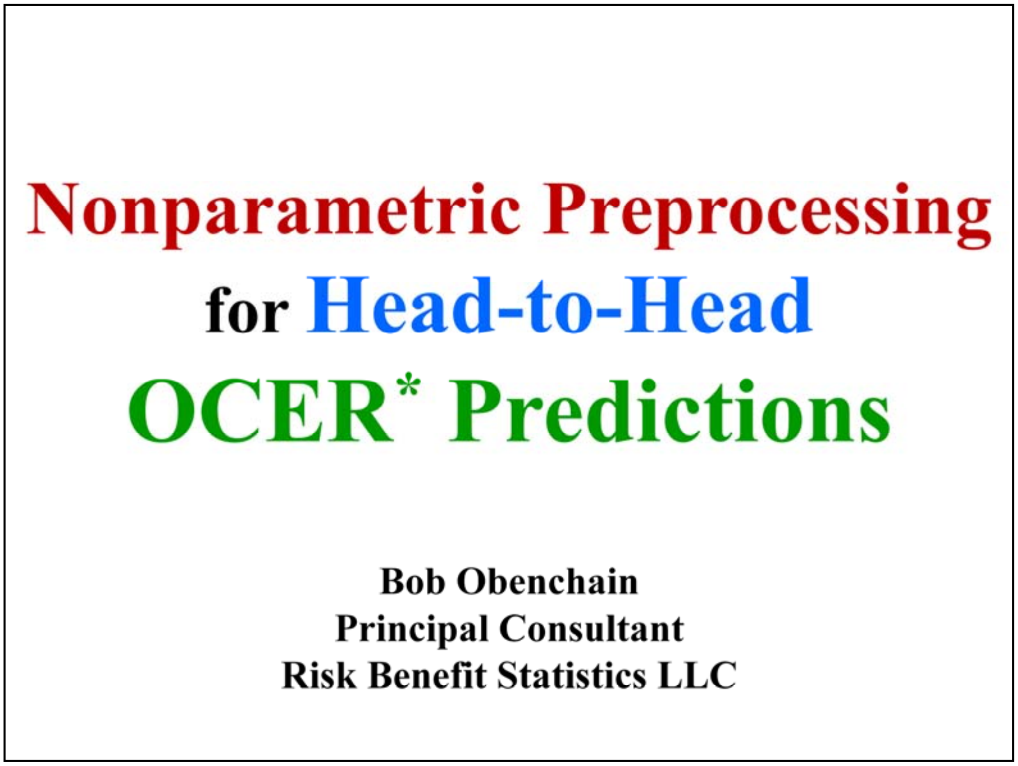
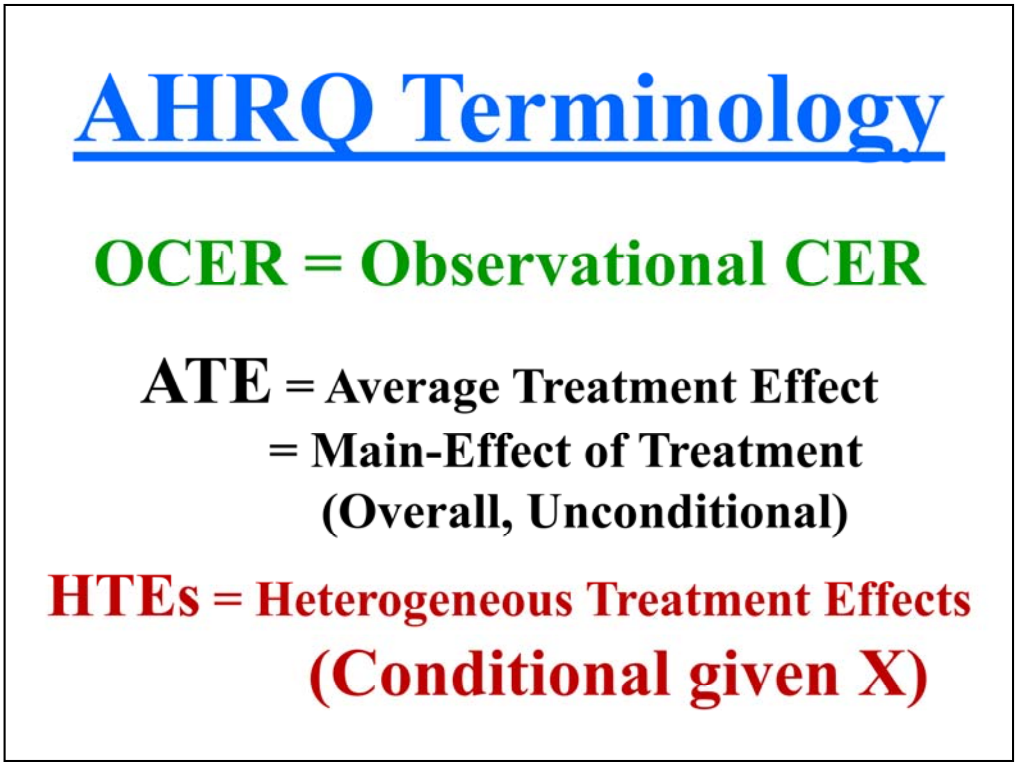


Nonparametric Preprocessing for Head-to-Head OCER* Predictions

Bob Obenchain
Principal Consultant
Risk Benefit Statistics LLC

At least three distinct approaches that feature **Nonparametric Preprocessing** have been proposed over the last 18 years for making head-to-head treatment comparisons within (big) observational datasets. Each of these approaches starts out by dividing patients into hundreds or even thousands of subgroups (blocks) of patients who are relatively well-matched on their pre-treatment X-covariates. The approach of McClellan et al. (1994) requires that all X-variables be instruments, while Ho et al. (2007) and Iacus et al. (2009a,b) make closely related proposals that proceed by matching treated and control patients in some fixed ratio within all subgroups before then resorting to traditional parametric modeling. In other words, these two types of approaches must make rather strong assumptions that can be quite unrealistic; see van der Laan and Rose (2010).

In sharp contrast, the Local Control (LC) approach of Obenchain (2010) computes local Average Treatment Effects (ATEs) and forms a full distribution of heterogeneous effect-sizes. LC thus persists in making only minimal assumptions; its statistical model is equivalent to simple nested ANOVA (treatment within subgroup.) Today's presentation will focus on the somewhat surprising result, verified via extensive simulations of "realistic" observational data, that LC predictions can be more accurate than those from parametric fits. After all, LC lends itself to relatively harmless forms of deliberate over-fitting that, by forfeiting causal "interpretability," can adequately handle all forms of bias expected in observational data. In particular, I will argue that LC analyses yield realistic measures of minimal uncertainty in observational data and help detect presence of unmeasured confounders.



AHRQ Terminology

OCER = Observational CER

ATE = Average Treatment Effect
= Main-Effect of Treatment
(Overall, Unconditional)

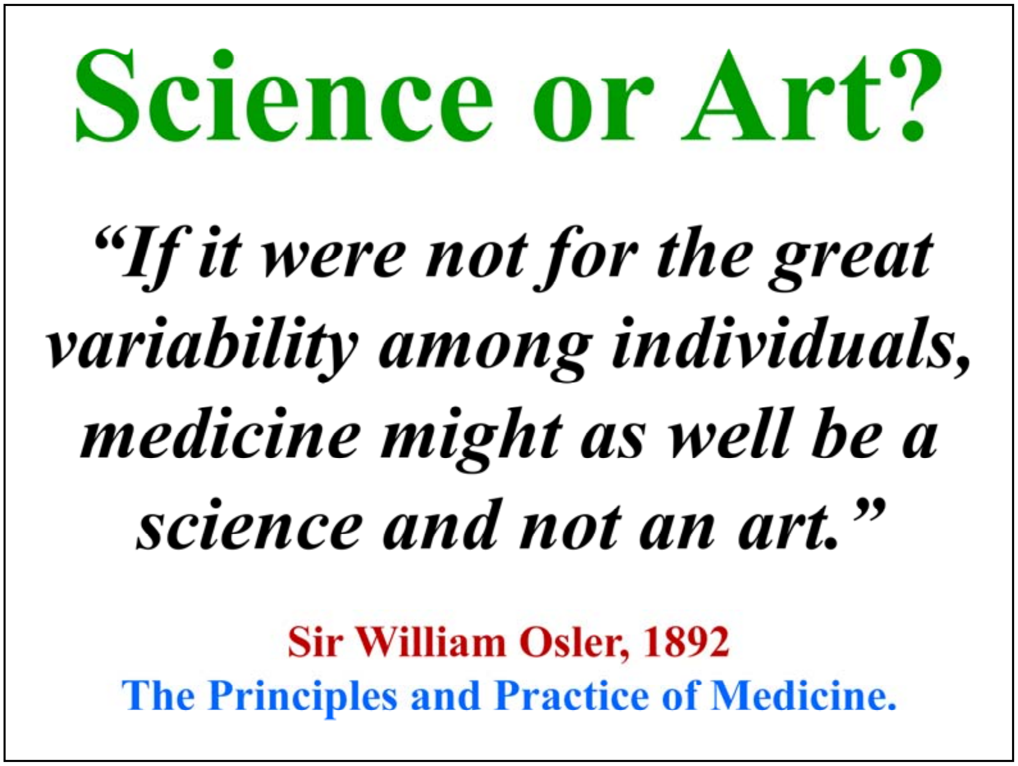
HTEs = Heterogeneous Treatment Effects
(Conditional given X)

AHRQ (2012) **Developing a Protocol for Observational Comparative Effectiveness Research (OCER): A User's Guide** (Comment period for Draft Version, May 11 – June 8, 2012.)

