Advancing Statistical Thinking in Observational Health Care Research

Bob Obenchain, RBstats

Stan Young, NISS

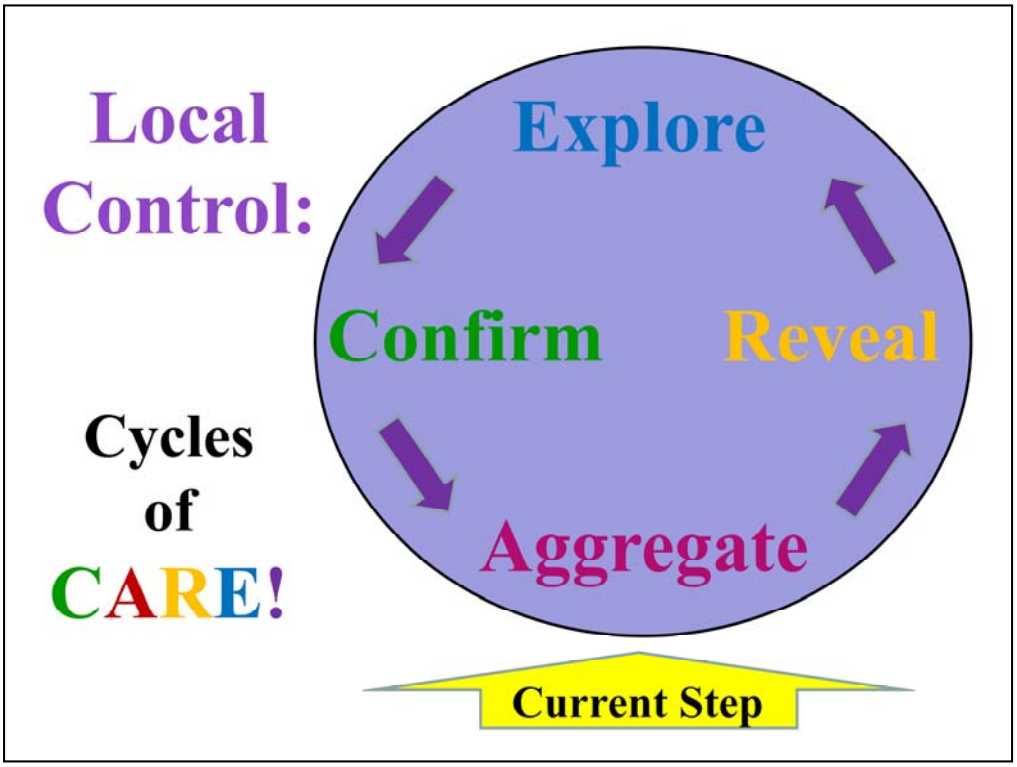
Journal of Statistical Theory and Practice 7, 456-469, April 2013

The Local Control (LC) approach to adjustment for treatment selection bias (patient channeling) and confounding in Observational Comparative Effectiveness Research uses patient matching / clustering methods (unsupervised learning.) The theoretical basis for LC is that cluster membership is guaranteed to become a “blocking / factoring score” that is finer (more detailed) than the unknown true propensity score in the limit as clusters become small, compact and numerous. The LC approach uses systematic sensitivity analyses to explore alternative Local Treatment Difference (LTD) distributions (which are purely objective-but-Bayesian-like analyses), while using resampling methods to confirm that these LTD distributions are clearly different from what results from purely random patient clusterings.

Key "Local" Statistic:

Within-Patient-Subgroup
Local Treatment Difference

$$\text{LTD: } \Delta Y = \bar{Y}_{Treated} - \bar{Y}_{Control}$$



The LC Carousel is usually depicted as rotating **counter-clockwise** as the health outcomes researcher repeats four basic types of analysis steps (or phases), where LC parameter settings **change** with each new Aggregate step …thereby performing SYSTEMATIC SENSITIVITY analyses.

These “Cycles of CARE” assure the OBJECTIVITY and VALIDITY of the full LC PROCESS.

Depending upon which way the LC Carousel rotates (clockwise or counter-clockwise), its four letters can appear to spell out either C-A-R-E or A-C-E-R. This latter sequence of steps: A=1, C=2, E=3 and R=4 is most typically followed …but researchers can, of course, proceed and/or “jump-around” any way they wish.

