NISS: Statistical Method Development for Observational Comparative Effectiveness Research (OCER)

Alan Karr, Director
Bob Obenchain, Research Fellow
Stan Young, Assistant Director for Bioinformatics

• **Background:** Recent Institute of Medicine (IOM) initiatives are generating increasing interest in Rapid Medical Learning approaches based upon ***patient micro-aggregation*** concepts. Emerging patient registry systems, such as the ASCO ***CancerLinQ*** prototype, have demonstrated the feasibility of "clustering" patients on pretreatment characteristics and displaying historical treatment and corresponding outcome information for any given patient's "nearest neighbors." This nonparametric approach is "bottom-up" in the sense that many local comparisons can be agglomerated to develop a clear overall picture; traditional parametric methods are "top-down" in the sense that a global model is sought to make predictions for individual patients. New statistical concepts and treatment effect-size visualization tools are needed to make both local and global comparisons more understandable to all health care stakeholders. Truly local comparisons will inform two-way doctor-patient conversations on treatment choice, while the corresponding more global comparisons will use real world evidence to provide a more objective basis for health care policy decisions and regulation.

• **Objectives:** NISS will continue its ongoing OCER methodological research efforts with increased emphasis on the ***Local Control*** (LC) approach that is based upon patient micro-aggregation strategy. These efforts will be designed to confirm that the least biased and most relevant information from OCER data comes from making treatment comparisons only within numerous, small subgroups of highly similar patients. NISS researchers will also collect feedback from patients, caregivers, physicians and health policy makers to access and, ultimately, optimize human understanding of results from our proposed analyses and visual displays. Finally, NISS will create and validate open-source software that fully implements our proposals. Parallel development of proprietary "enterprise" software systems by potential NISS collaborators is actively encouraged.

• **Methods:** Methods for bias reduction in analyses of observational data have been actively studied over the last 40 years, but widely used OCER methods have tended to remain model-based (distinctly top-down) with ever increasing

complexity. Until recently, there was neither enough electronic health care data nor enough computer processing power available to demonstrate the practical advantages of going "back-to-basics" …using the simple "blocking" concepts championed by Cochran in his seminal papers on observational research of the 1960s. In the same way that bootstrap (re-sampling) algorithms have revolutionized statistical inference in practical applications where traditional statistical "theory" fails to provide viable solutions, the "bottom-up" approach of micro-aggregation has clear potential to become a disruptive technology for analysis of Big OCER data. The NISS team is uniquely well-qualified to provide the statistical expertise, practical experience in algorithm development, and communication skills necessary for long-term project success.

• **Results:** The LC approach is based upon patient micro-aggregation and uses key information from Electronic Medical Records to reveal ***distributions*** of nonparametric, "local" effect-size estimates. This makes LC strategy unique in providing a truly objective basis for individualized treatment choice. LC yields highly-credible answers to the key PCOR question: "Which treatment choice is most likely best for me?" Continuing NISS research is aimed at validating the observed heterogeneity in LC treatment effect-size distributions. The key question to be addressed is: "Can this heterogeneity literally be ***predicted*** from patient pretreatment X-characteristics?" Being able to answer this question will empower health care providers and policy makers, allowing them to literally "see" and understand the objective answers that Big OCER data can economically provide.

• **Conclusions:** Because patient micro-aggregation is computationally intensive, the LC approach failed to "scale up" adequately to be included within OMOP 2010 comparisons of methods for drug safety surveillance. Of course, only those (many fewer) patients with both a specific disease diagnosis and who chose one of the two treatments being compared, head-to-head, can provide relevant OCER information supporting that treatment choice. Furthermore, OCER methodology really needs to focus on key differences in statistical thinking between the "Science of Data Analysis" and the "Art of Model Fitting"