No longer Posted at: http://www.effectivehealthcare.ahrq.gov/index.cfm/research-available-for-comment/

Page 54 of 228:

When this variability encompasses treatment effects of different directions, i.e., both benefit and harm, this is sometimes called a **qualitative** treatment interaction, whereas differences in the magnitude of treatment effect in the same direction are called **quantitative** interactions.



Science or Art?

"If it were not for the great variability among individuals, medicine might as well be a science and not an art."

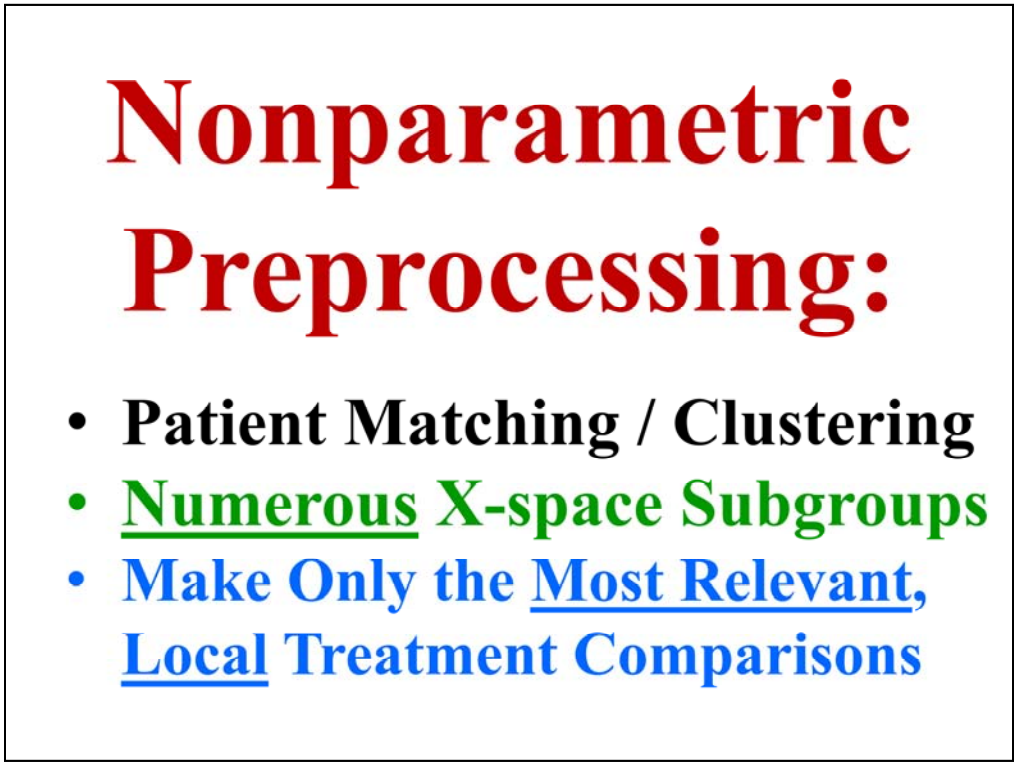
Sir William Osler, 1892
The Principles and Practice of Medicine.

Statisticians have historically considered themselves to be self-appointed "guardians of the scientific method." To retain any semblance of that role, it is about time for US to truly **step up** and **answer** the above **challenge**:

In the absence of HTE estimates as an OBJECTIVE basis for individualized medicine, medicine will remain an ART and not a SCIENCE.

Causal inference uses only ATEs and is actually a highly SUBJECTIVE art form.

Osler was quoted in Kaplan et al. (2010) "Who Can Respond to Treatment?" Medical Care • Volume 48, Number 6 Suppl 1, June …work sponsored by AHRQ.



Nonparametric Preprocessing:

- Patient Matching / Clustering
- Numerous X-space Subgroups
- Make Only the Most Relevant, Local Treatment Comparisons

Ho, Imai, King and Stuart Matching as nonparametric preprocessing… ***Political Analysis*** 2007

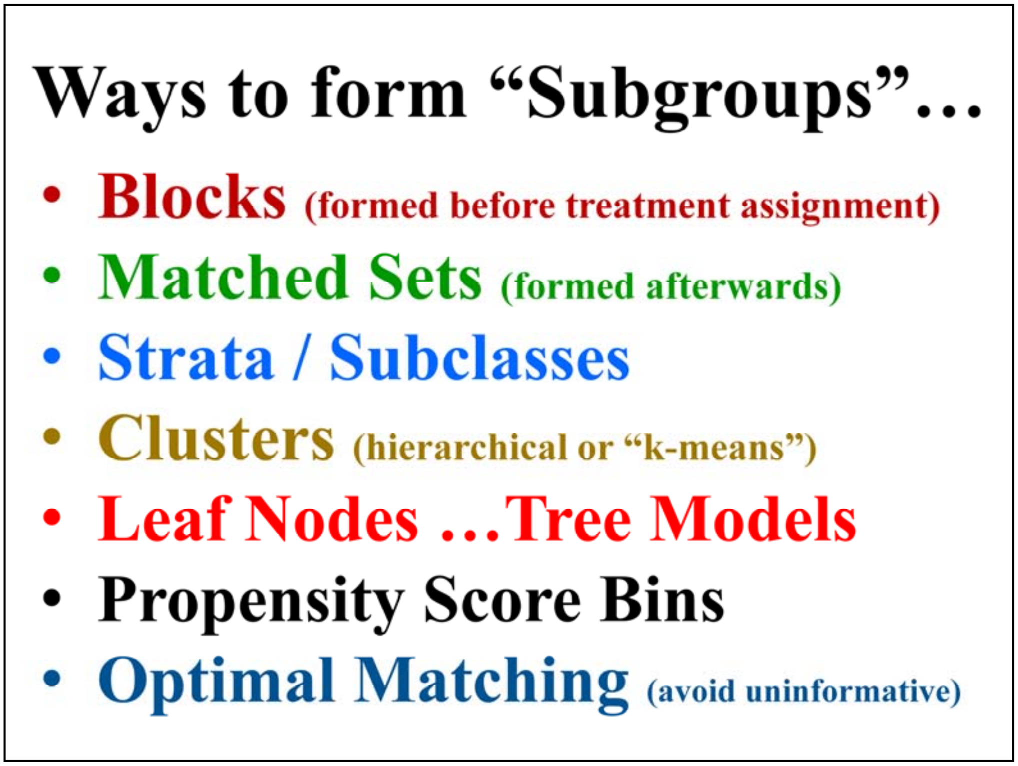
Iacus, King and Porro 2009 Coarsened Exact Matching: Casual Inference Without Balance Checking.

Rather than fitting parametric equations, adopt a **simple "divide and conquer" strategy.**

Future observational datasets promise to be MASSIVE. There may be no need to interpolate between or extrapolate beyond the available data; the data will NOT consist of only a few, scattered observations. Therefore, a completely valid strategy will be to… LET THE DATA SPEAK FOR THEMSELVES !!! What they "say" may not be very simple and totally coherent, but it almost certainly will be quite REALISTIC.

Patient Level LTD Estimates = HTEs --> KEY, NEW OUTCOME measure for head-to-head treatment comparisons.

Finally, here is a fundamentally "simple" health care data analysis STRATEGY (NPP) that doctors and patients -- and everybody else -- can truly understand and easily appreciate.



Ways to form “Subgroups”…

- Blocks (formed before treatment assignment)
- Matched Sets (formed afterwards)
- Strata / Subclasses
- Clusters (hierarchical or “k-means”)
- Leaf Nodes …Tree Models
- Propensity Score Bins
- Optimal Matching (avoid uninformative)

There are many alternative ways to define or describe them. Here, subgroups of patients are assumed to be mutually exclusive and exhaustive.

[1] Patents within a single subgroup are to have some common characteristic(s) or be as similar as possible.

[2] Patents in all other subgroups are to have different characteristic(s) or be as dissimilar as possible from the patients in any given subgroup.

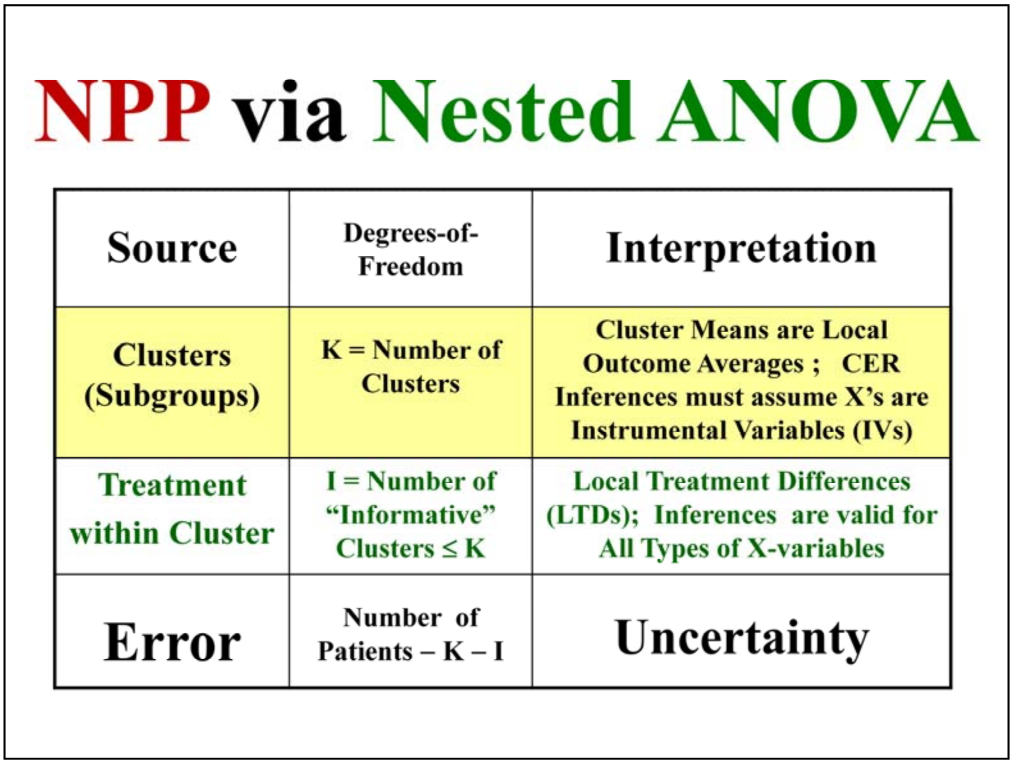
Subgroups are most typically formed in an “unsupervised” way; i.e. based **only** upon known patient pre-treatment X-characteristics.