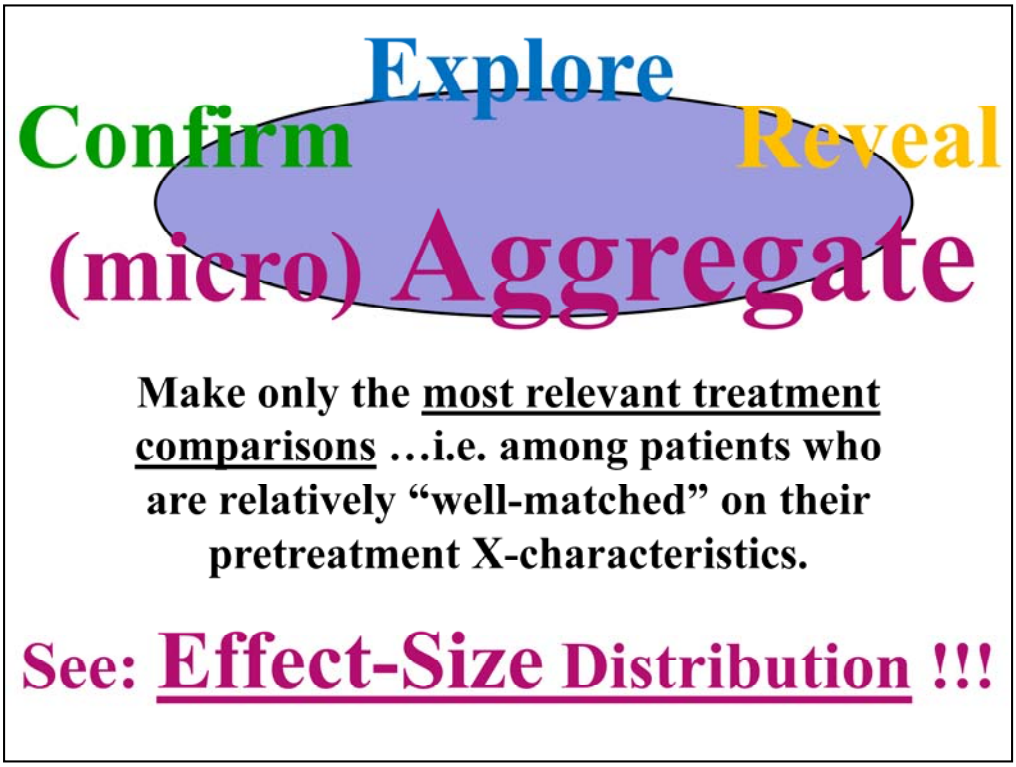
In this presentation, we will start with the A step for (micro-) AGGREGATION.

An Adobe Flash (video) animation for the **LC Carousel** can be viewed by clicking the link on the first **Local Control** page at **localcontrolstatistics.org**

Guests may also download an archive containing all files needed to install the **LC Carousel** animation on their own Microsoft Windows computer using the link at the very bottom of the “LC References and Software Downloads” page.

As the viewer moves his/her cursor left or right across the window, the carousel will speed up, slow down or change direction.

Click on one of the four letters to view a brief description of each phase of LC. See slides #7, #10, #13 and #16 in this presentation.