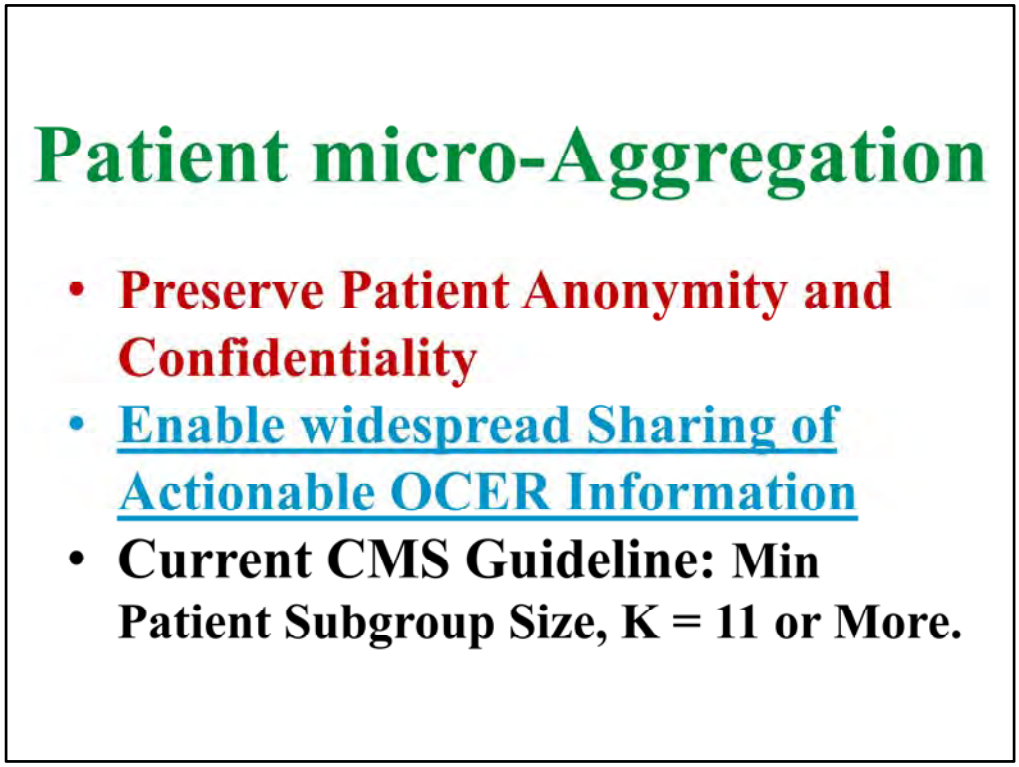
PCORI Methods PFA submission, R-1306-00788, August 2013
Patient Micro-Aggregation: "Bottom-Up" Statistical Methodology for Rapid Medical Learning

Patient Micro-Aggregation:

"Bottom-Up" Statistical Methodology for Rapid Medical Learning

Alan Karr, Director
Bob Obenchain, Research Fellow
Stan Young, Assistant Director for Bioinformatics

National Institute of Statistical Sciences (NISS)



A key focus of the statistical methods research proposed here is to demonstrate that micro-aggregation strategy is as effective as (or even more effective than) existing methods for protecting patient anonymity and confidentiality.

The key feature of micro-aggregation for preserving patient anonymity is that the size of each of many mutually exclusive and exhaustive patient subgroups can be restricted to be at least some minimum number, K. When K is sufficiently large, summary statistics computed within a subgroup tend to be considered remotely "safe" to publish. In other words, similar statistics from subgroups smaller than K are viewed as potentially compromising the anonymity and confidentiality of individual patients.

Under recently stated CMS publication policy, this minimum subgroup size is currently considered to be K=11. On the other hand, some outcomes researchers apparently feel that K=3 could even be adequate. There already is a considerable literature in this general topic area, and researchers associated with NISS have made many contributions to this literature. For example, see Karr, Kohnen, Oganian, Reiter and Sanil (2006).

Other than the minimum subgroup size restriction (K) discussed above, LC subgroups should be taken to be as numerous (and small) as is possible consistent with two key factors:

