}$ "diameter" of cluster $\boldsymbol{C}$ shrinks to zero, the given $\boldsymbol{x}$ becomes the only interior point of $\boldsymbol{C}$, and expression (A.3) holds asymptotically.

The main implication of (A.3) is that the conditional distributions of the $\boldsymbol{x}$-vector and the $\boldsymbol{t}$-choice are *asymptotically independent* within each given cluster ...making cluster membership an asymptotic "balancing score"[29].

Part II: Asymptotically, Cluster Membership is either "equivalent to" or else actually "finer than" the unknown, true Propensity Score.

True propensity scores are the most "coarse" of all possible balancing scores[29]. Since cluster membership is an asymptotic balancing score, it follows that cluster membership is either equivalent to the unknown, true propensity score or else is more "fine" than true propensity. In fact, the given $\boldsymbol{x}$-confounder vectors of individual experimental units are known to be the "most fine" possible balancing scores[29].

The true propensity score, $\mathbf{p} \equiv \Pr(\boldsymbol{t} = 1 \mid \boldsymbol{x})$, is typically *unknown* and, thus, needs to be *estimated* (say, via a logit or probit model) in all practical applications of propensity scoring. Unfortunately, propensity *estimates* can fail to have the highly desirable properties of true propensity scores! Thus, besides being asymptotically equivalent to or finer than true propensity, cluster membership has the added practical advantage of being a known (observable) characteristic of the $\boldsymbol{x}$-confounders of experimental units.

References

1. Katsouyanni K, Samet J, Anderson HR, et al. Air Pollution and Health: A European and North American Approach (APHENA). HEI Research Report 142. Health Effects Institute, Boston, MA. 2009.

2. Pope III CA, Ezzati E, Dockery DW. Fine particulate air pollution and life expectancy in the United States, *N Engl J Med* 2009; 360:376–386.

3. Chay K, Dobkin C, Greenstone M. The Clean Air Act of 1970 and adult mortality. *J Risk Uncertainty* 2003; 27:279-300.

4. Enstrom JE. Fine particulate air pollution and total mortality among elderly Californians, 1973-2002. *Inhal Toxicol* 2005;17:803-816.

5. Janes H, Dominici F, Zeger S. Trends in air pollution and mortality: an approach to the assessment of unmeasured confounding. *Epidemiol* 2007;18:416-423.

6. Greven S, Dominici F, Zeger S. An approach to the estimation of chronic air pollution effects using spatio-temporal information. *J Amer Stat Assoc* 2011;106:396-406.

7. Cox LA Jr, Popken DA, Ricci PF. Warmer is healthier: effects on mortality rates of changes in average fine particulate matter (PM2.5) concentrations and temperatures in 100 U.S. cities. *Regul Toxicol Pharmacol* 2013;66:336-346.

8. Young SS, Xia JQ. Assessing geographic heterogeneity and variable importance in an air pollution data set. *Stat Anal Data Mining* 2013; 6:375-386.

9. Obenchain RL. SAS macros for local control (initial-phase). Observational Medical Outcomes Partnership (OMOP), Foundation for the National Institutes of Health (Apache 2.0 License). 2009.
10. Obenchain RL. The local control approach using JMP. In *Analysis of observational health care data using SAS*, ed. D. E. Faries, A. C. Leon, J. M. Haro, and R. L. Obenchain, pp. 151–192. Cary, NC: SAS Press, 2010.

11. Obenchain RL, Young SS. Advancing statistical thinking in observational health care research. *J Stat Theory Pract* 2013; 7:456–469.

12. Lopiano KK, Obenchain RL, Young SS. Fair treatment comparisons in observational research. *Stat Anal Data Mining* 2014; 7:376-384.

13. Wolfinger RD. JMP Add-In for Local Control calculations. Cary, NC: SAS Institute Inc. 2015 https://community.jmp.com/docs/DOC-7453

14. Obenchain RL. *Local Control Analysis Strategy*: Website. 2015 http://localcontrolstatistics.org/

15. Leamer EE. *Specification Searches: Ad Hoc Inference with Nonexperimental Data.* New York: Wiley, 1978.

16. Glaeser EL. Researcher incentives and empirical methods. Harvard Institute of Economic Research, Discussion paper 2122. 2006.

http://scholar.harvard.edu/files/glaeser/files/researcher_incentives_and_empirical_methods.pdf

17. Stuart EA. Matching methods for causal inference: a review and a look forward. *Stat Sc* 2010; 25:1–21.

18. Clyde M. Model uncertainty and health effect studies for particulate matter. *Environmetrics* 2000; 11:745–763.

19. Young SS, Karr A. Deming, data and observational studies. *Significance* 2011; 8:116–120.

20. Gelman A, Loken E. The statistical crisis in science. *Am Sci* 2014; 102:460-465.

21. van der Laan MJ, Rose R. Statistics ready for a revolution: next generation of statisticians must build tools for massive data sets. *AMStat News* 2010; 399:38–39.

22. Craig P, Cooper C, Gunnell D, et al. Using natural experiments to evaluate population health interventions: new Medical Research Council guidance. *J Epidemiol Community Health.* 2012; 66: 1182-1186.

23. Ho D, Imai K, King G, Stuart EA. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Anal* 2007; 15: 199-236.

24. Iacus SM, Gary King G, Porro G. Causal Inference without Balance Checking: Coarsened Exact Matching. *Political Anal* 2012; 20: 1-24.

25. Guha S, Hafen R, Rounds J, et al. Large complex data: divide and recombine (D&R) with RHIPE. *Stat* 2012; 1: 53–67.

26. Domingo-Ferrer J, Mateo-Sanz JM. Practical data-oriented microaggregation for statistical disclosure control. *IEEE Trans Knowledge Data Eng* 2002; 14: 189-201.

27. Hastie T, Tibshirani R, Friedman J. Unsupervised Learning. *The Elements of Statistical Learning: Data Mining, Inference and Prediction* (2nd ed.) New York: Springer, 2013; 485-586.

28. Rubin DB. For objective causal inference, design trumps analysis. *Ann Appl Stat* 2008; 2: 808-840.

29. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika* 1983; 70: 41-55.

30. McClellan M, McNeil BJ, Newhouse JP. Does more intensive treatment of acute myocardial infarction in the elderly reduce mortality? analysis using instrumental variables. *JAMA* 1994; 272: 859-866.

31. Dwass M. Modified randomization tests for nonparametric hypotheses. *Ann Math Stat* 1957; 28: 181–187.

32. Welch WJ. Construction of Permutation Tests, *J Amer Stat Assoc* 1990; 85: 693-698.

33. Greenland S. Basic methods for sensitivity analysis of biases. *Int J Epidemiol* 1996; 25: 1107-1116.

34. Davidoff F. Heterogeneity is not always noise: lessons from improvement. *JAMA* 2009; 302(23): 2580-2586.

35. Hawkins DM. Recursive partitioning. *Comput Stat* 2009; 1: 290-295.

36. Tukey JW. Sunset salvo. *Amer Stat* 1986; 40: 72-76.