“Matched Sets” typically do not (and **SHOULD not**) require any fixed ratio of “treated” to “control” patients.

Knowledge of treatment choice (trtm = 0 or 1) is used in the last three **supervised** approaches: classification trees, discrete choice models (such as logistic regression) and optimal matching.

However, knowledge of patient responses (y-outcomes) should almost never be used in forming subgroups, especially when “matching” patients.

A subgroup is said to be “uninformative about its local treatment difference” when it is PURE in the sense that it contains either ONLY trtm = 0 patients or else ONLY trtm = 1 patients.



NPP via Nested ANOVA

Source	Degrees-of-Freedom	Interpretation
Clusters (Subgroups)	K = Number of Clusters	Cluster Means are Local Outcome Averages ; CER Inferences must assume X's are Instrumental Variables (IVs)
Treatment within Cluster	I = Number of "Informative" Clusters ≤ K	Local Treatment Differences (LTDs); Inferences are valid for All Types of X-variables
Error	Number of Patients – K – I	Uncertainty

McClellan, McNeil, Newhouse (1994) and many economists have studied "instrumental variable" approaches. The key assumption is that observed X-covariates determine only treatment selection and do NOT influence outcome, Y, except through treatment choice. MM&C (1994) proposed that cluster means be plotted vertically against a horizontal axis depicting within-cluster fraction treated (**observed propensity for treatment**.) This approach uses information only from the "Clusters" row of the ANOVA table. MM&C (1994) further contended that trends (up or down) in the displayed values from left-to-right across this plot are **interpretable as causal effects when all X-variables used to form patient clusters are instrumental variables**.

The Local Control approach uses information only from the "Treatment within Cluster" row of the ANOVA table and yields the "observed LTD" distribution shown at the bottom of Slide 15. Interpretation of an LTD distribution does NOT require X-variables to be "instruments" or make any other sorts of un-testable assumptions.

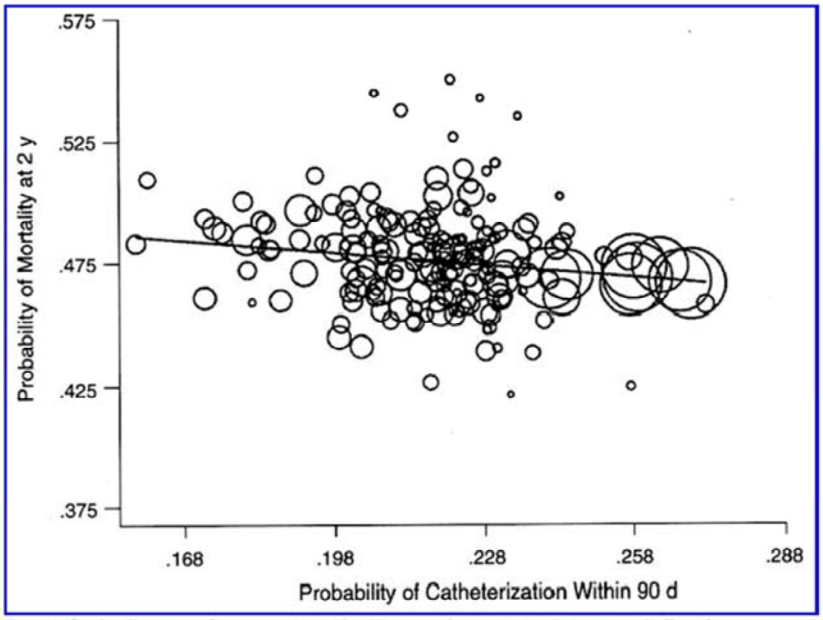


Figure 1. Graphic illustration of instrumental variables (IV) approach to estimating the incremental effect of invasive procedure use on acute myocardial infarction mortality at 2 years. Each circle represents a group of patients in a particular differential-distance group with a particular set of age, gender, and race characteristics. The model is thus fully adjusted for effects of observable patient demographic characteristics, and remaining differences in probabilities of catheterization across groups are strictly the result of different IV values for patients who are identical in terms of observable characteristics. The area of each circle is proportional to the number of patients in the group. The slope of the line fitted to the cells estimates the weighted-average effect of invasive procedure use

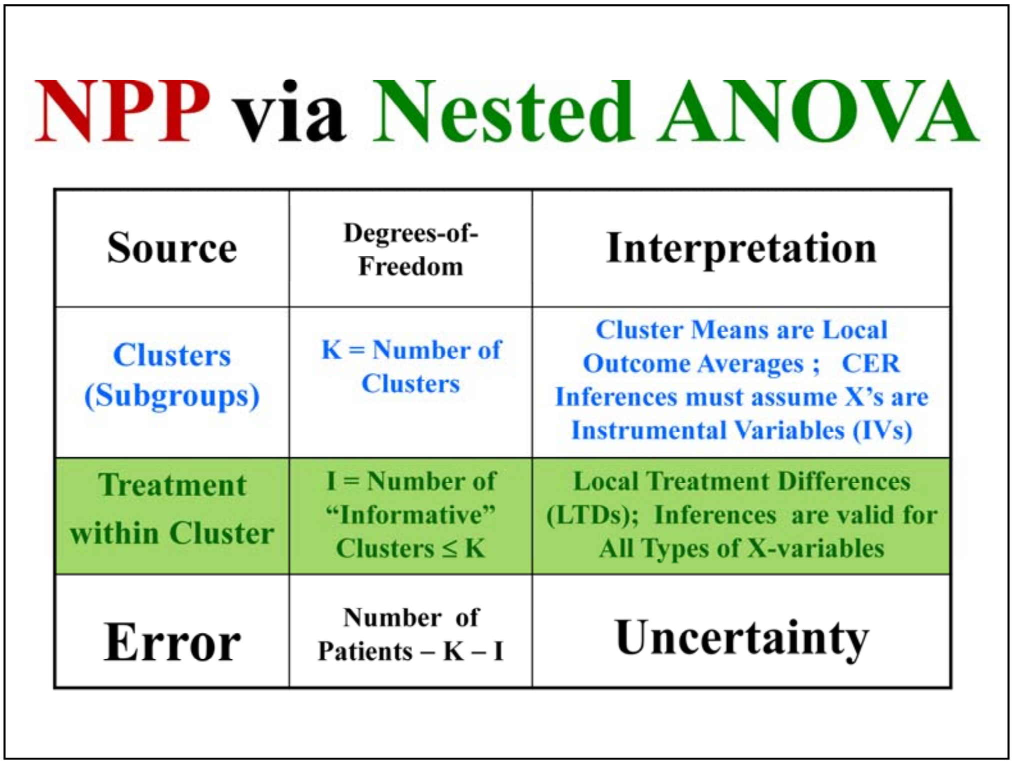
Once Mark McClellan became commissioner of the FDA, this paper became an instant “classic.”

N = 200K elderly AMI patients for whom “distance from the hospital” of admission could be computed from ZIP code information.

Note that “distance from the hospital” is a plausible instrument here because patients who live in a big city near to a big hospital are more likely ti receive an (expensive) invasive procedure (whether they really need it or not.) Thus “distance from the hospital” should be predictive of choice of treatment without being predictive of outcome (except through treatment.)

Here clusters were (apparently) formed primarily by dividing patients into “distance from the hospital” bands. Clusters were then formed within bands by matching on sex and race and grouping into (elderly) age subclasses.

The Wald(1940) “Grouping Estimator” is the slope of the line connecting any two cluster centroids. The OLS fit displayed here represents an appropriately weighted average over all possible pairs of patient subgroups.