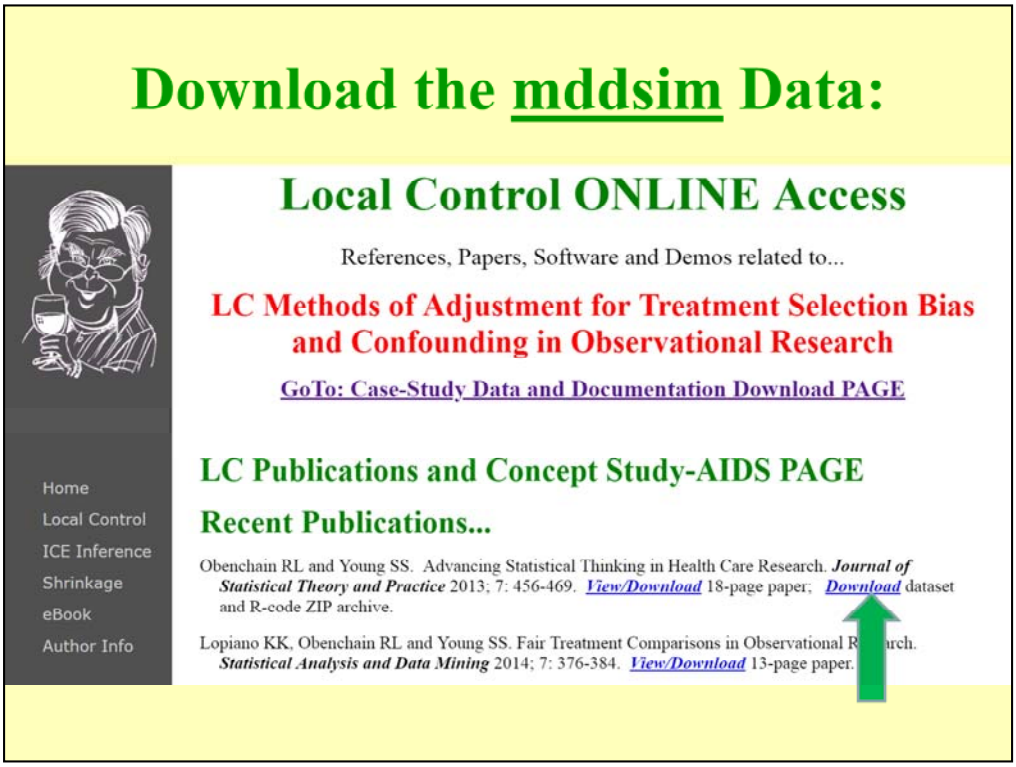
LOCAL CONTROL => Post Hoc BLOCKING of patients.

"micro" Aggregation => considering using more and more (smaller and smaller) patient subgroups.

It's NEVER too late to consider doing a "well designed" and "more relevant" analysis!

Each new cycle of CARE usually starts with considering a new, **alternative Aggregation** of patients.

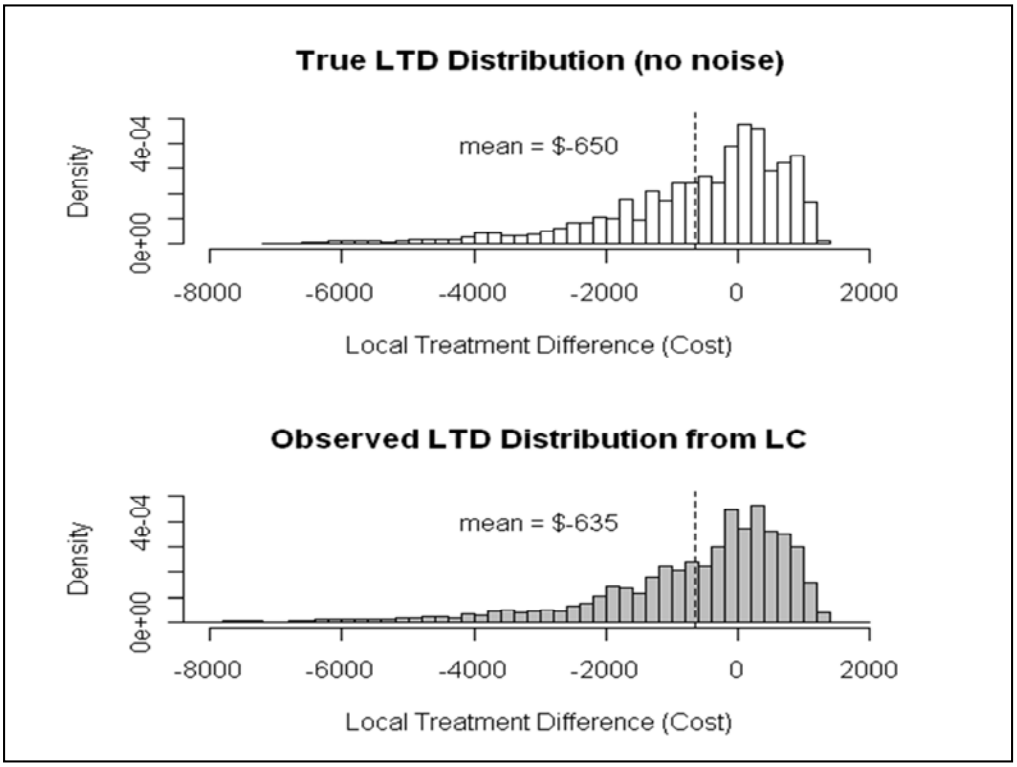
LC Phase One: micro-AGGREGATE. Use patient pretreatment X-characteristics to divide patients into K relatively SMALL and homogeneous subgroups (BLOCKS). A Local Treatment Difference (LTD) of the form [(Mean Y-outcome when Treated) minus (Mean Y-outcome when on Control)] is then computed within each subgroup. Usually, some LTD estimates are positive while others are negative, but all LTD estimates are displayed together as a DISTRIBUTION of Effect-Sizes in a Histogram. While this distribution may look like a Bayesian Posterior, it is actually constructed using only OBJECTIVE information that quantifies patient differential response to treatment and provides sound data-based evidence for Individualized Medicine.