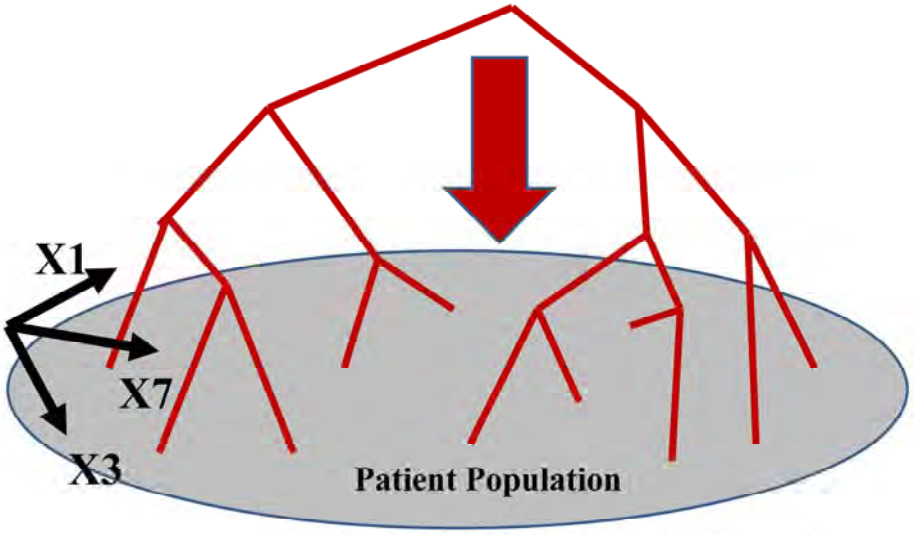
1. How much computational time and storage space are available for formation of realistic patient subgroups within a potentially enormous database of observational data?

Because these computations are parallelizable whenever a few "exact matches" are imposed (and found via simple sorting techniques), we anticipate that these sorts of restrictions will prove to be minimal. Is this truly the case?

2. What are the "local" reliability needs of an intended OCER analysis?

For example, suppose one needs to compare two treatments on time-to-event outcomes that are delayed, providing many censored "survival" times. This longitudinal setting could require using somewhat larger patient subgroups than those ideal in purely cross-sectional OCER analyses. After all, Local Control computations would then typically require estimation of within-subgroup Kaplan-Meier (1958) survival curves because LTDs will compute differences in, say, Restricted Mean Survival Times (RMSTs), Royston and Parmar (2011). Reliable estimates of these LTDs could then require from 100 to 300 patients within each subgroup.

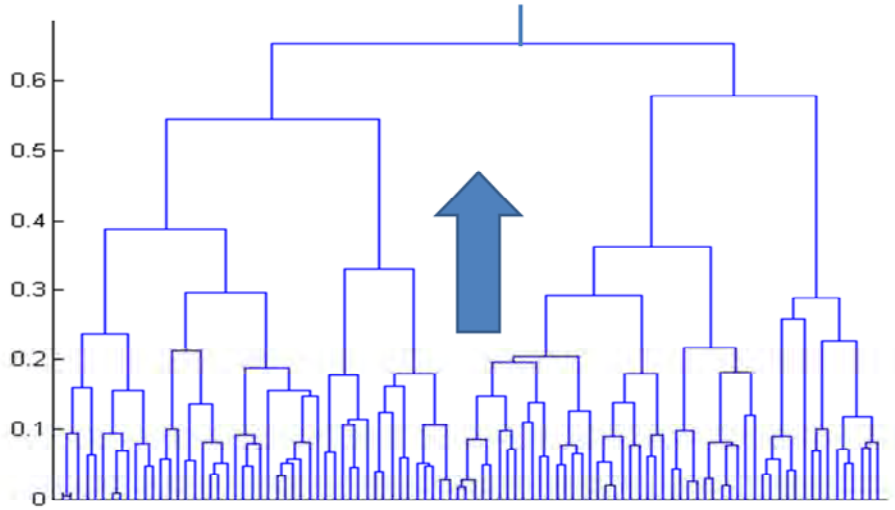
"Top-Down" Methods of Parametric Model Fitting