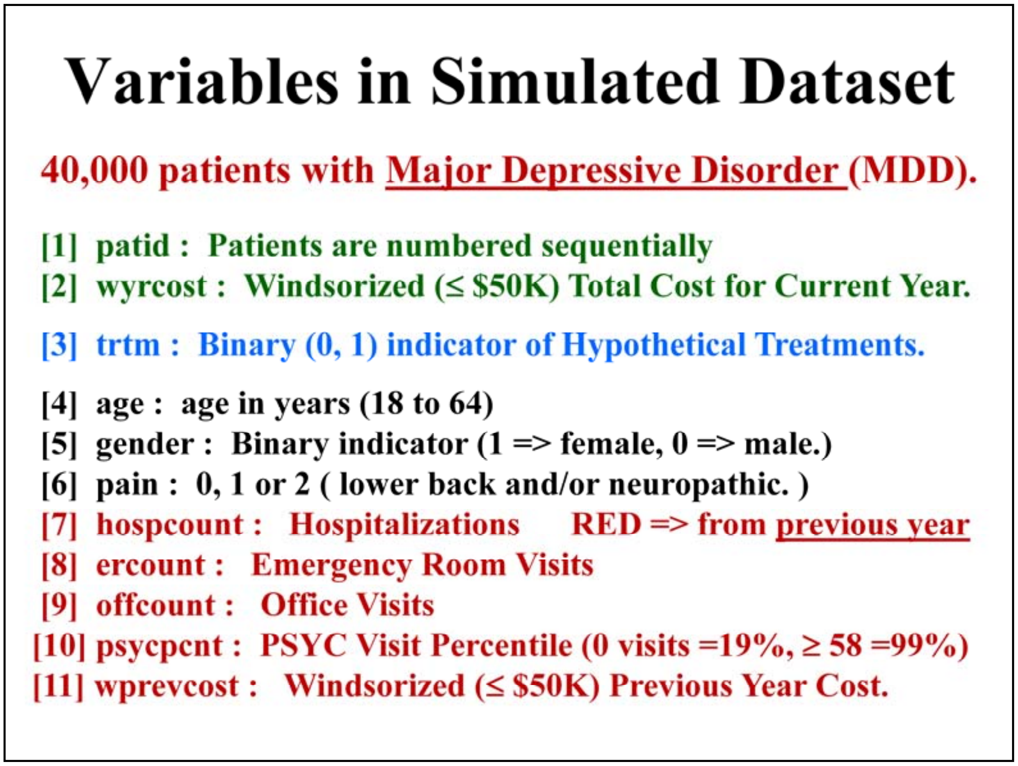


NPP via Nested ANOVA

Source	Degrees-of-Freedom	Interpretation
Clusters (Subgroups)	K = Number of Clusters	Cluster Means are Local Outcome Averages ; CER Inferences must assume X's are Instrumental Variables (IVs)
Treatment within Cluster	I = Number of "Informative" Clusters ≤ K	Local Treatment Differences (LTDs); Inferences are valid for All Types of X-variables
Error	Number of Patients – K – I	Uncertainty

McClellan, McNeil, Newhouse (1994) and many economists have studied "instrumental variable" approaches. The key assumption is that observed X-covariates determine only treatment selection and do NOT influence outcome, Y, except through treatment choice. MM&C (1994) proposed that cluster means be plotted vertically against a horizontal axis depicting within-cluster fraction treated (**observed propensity for treatment**.) This approach uses information only from the "Clusters" row of the ANOVA table. MM&C (1994) further contended that trends (up or down) in the displayed values from left-to-right across this plot are **interpretable as causal effects when all X-variables used to form patient clusters are instrumental variables**.

The Local Control approach uses information only from the "Treatment within Cluster" row of the ANOVA table and yields the "observed LTD" distribution shown at the bottom of Slide 15. Interpretation of an LTD distribution does NOT require X-variables to be "instruments" or make any other sorts of un-testable assumptions.



Variables in Simulated Dataset

40,000 patients with Major Depressive Disorder (MDD).

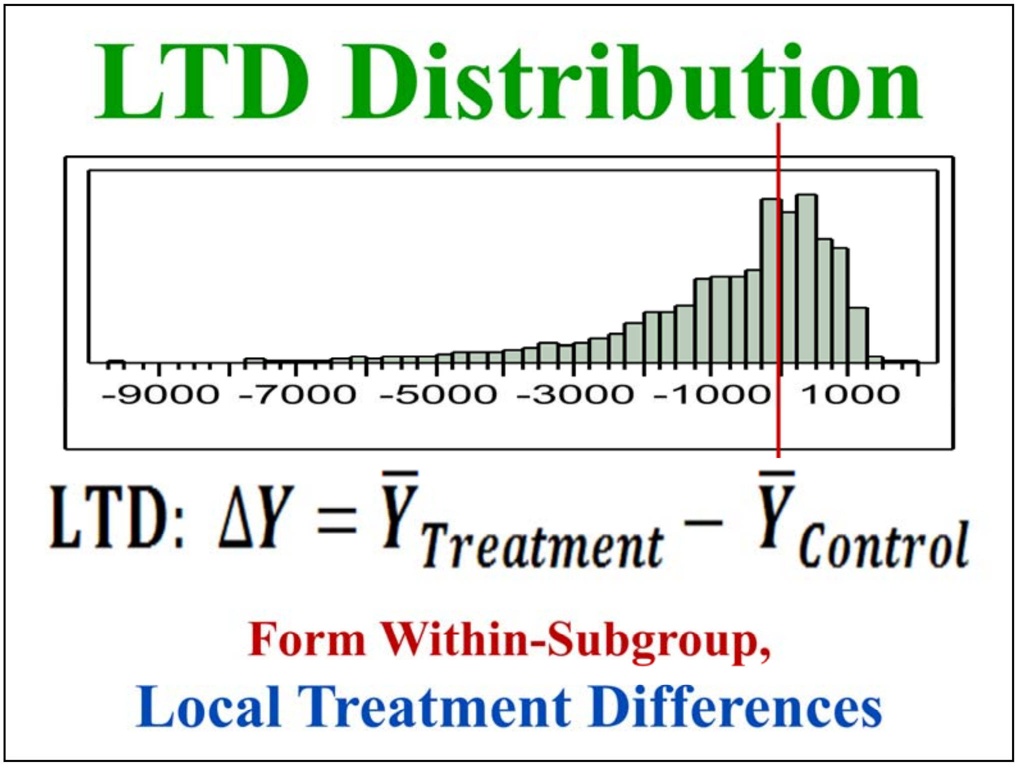
[1] patid : Patients are numbered sequentially
[2] wyrcost : Windsorized (≤ $50K) Total Cost for Current Year.

[3] trtm : Binary (0, 1) indicator of Hypothetical Treatments.

[4] age : age in years (18 to 64)
[5] gender : Binary indicator (1 => female, 0 => male.)
[6] pain : 0, 1 or 2 (lower back and/or neuropathic.)
[7] hospcount : Hospitalizations RED => from previous year
[8] ercount : Emergency Room Visits
[9] offcount : Office Visits
[10] psycpcnt : PSYC Visit Percentile (0 visits =19%, ≥ 58 =99%)
[11] wprevcost : Windsorized (≤ $50K) Previous Year Cost.

The 8 patient baseline X-characteristic variables (#s 4 to 11) show highly realistic variation in the sense that they are quite similar to those of actual MDD patients in an administrative claims database.

All of the “wyrcost” values are simulated. The treatments being simulated here are both hypothetical; in particular, no current MDD medication may behave much like trtm = 1 does here …relative to trtm = 0 = “control.”



LTD Distribution

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

Form Within-Subgroup,

Local Treatment Differences

LTDs = Local Treatment Differences. Data on total yearly cost of treatment for MDD from 40K patients hierarchically clustered into 2K relatively homogeneous subgroups in patient X-characteristics from the previous-year (average subgroup size = 20 patients.) Control = current standard of care; Treatment = hypothetical new and more effective but expensive alternative.

Note that 41% of the 39,585 patients with estimated LTDs are **positive**, but the mean is **negative** $635.

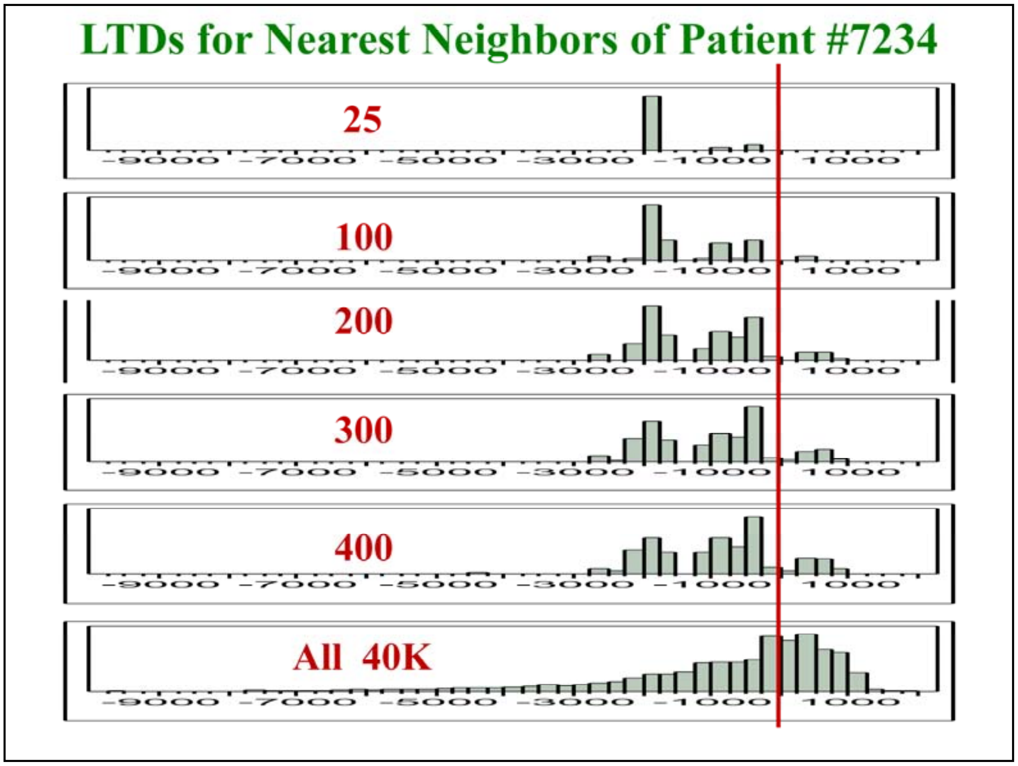
This sort of display provides an objective basis for individualized treatment choices. It depicts the distribution of local, observed effect-sizes estimates …using a simple histogram.