From the **LocalControlStatistics.org** "home" page, click first on the **Local Control** main-menu item, then click on the **LC References and Software Downloads** link to arrive at the page displayed above.

Next, click on the **Download** link pointed to by the Green Arrow above to get your copy of the **mddsim.zip** data archive (1,785 KB). These are the data used to generate the LC graphical displays presented in these Powerpoint slides.

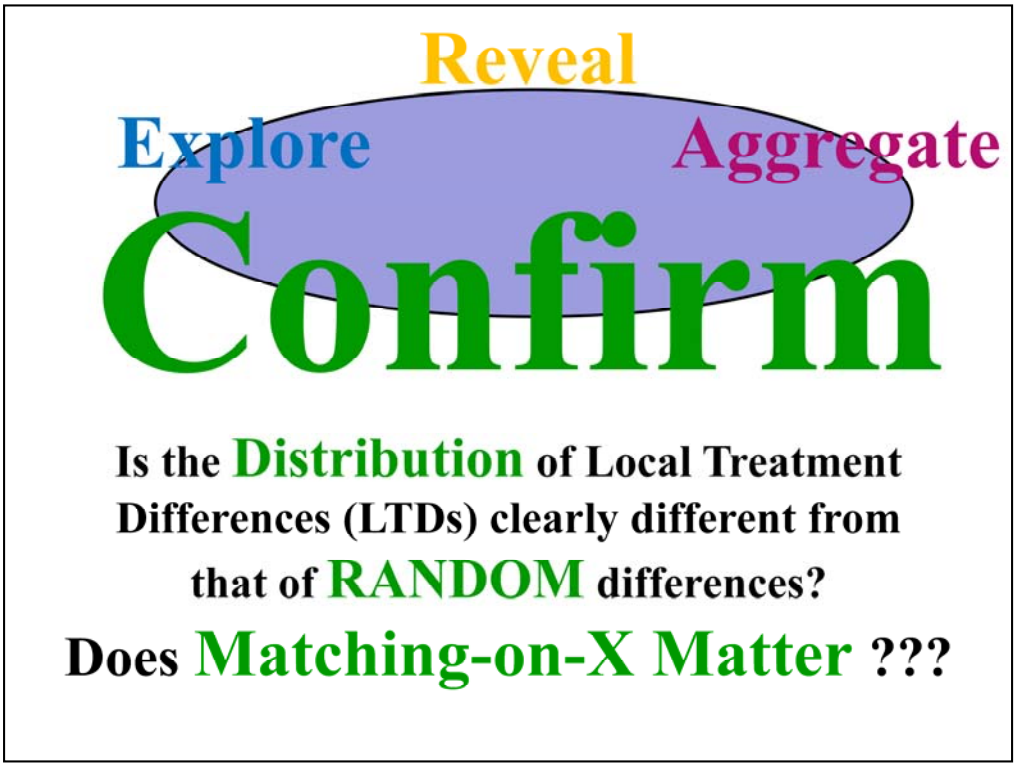
If you download these data, you may analyze them any way you wish. However, it is important to note that, of the 26 variables listed, only the **first 11** represent typical observational data. The next 9 variables in the mddsim26.csv data file represent unknown, unobserved "true" causal relationships and outcomes. Lastly, the final 6 variables were created using **LC analysis strategy**. For example, variable # 21 (Clus2k) is the number (1 to 2,000) of the cluster into which each of the 40,000 patients fell …within the 8-dimensional X-space of variables #4=age through #11=wprevcost.