Supervised Learning

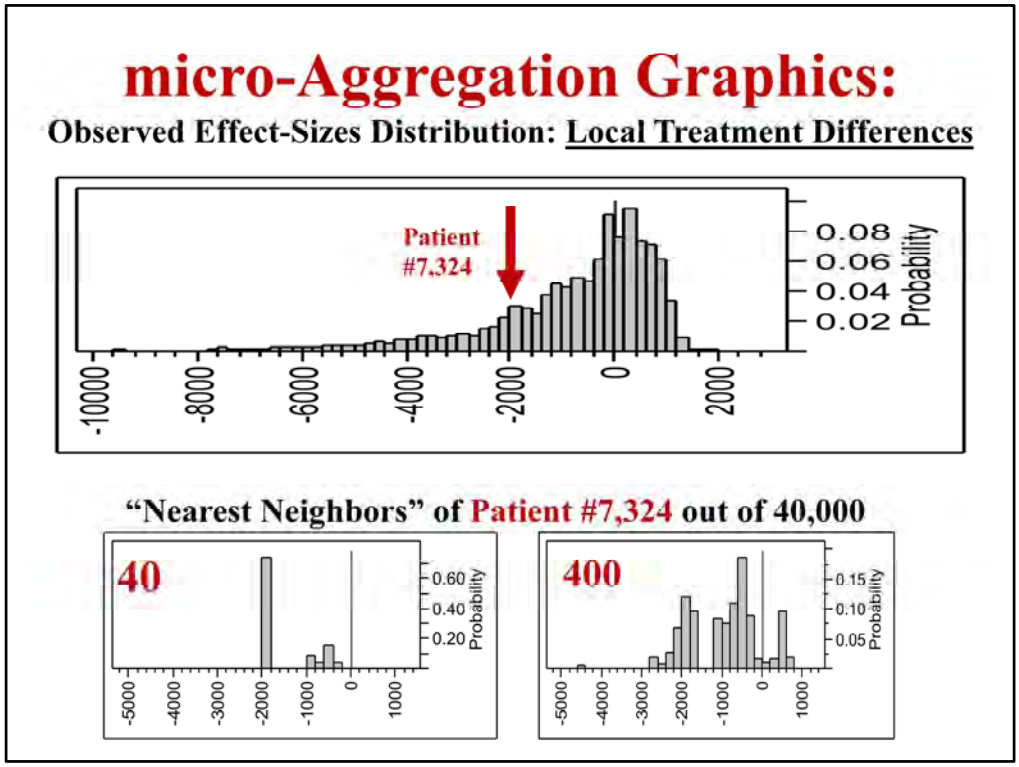
Unit of Information:
Finding from a "Study"

"Bottom-Up" Analysis: Patient micro-Aggregation



Unsupervised Learning …Nonparametric Preprocessing

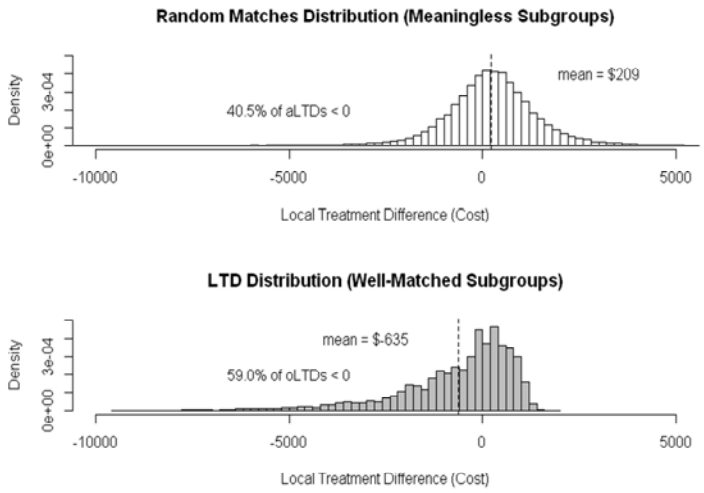
Unit of Information: Local Treatment Effect-Size (given X)



The initial phase of LC strategy is called "micro-Aggregation" and consists of dividing up a truly large dataset into literally thousands of small subgroups of well-matched patients. Each such subgroup may contain as few as 12 to 20 patients who are highly similar in terms of their known, pre-treatment X-characteristics. Either traditional "clustering" algorithms or any of the newer **unsupervised learning** techniques for near-neighbor matching can be used here. Note, in particular, that this phase (formation of patient subgroups in X-space) can be completed without knowing either [i] which Y-outcome variable or [ii] which T-treatment choice indicator will be used in subsequent LC analyses. This makes the LC approach more objective than alternative approaches that use information from Y-variables and/or T-indicators to guide their analyses, making their results less robust in the sense of being more sensitive to data outliers, leverage points and other quirks or anomalies within the available data.