In observational research, it's "too late" to rely on randomization to make treatment cohort comparisons more fair.

But it's never too late to use BLOCKING. This strategy yields LOCAL comparisons that are as UNBIASED as possible relative to all OBSERVED patient pre-treatment characteristics.

Display the full Distributions => Retain all of the information you can from all patients in all blocks.

In other words, blocking makes only the **most relevant treatment comparisons**


LTDs for Nearest Neighbors of Patient #7234

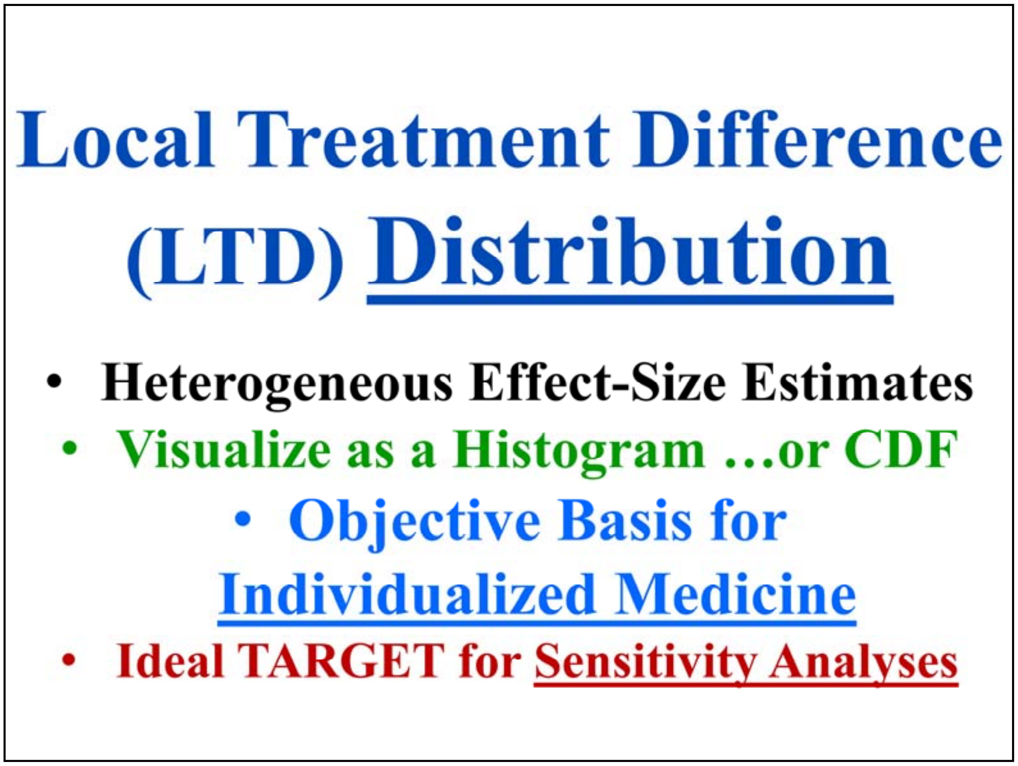
Individualized medicine demands a full, meaningful **distribution of effect-size estimates** as patient X-characteristics vary …in ways that are not (and cannot be) made explicit within a single, simple-to-understand LTD display.

Still, as illustrated in the above 6 histograms, LTDs for the “patients most like me” can be displayed as the definition of “nearest neighbors” (in X-space) is relaxed.

Here, 21 of the 24 Nearest Neighbors of patient #7234 save approximately $2K per year by taking the newer, more expensive treatment for MDD. Two other NN’s save only $400 and one saves $800.

On the other hand, 1% of 40K patients is 400, and these more distant NN’s include 53 (12%) with positive LTDs.

Allowing observational data to “speak for themselves” in this way provides doctors and patients with the essential information about **outcome heterogeneity and uncertainty** that they truly need to confidently make treatment recommendations and choices.

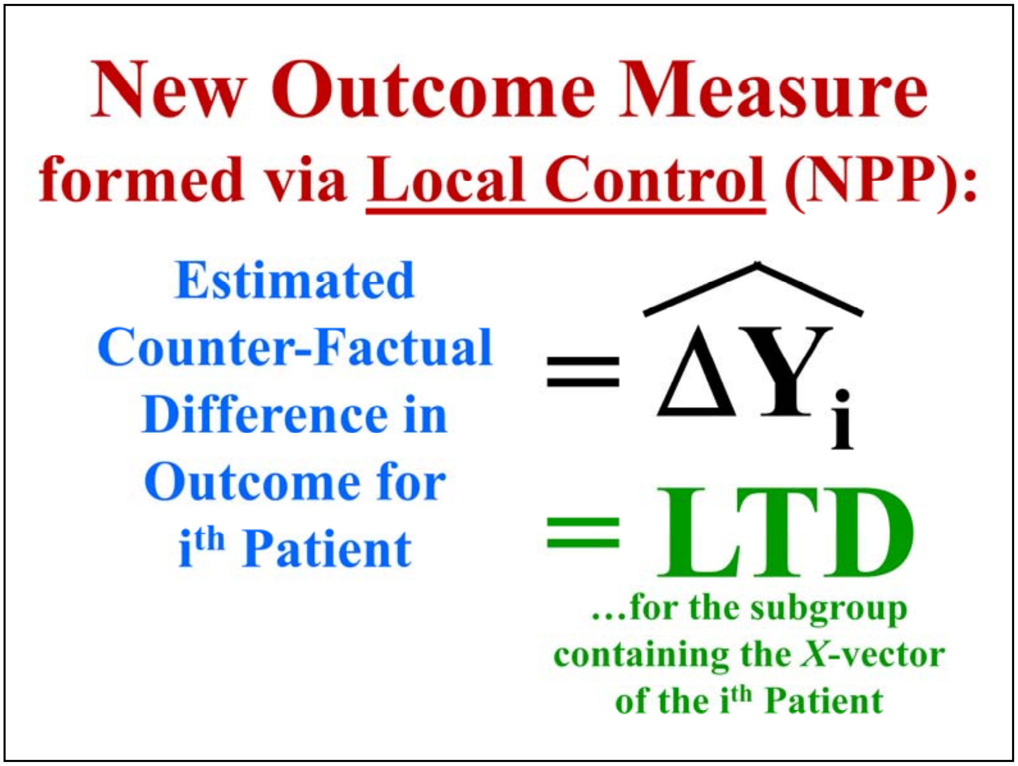


Local Treatment Difference (LTD) Distribution

- Heterogeneous Effect-Size Estimates
- Visualize as a Histogram …or CDF
- Objective Basis for Individualized Medicine
- Ideal TARGET for Sensitivity Analyses

The “main-effect” of treatment is the MEAN of the LTD distribution. In the above example, this main-effect is **negative** and **highly significant**, statistically. Unfortunately, this traditional sort of “overall finding” is frequently somewhat misleading. For example, it certainly does NOT mean here that all MDD patients are expected to reduce their total yearly costs by taking the newer and more expensive treatment.

With neither a well defined “target parameter” nor an explicit model allowing attention to be focused on it’s key propertpes, sensitivity analysis could not be performed.



New Outcome Measure formed via Local Control (NPP):

Estimated Counter-Factual Difference in Outcome for ith Patient $= \widehat{\Delta Y_i}$

= LTD

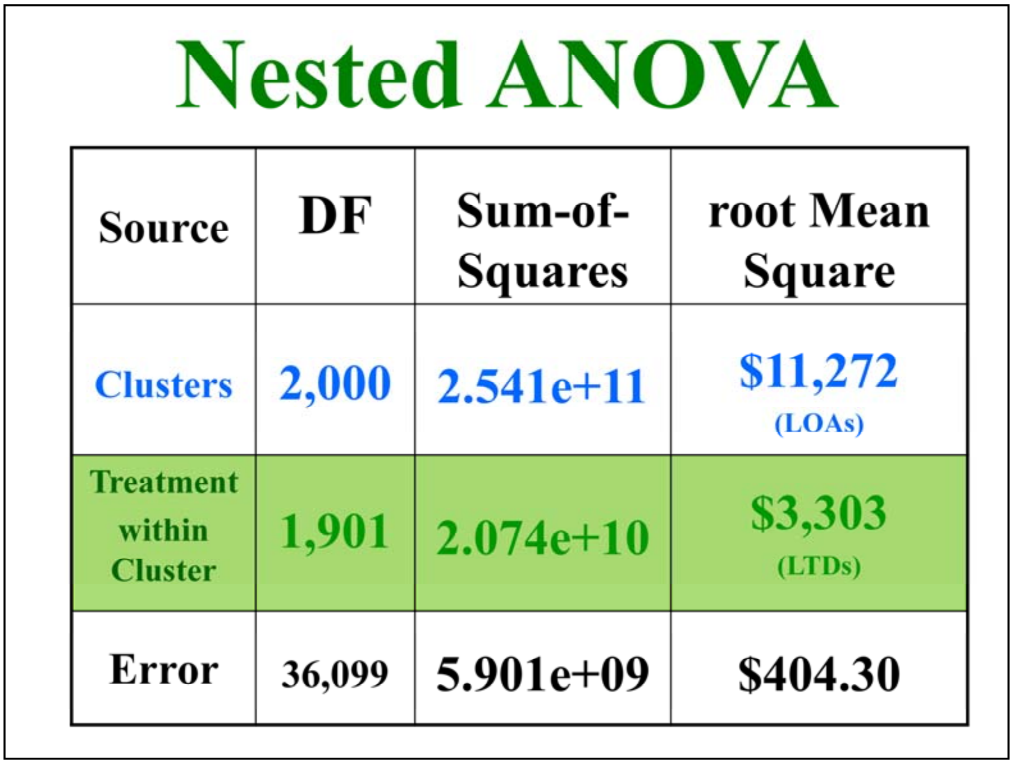
…for the subgroup containing the *X*-vector of the ith Patient

NPP --> HTE via calculation of hundreds or even thousands of “LOCAL” ATEs.