This comparison typically cannot be made because the information depicted in the upper histogram is usually **unknown**.

The comparison is possible here only because much of the data within the mddsim26.csv file were generated by **simulation**; see the file “mdd40k.r” for the R-code used. To regenerate the provided mddsim data exactly, the correct random-number-generator “seed” value (1249) must be specified.

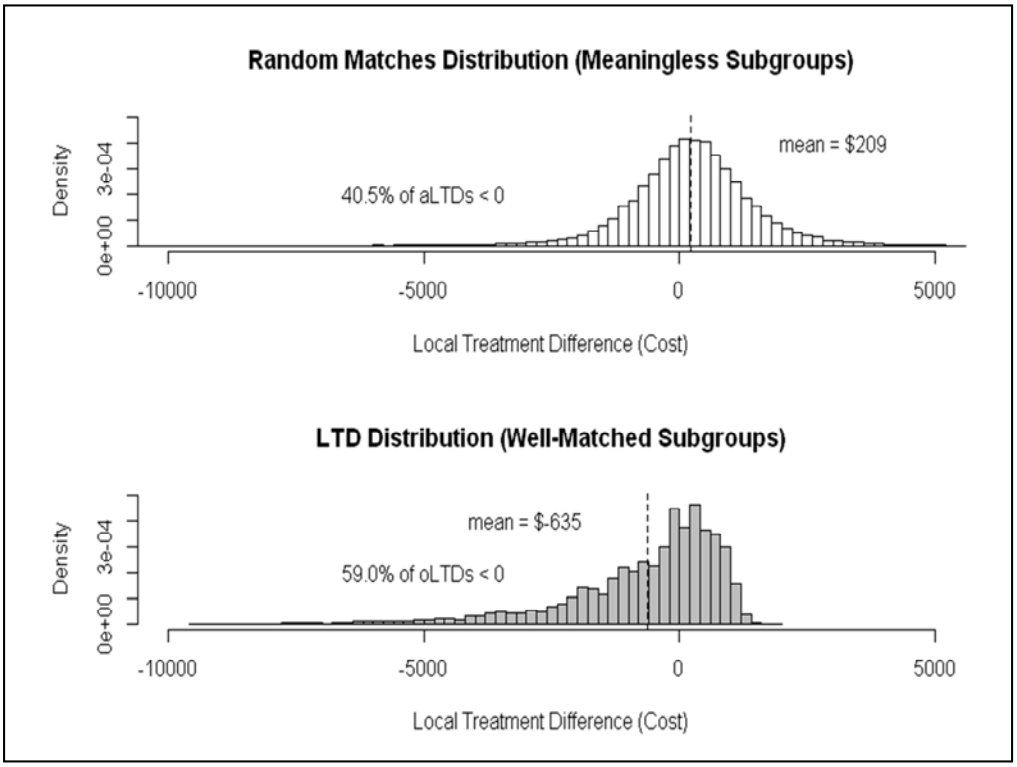
The top histogram then depicts the distribution of the (true) “ltdcost” variable …while the bottom histogram displays the “LTDobs” distribution with many tied values. **This two distributions are clearly quite similar!**


Reveal
Explore
Aggregate
Confirm
Is the Distribution of Local Treatment Differences (LTDs) clearly different from that of RANDOM differences?
Does Matching-on-X Matter ???

C

==2==

LC Phase Two: CONFIRM. An observed LTD Distribution needs to be clearly DIFFERENT from the corresponding distribution generated via RANDOM subgroup formation. Purely RANDOM comparisons using K subgroups of given, fixed sizes N1, N2, ..., NK actually provide NO ADJUSTMENT for Treatment-Selection BIAS or CONFOUNDING. In LC Phase Two, a researcher VISUALLY compares empirical CDFs and BoxPlots for these two types of distributions, looking for OBVIOUS differences. When there are NO clear differences, the observed AGGREGATION of patients well-matched on their pretreatment X-characteristics truly does NOT really MATTER! Phase Two thus yields a GO or NO-GO decision for the current LC iteration.