This "micro-aggregation" strategy yields nonparametric conditional inferences and provides powerful, visual insights into treatment effect-size distributions. Specifically, a **Local Treatment Difference**, LTD = (Average Outcome on Treatment) minus (Average Outcome on Control), is computed within every subgroup that contains at least one treated and at least one control patient. The resulting collection of LTD estimates constitutes a detailed statistical distribution that provides an objective, quantitative basis for individualized medicine.

Confirm that X-Matching "Matters"

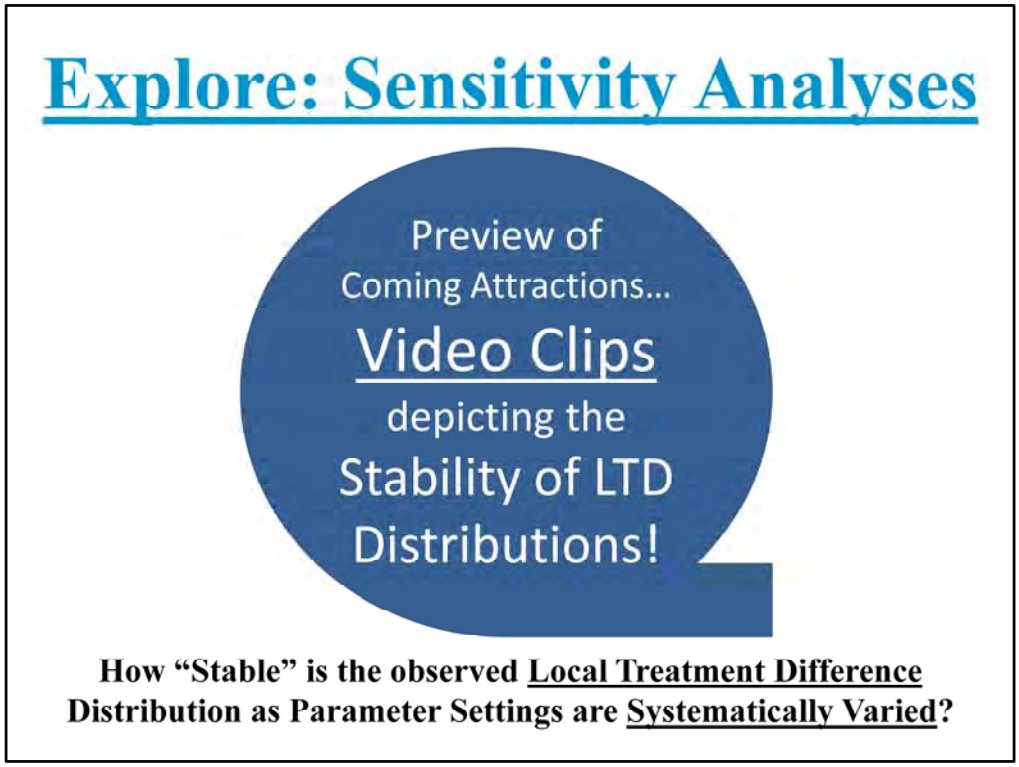


The top distribution of purely Random Matches (ignoring X) needs to be clearly DIFFERENT from that of Local Treatment Effect-Sizes

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

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

To eliminate any effects of the choice of the number of subgroups being formed, the sizes of these subgroups, or even the fractions of treated patients within these subgroups, the random subgroup formation process can exactly mimic the number, size and treatment-fraction statistics of the observed subgroups of well-matched patients.