In our example, we assume that LC has formed 2,000 patient subgroups in an 8-dimensional X-space …yielding 1,901 informative subgroups containing 39,585 patients (0.989625 => 99.0% of the original 40K patients.)

In other words, LC has created CFD estimates for 39,585 patients that consist of only 1,901 potentially distinct numerical values. But this rather clearly corresponds to **considerable over-fitting** of a Nested ANOVA model that, by deliberate choice, makes relatively FEW, WEAK and REALISTIC assumptions!

Let us now explore the question of how **GOOD** these LTD predictions might actually be!



Nested ANOVA

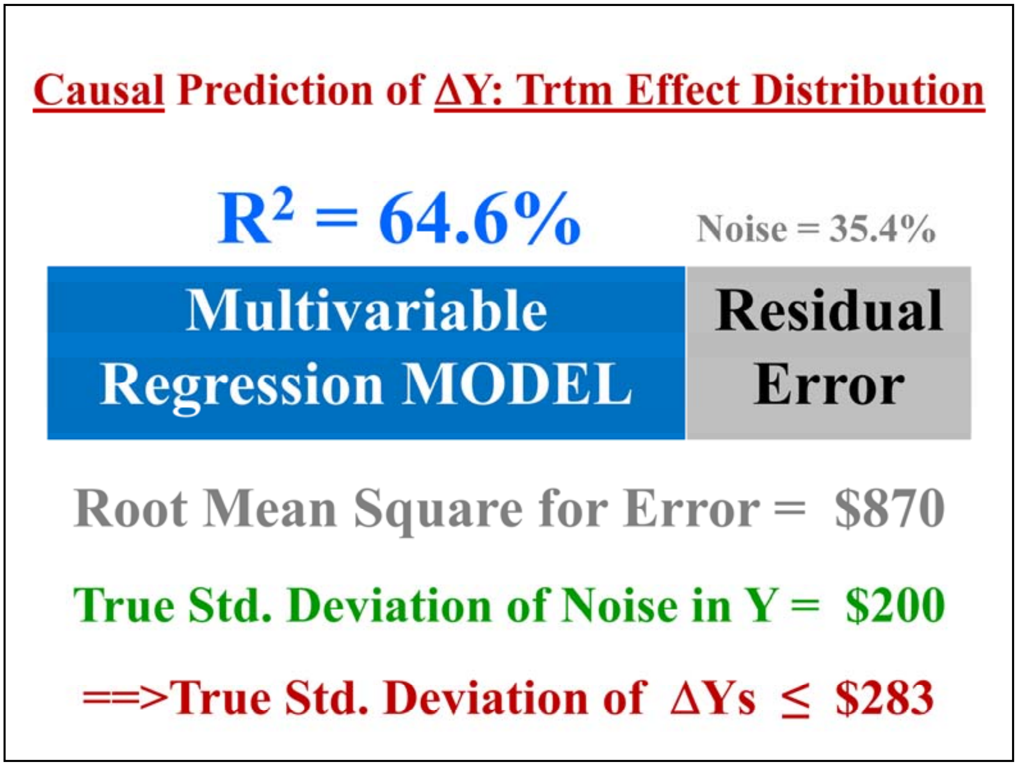
Source	DF	Sum-of-Squares	root Mean Square
Clusters	2,000	2.541e+11	$11,272 (LOAs)
Treatment within Cluster	1,901	2.074e+10	$3,303 (LTDs)
Error	36,099	5.901e+09	$404.30

NPP encourages a relatively harmless for of deliberate over-fitting that can yield truly excellent predictions.

The R-square for this 2K subgroup Nested ANOVA model is a whopping 98.0% here.

Technically, the ANOVA tradition is to “remove” a single degree-of-freedom for the overall mean from the table. Here, I have included that effect within the “clusters” row …where I think it “belongs” from a sequential sum-of-squares perspective. Clearly, I have also not separated out the “main effect” of treatment across blocks!

However. the computed **root Mean Square for Error** is roughly double the true **Std. Error = $200** of the additive white nose actually added to true expected values in the simulation that generated these data.



Causal Prediction of ΔY: Trtm Effect Distribution

$R^2 = 64.6\%$ Noise = 35.4%

Multivariable Regression MODEL

Residual Error

Root Mean Square for Error = $870

True Std. Deviation of Noise in Y = $200

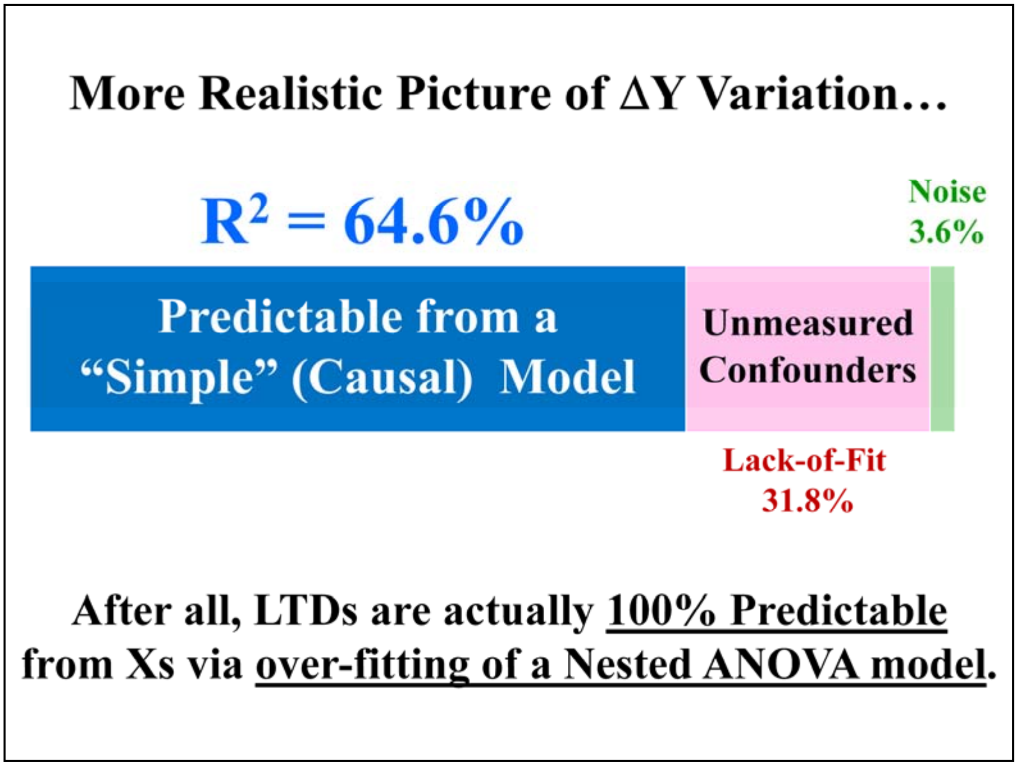
==>True Std. Deviation of ΔYs ≤ $283

Quite simply stated: This is a “WRONG” model. The simulation inserted “unmeasured confounders” (systematic lack-of-fit) as well as measurement error (white noise) into the resulting synthetic yearly cost data !!!

The difference of two independent variables has variance equal to the sum of its component variable variances. Thus, forming an LTD can increase the standard deviation by as much as sqrt(2) = 1.414. On the other hand, the standard deviation of LTDs can (alternatively) be greatly reduced by **averaging Y-outcomes over several patients choosing the same treatment** before forming a counter-factual difference.

Thus a rather clear trade-off is quite possible. More but smaller subgroups ==> better bias correction. Fewer but larger subgroups ==> higher precision in LTD estimation.

The model fitted to LTD estimates here uses a total of 50 degrees-of-freedom, is second-degree factorial in all 8 X-variables and includes squared terms in all 6 continuous X-variables. (female is binary; pain is ordinal with 3 levels.)