Considerable information is provided by this numerical example: 40,000 patients spread across 2,000 clusters. This abundance of information allows us to use **simple histograms** to compare the distribution of **observed LTDs** (bottom) to a **NULL distribution** (top) resulting from **purely random assignment** of patients to 2,000 subgroups of the same sizes as the observed clusters. Since this NULL distribution is generated under the assumption that the observed X-confounder variables used to cluster patients are actually **IGNORABLE** (irrelevant), the observed LTD distribution provides strong evidence here to the contrary.

When much less data are available, the observed and random (artificial) distributions are probably best compared (visually) using empirical CDF plots or even simple I-plots (Box and Whisker diagrams.) Use of CDF plots can be augmented with statistical inference based upon random permutations to generate the NULL distribution of Kolmogorov-Smirnov D-statistics. Use of either histograms or density estimates to make such comparisons is confounded with choice of the required cell-wifth or band-width parameter.

Systematic Sensitivity Analyses…

E

==3==

LC Phase Three: EXPLORE. LC helps researchers develop CONSENSUS views using Systematic Sensitivity Analyzes (SSA). Researchers vary three LC Parameter Settings: [1] What number of subgroups, K, optimizes Variance-Bias Trade-Offs? [2] Which X-characteristics determine patient similarity? And [3] What clustering algorithm is used to form subgroups? These choices determine the LOCATION of the LTD Distribution relative to ZERO, its SPREAD, and its SKEWNESS / SHAPE. Phase Three thus determines the full range of alternative LTD distributions supported by the available data AND determines when ITERATION through all four LC Phases should CONTINUE and when it should STOP.

EXPLORE:

Systematic Sensitivity Analyses

[1] How many patient subgroups will be formed?

[2] Which pretreatment X-characteristics will determine similarity of patients?

[3] What patient matching / clustering algorithm?

Cluster membership provides not only [1] what Rosenbaum and Rubin (1983) called a “balancing score” (but is better described as a “blocking-factor”) but also [2] in the limit as clusters become small, compact and numerous, these scores are GUARANTEED to be “finer” (more relevant for inference) than the UNKNOWN, TRUE **Propensity Score.**

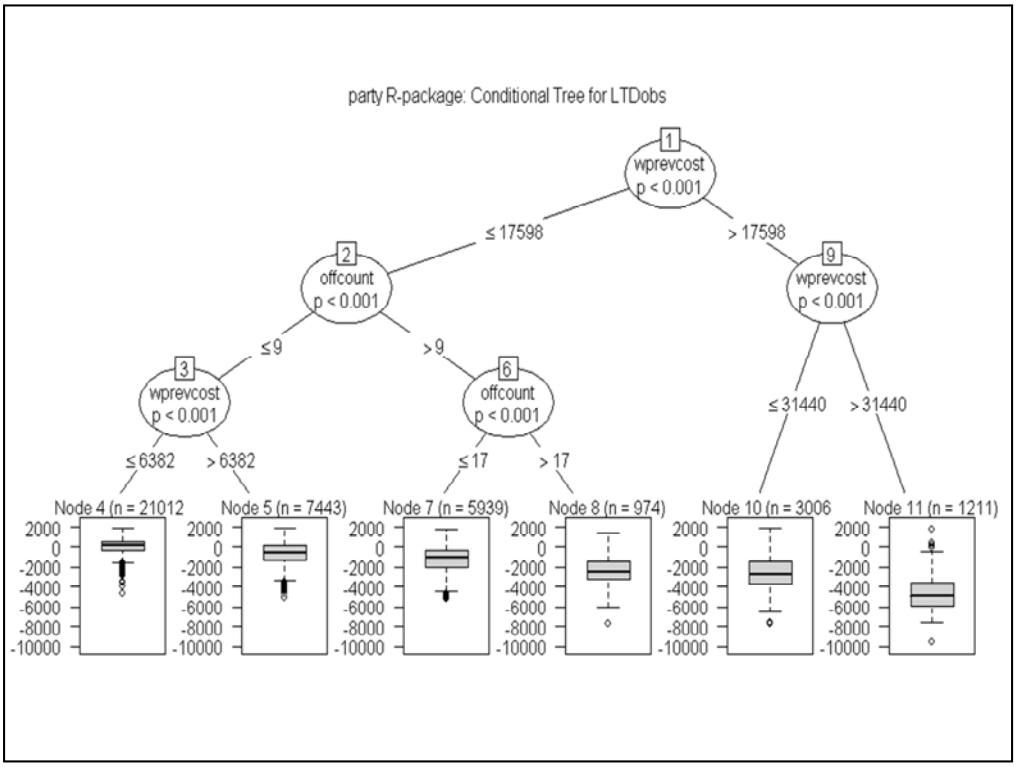
[1] What fraction of patients do better on which treatment?