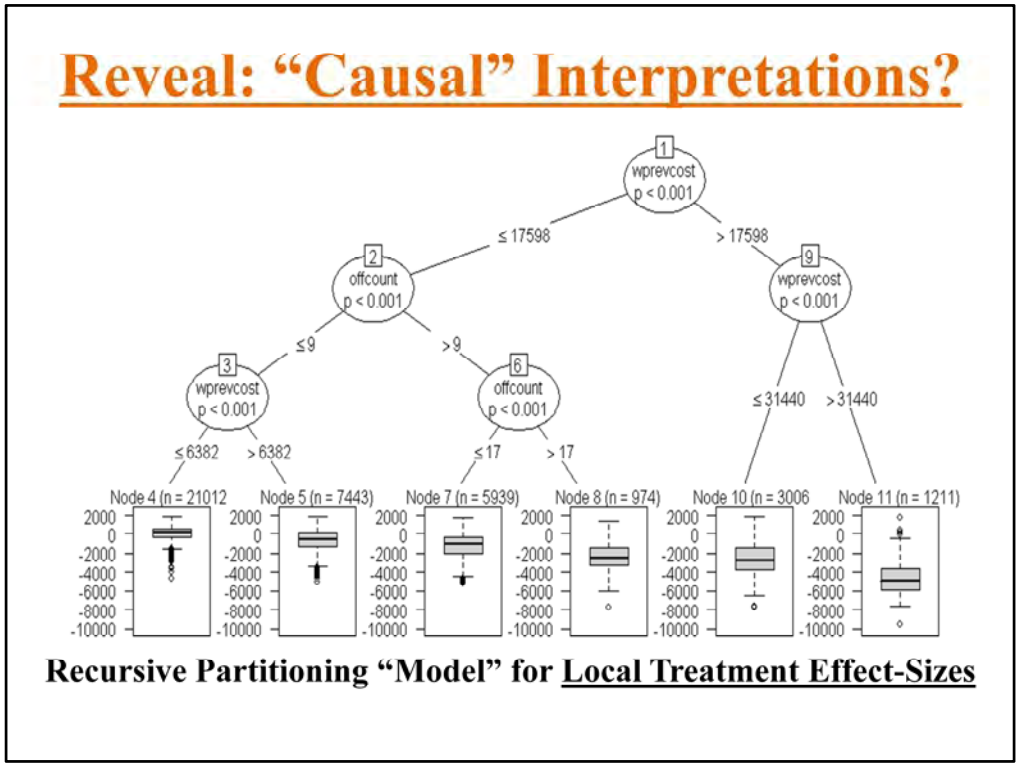


How does the observed LTD distribution change when…

[1] The X-characteristics used to determine patient pretreatment similarity are varied.
[2] The algorithm for patient matching / clustering changes.
[3] The number of mutually exclusive and exhaustive patient subgroups increases or decreases?

An advantage of the LC approach to conditional inference via micro-aggregation is that it requires the analyst to make only relatively few and simple choices to fine-tune an LC analysis. Only the three above basic types of LC "parameter settings" introduced above need be varied.

How stable are LTD distributions under the above sorts of changes? For a given dataset, can alternative LC settings change the location, spread, skewness or kurtosis of these LTD distributions? **Video Clips would allow health services researchers to literally "see" these effects.**



The fourth Phase of LC analysis is somewhat optional and has been rightfully postponed until last because it is, potentially, most likely to end in relative frustration, without revealing reliable HTE **predictions**. No patient micro-aggregation that failed in Phase II of LC analysis is even a candidate for consideration here in Phase IV. After all, if an observed LTD distribution is not clearly different from its corresponding, purely Random-Matches distribution, the X-variables it uses clearly do not "really matter" for predicting the Y-outcome currently under evaluation.

Quite naturally, the questions addressed in Phase IV are: How much do these X-variables matter? -- and/or-- Is there reasonable evidence that the observed LTDs do indeed vary with X in **understandable** and **predictable ways**? That is, in ways interesting enough to stimulate interest in doing confirmatory, follow-up studies using different data sources (observational or otherwise.)

The good news at the start of Phase IV is that the explorations performed within the first three phases of LC analysis usually have provided an interesting variety of observed LTD effect-size distributions to now model. In fact, these (nonparametric) **derived Y-outcomes** may prove to be much easier to "model" than the original observed Y-outcomes. Note, in particular, that all treatment effects have already been moved to the left-hand-side of the Phase IV model equations for predicting LTDs; it is unnecessary (and inappropriate) to include treatment indicator variables in the right-hand-side of a Phase IV model. This is one distinct advantage of LC over the earlier "nonparametric preprocessing" proposal of Ho, Imai, King and Stuart (2007).

While "success" in Phase IV modeling can signal genuine progress, "failure" is simply indicative of the relative importance of **Unmeasured Confounder** (UC) variables, showing that they genuinely need to be identified and added to OCER databases. The fact that patient pretreatment X-characteristics that are available apparently "do matter" means only that they are somewhat correlated with important UCs; "better" X-variables are needed to make "better" predictions of LTDs.