More Realistic Picture of ΔY Variation…

$R^2 = 64.6\%$

Predictable from a "Simple" (Causal) Model

Unmeasured Confounders

Noise 3.6%

Lack-of-Fit 31.8%

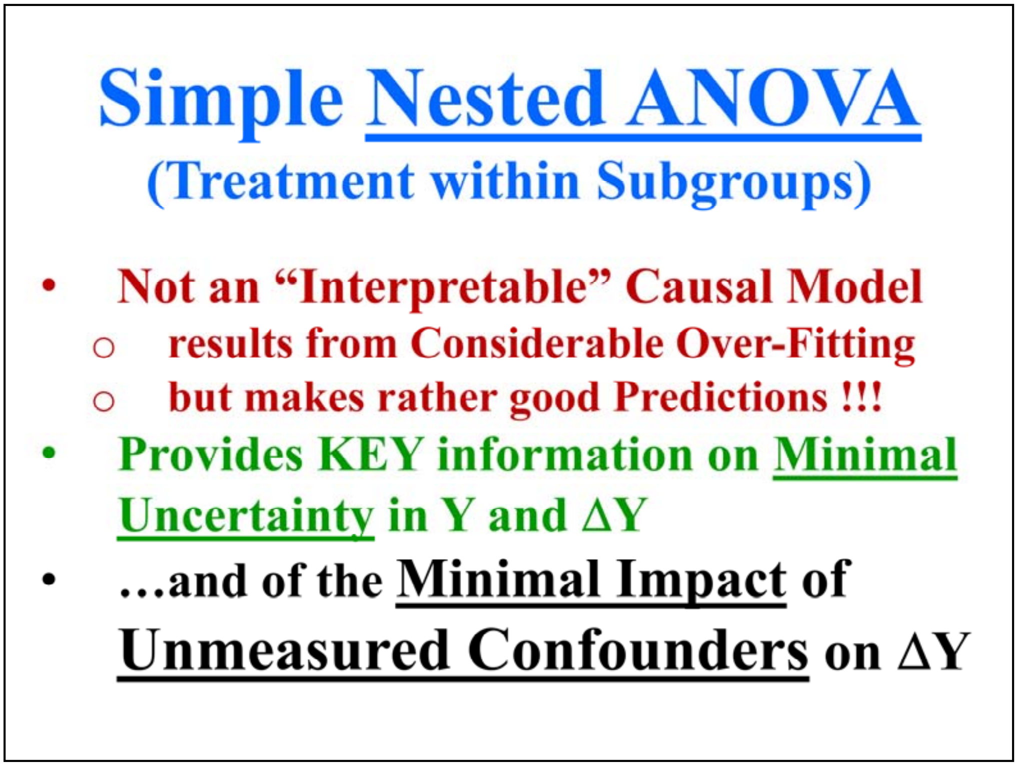
After all, LTDs are actually 100% Predictable from Xs via over-fitting of a Nested ANOVA model.

NPP goes well beyond traditional global, parametric models in realistically assessing Minimal Uncertainty in the original Y-outcome in observational data …event though NPP is still tends, quite naturally, to be somewhat conservative (over-estimation of noise.)

Boundary between the "signal" components that either **are** or **are not** predictable from observable pre-treatment characteristics of patients tends to be less clear-cut. Here the 8 X-variables are aggressively claiming 64.6% of the "credit" …percent of total variability in Windsorized total yearly cost incurred by MDD patients.

I have not shown any Systematic Sensitivity visualizations today. Good algorithms are badly(?) needed for this …or current capability PLUS **cloud computing** ???

TECHNICAL BACKGROUND: Always remember that the Error Mean Square potentially has two distinct variance components …which may be difficult to separate and identify. These two additive components are: Lack-of-Fit of one's model (especially when a global model is too simple and smooth or when unmeasured confounders are active) and True Noise (measurement error) in the observed response variable.



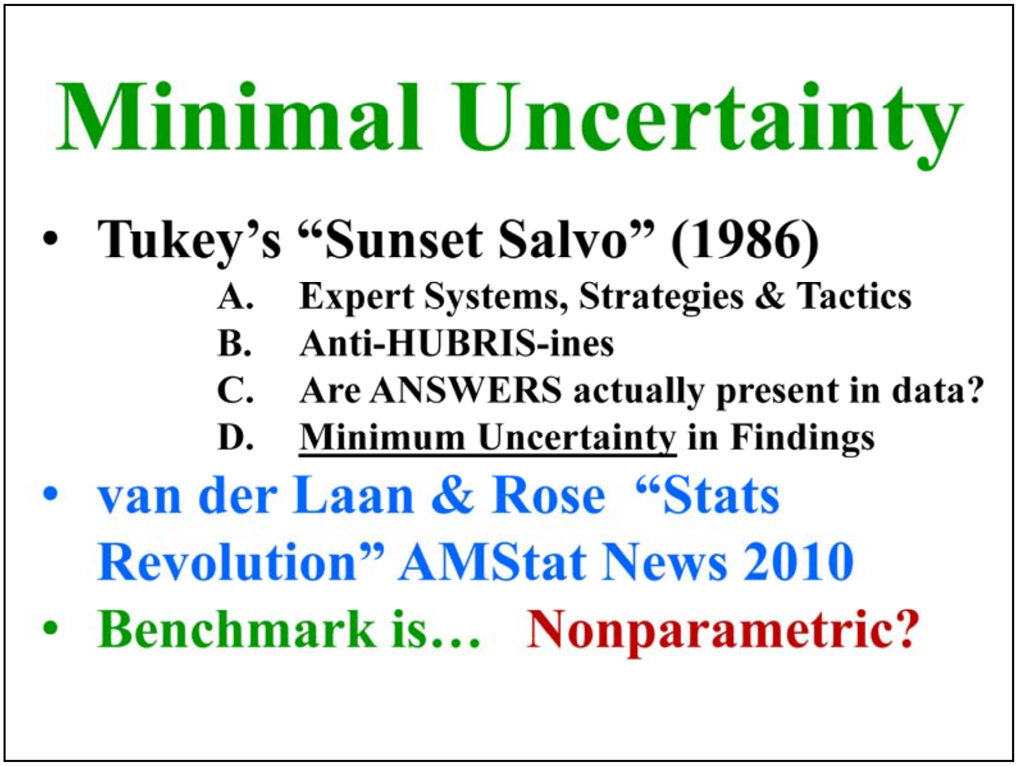
Simple Nested ANOVA

(Treatment within Subgroups)

- Not an "Interpretable" Causal Model
  - results from Considerable Over-Fitting
  - but makes rather good Predictions !!!
- Provides KEY information on Minimal Uncertainty in Y and ΔY
- …and of the Minimal Impact of Unmeasured Confounders on ΔY

The "model" implied by Nested ANOVA depicts LTDs as being constant within hundreds (or even thousands) of patient X-space cluster …possibly visualized (following standardization of scales along X-space axes) as hyper-cubes or hyper-spheres.

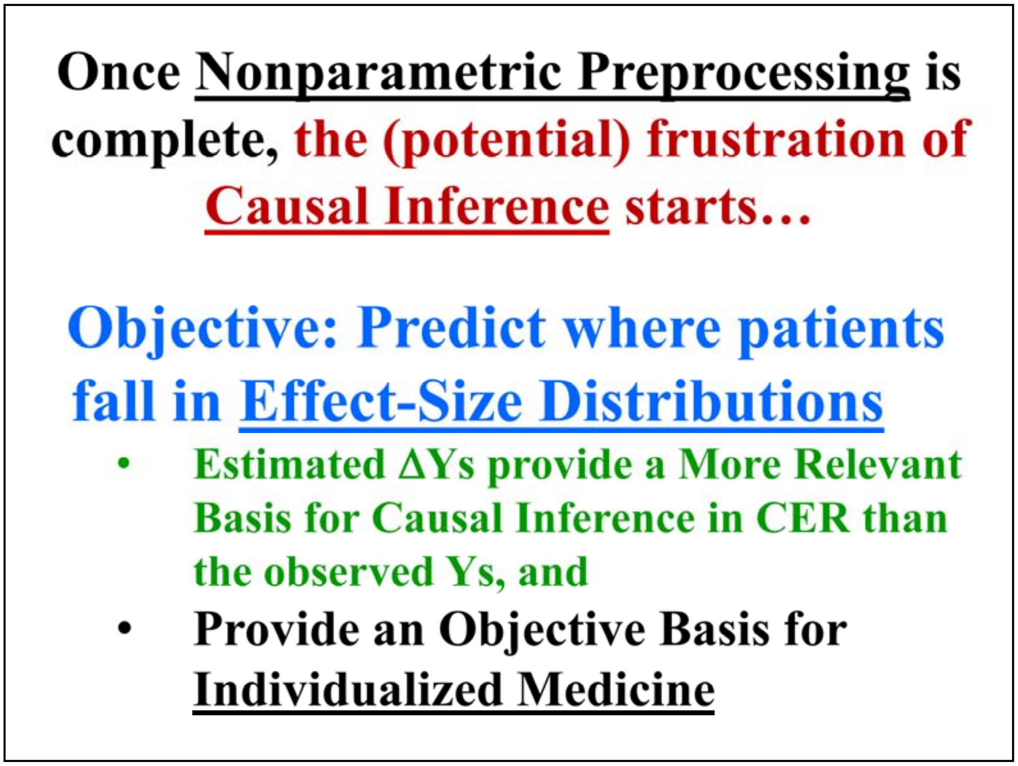
No "smoothing" of LTD predictions (using, say, results from "nearest neighbor" clusters) has been imposed. Some might think that this sort of tactic would increase interpretability (in terms of observed X-characteristics). But, due to unmeasured confounders, patients who "appear" to be nearest neighbors may actually be quite "distant."



Minimal Uncertainty

- Tukey’s “Sunset Salvo” (1986)
  - A. Expert Systems, Strategies & Tactics
  - B. Anti-HUBRIS-ines
  - C. Are ANSWERS actually present in data?
  - D. Minimum Uncertainty in Findings
- van der Laan & Rose “Stats Revolution” AMStat News 2010
- Benchmark is… Nonparametric?