[2] Are pretreatment X-characteristics predictive of LTD effects?

R

===4

LC Phase Four: REVEAL. LC analyses typically end with an attempt at formal CAUSAL INFERENCE, usually via parametric statistical MODEL fitting. Researchers want to know whether observed LTD variation represents [a] Purely Random deviations around Main-Effects or [b] Predictable Fixed Effects of X-confounders! Do at least some of the observed LTDs represent interpretable SIGNAL? Or, is some One-Size-Fits-All treatment Policy justified? LTD estimates from LC are NONPARAMETRIC and always make inherently FAIR (apples-to-apples) Treatment Comparisons. Because LC REVEAL analyses represent Stretch-Goals, they are attempted LAST because they can be frustrating and inconclusive.



Recursive Partitioning methods generate an easily visualized “Tree” model, but any sort of **Supervised Learning** algorithm can be used to **predict** observed **LTD estimates** from the available X-confounder characteristics of individual experimental units (patients).

ANALYSIS SUMMARY

Our numerical example illustrates that Local Control methods are sufficiently flexible to do a much better job of removing the effects of treatment selection bias and confounding than smooth, global multivariable models.

This example also dramatically illustrates that some of the most interesting effects of treatment are NOT main-effects. Evidence of patient differential response to treatment is always important …even when the only available X-confounders are surrogates for True Causes.

LC Advantages

[i] LC analysis is easy to explain and illustrate (visually) to nontechnical audiences,

[ii] Statistical professionals should find the LC approach to be both intuitively appealing and widely applicable,

[iii] Most LC results are non-parametric and are based upon more realistic assumptions than global, parametric models, …and

[iv] LC yields reasonable inferences even in the limit as "N" increases towards infinity.

We conclude by stressing the (four) main advantages of LC:

Introductory Remarks about our Next (and Final) Slide:

The cycles of analyses described here differ from those of "Exploratory Data Analysis" (EDA) practiced for many years at AT&T Bell Labs and epitomized in John Tukey's 1977 book. Specifically, in the physical and engineering sciences, experimental data typically follow some sort of law-like-relationships. These kinds of data can "suggest" reasonable statistical models …with R-squared (goodness-of-fit) statistics in the 0.75 to 0.95 range.

In human health care studies, parametric models (even rather complicated ones) are rarely anywhere near this good (accurate, definitive.) Although health outcomes models typically identify effects that are declared to be (highly) significant, statistically, they typically have R-squared statistics of less than 0.25 or even 0.10. In other words, these are actually rather weak, overly-simplified "signals" embedded within a whole lot of NOISE!

The 4 phases of each complete cycle described here are easily characterized (described, explained) in general terms. However, they cannot be COMPLETELY PRE-SPECIFIED, as in a RCT. We contend that, by consistently and repeatedly following the analysis cycles described here, a health care researcher can and will become confident that he/she is performing OBJECTIVE and SCIENTIFIC data analyses as long as the available data have 3 properties.

The data [1] must be REAL and REPRESENTATIVE of current medical practice, [2] must be RICH in information about patient X-characteristics, and [3] must provide VOLUMINOUS and DIVERSE information about all patient sub-populations of interest. On the other hand, all patients being analyzed together usually need to have been treated for a single, common disease / diagnosis (possibly with multiple etiologies.)

Repetition of the three A-C-E steps of Local Control assure scientific OBJECTIVITY. The ultimate (FINAL) R step allows researchers to express their full talents in statistics and data science in attempting CAUSAL INFERENCE.