Traditional MODEL Fitting…

$$y = f(x'\beta) + e$$

$$\text{Data} = \boxed{\begin{matrix}\text{SIGNAL} \\ \text{(Expected Value)}\end{matrix}} + \begin{matrix}\text{Noise} \\ \text{(Error)}\end{matrix}$$

Typical Model Equation

$$E(Y \mid t, X) = \text{Overall Mean} + \text{Main Effect of Treatment}(t) + \text{Effects of X-covariates} + \text{Interactions (between } t \text{ and } X)$$

Technically, the binary t-choice variable is just another element (or column) of the X-vector (or matrix).

Interactions among X-variables (and powers of X-variables) are implicitly part of the "X effect" terms above.

Model fitting is typically a "zero-sum game" where the t and X variables are locked in deadly combat, vying for credit in "causing" Y-outcomes to vary across subjects (patients.) The artist's tools are his/her imagination, experience, perspective and initiative …plus Type IV**XYZ** ANCOVA sums-of-squares and corresponding p-values (for a possibly wrong theoretical distribution and a hopefully at least approximate model) accurate to at least 6 decimal places!

Model Equation for a Local Treatment Difference

$$E[\,(Y|t=1) - (Y|t=0) \mid X\,] =$$

Treatment Main-Effect (intercept) + **Effects of "centered" X-covariates**

Moving **both** the overall **Y intercept** and a **binary t-factor** to the left-hand side of the model equation can make a really BIG difference. In actual statistical practice, this becomes a realistic analysis strategy when the available data are truly BIG …lots and lots of diverse patients plus values of their relevant pretreatment X-characteristics.

Specifically, when sufficient Volume and Variety of subjects (patients) are available, the data become more like a census of a finite population, rather than a mere sample from an infinite one. After all, interpolations and extrapolations using some parametric model are then unnecessary!!!

The above left-hand **Estimand** is the very definition of a (local) treatment effect (at X). Tension / Competition between the t- and X-factors has been fully resolved (adjusted for). No t-factor terms are needed nor even ALLOWED on the right-hand side of parametric model equations for predicting LTDs (local effect-sizes).

Local Control

The Y-Outcome Estimand in LC:

$$LTD(X) = E[(Y \mid t = 1) - (Y \mid t = 0) \mid X]$$

Its Natural, Unbiased Estimator:

$$\overline{Y}_{t=1} - \overline{Y}_{t=0}$$ **…within a small patient cluster with centroid at *X*.**

This LTD estimand is truly **Heterogeneous** if and only if it not only (i) does truly vary as X varies but also (ii) represents a FIXED effect estimate (rather than only a RANDOM effect.) One obvious way to establish property (ii) is to demonstrate that LTDs are at least partially predictable from X.

Forming LTD effects does reduce precision. Unless "cross-over" data are available, (Y|t=1) and (Y|t=0) are counterfactual outcomes for every individual subject (patient); only one of these two outcomes can then be actually observed. In the most simple case where all Y-outcomes have the same variance (homoskedasticity), data from at least four subjects is needed to form an LTD estimate with the same precision or higher precision than a single Y. Again, this constraint is minimal when the available data are BIG.

Distinct "Statistical Thinking" Perspectives

Science of Data Analysis	Art of Model Fitting
Objective: Can be implemented via an "Expert System"	Subjective: Many too many ways to express "Research Initiative."
Cloud Computing for Sensitivity Analyses	Many Short Runs on a Workstation
Visualization, Clarity, Simplicity	Litany of Esoteric Methods
True Credibility	"Trust-Me" Science

Statistical Inference Perspectives

Science of Data Analysis	Art of Model Fitting
Robustness	Power
Effect-Sizes	p-Values
Treatment Effect Heterogeneity	"One-Size-Fits-All" Claims
Unsupervised …Let Data "Speak"!	Supervised …Deliberate Searches

Statistical Inference Perspectives

Science of Data Analysis	Art of Model Fitting
Accuracy-Unbiasedness	High Precision
Nested ANOVA …treatment within block	ANCOVA …"wrong" models
LOCAL	GLOBAL
Descriptive	Presumptive