EXPERT SYSTEMS are badly needed to implement either [1] very good strategies that are difficult to explain or else [2] strategies that are much too repetitive and tedious for statistical professionals to do with current packages.



Once **Nonparametric Preprocessing** is complete, the (potential) frustration of **Causal Inference** starts…

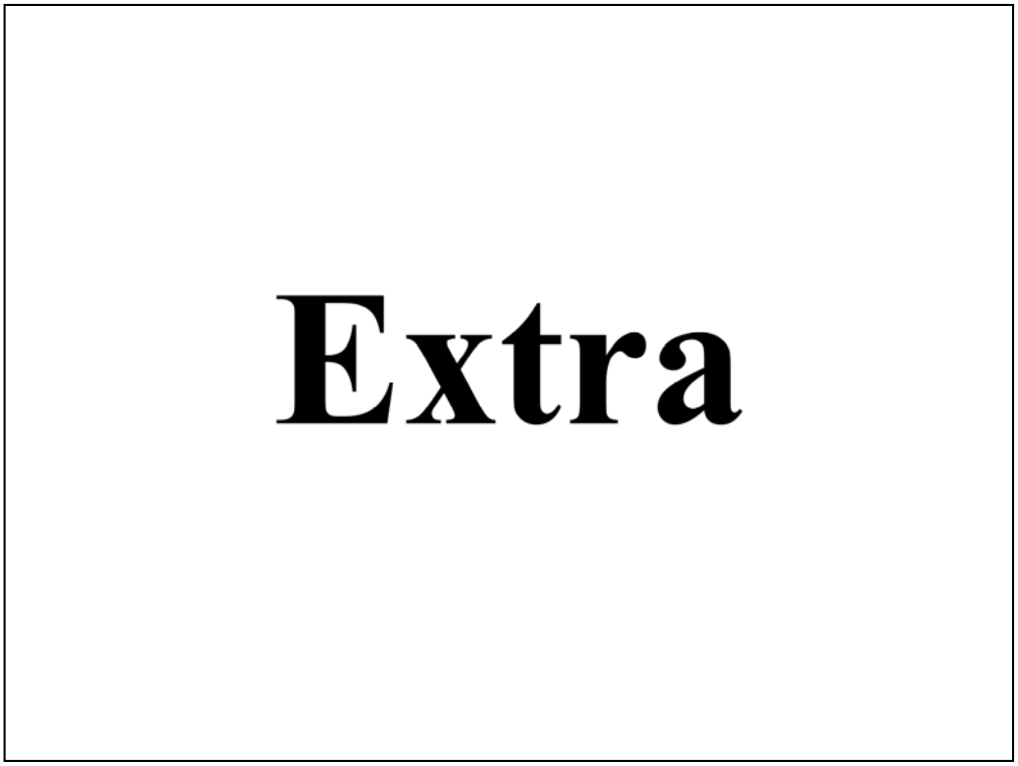
Objective: Predict where patients fall in **Effect-Size Distributions**

- Estimated ΔYs provide a More Relevant Basis for Causal Inference in CER than the observed Ys, and
- Provide an Objective Basis for **Individualized Medicine**

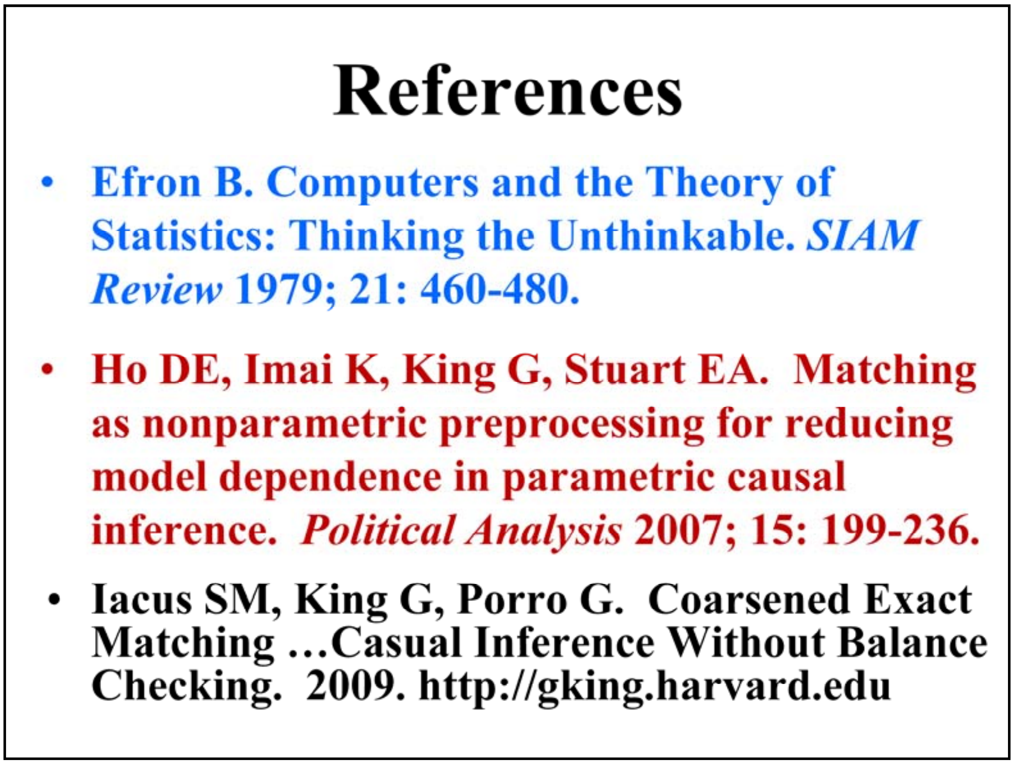
Following completion of “Systematic Sensitivity” analyses, health outcomes researchers have a relatively small number of alternative LTD distributions to take forward to attempt causal inference. Besides the “most typical” LTD distribution, he/she may also have (say) a least variance or most “peaked” distribution as well as the most variable and platokurtic distribution. Or “extreme” LTD distributions may be most skewed …either positive (favoring treatment A) or negative (favoring treatment B.)

With the “Nonparametric Preprocessing” gloves now potentially removed, how would you now analyze this “more relevant” information on patient level counterfactual differences ???

It’s a “brave new world” for young statisticians to develop “expert systems” that implement true OBJECTIVITY along the lines initially hinted at by Tukey (1986) and championed by van der Laan and Rose (2010).



Extra



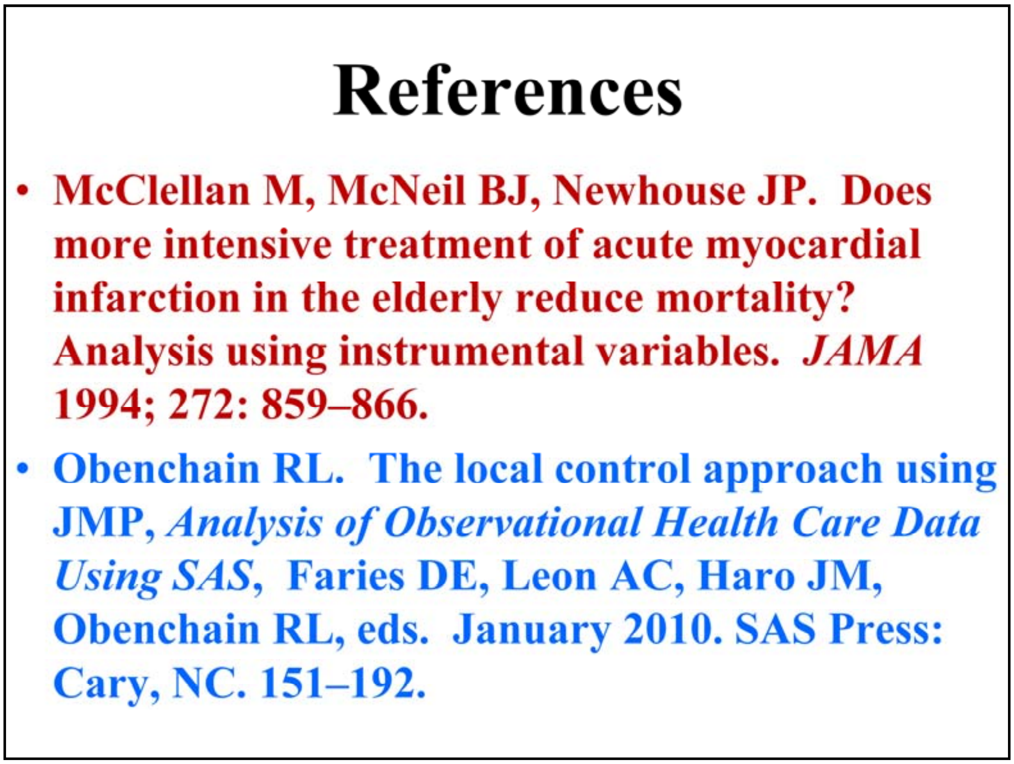
References

- Efron B. Computers and the Theory of Statistics: Thinking the Unthinkable. *SIAM Review* 1979; 21: 460-480.
- Ho DE, Imai K, King G, Stuart EA. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis* 2007; 15: 199-236.
- Iacus SM, King G, Porro G. Coarsened Exact Matching …Casual Inference Without Balance Checking. 2009. http://gking.harvard.edu

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

Iacus SM, King G, Porro G. cem: Coarsened Exact Matching. ***Journal of Statistical Software*** 2009; 30(9): 1–27.

Obenchain RL, Young SS. ***Advancing Statistical Thinking in Health Care Research.*** 2012 (under review)

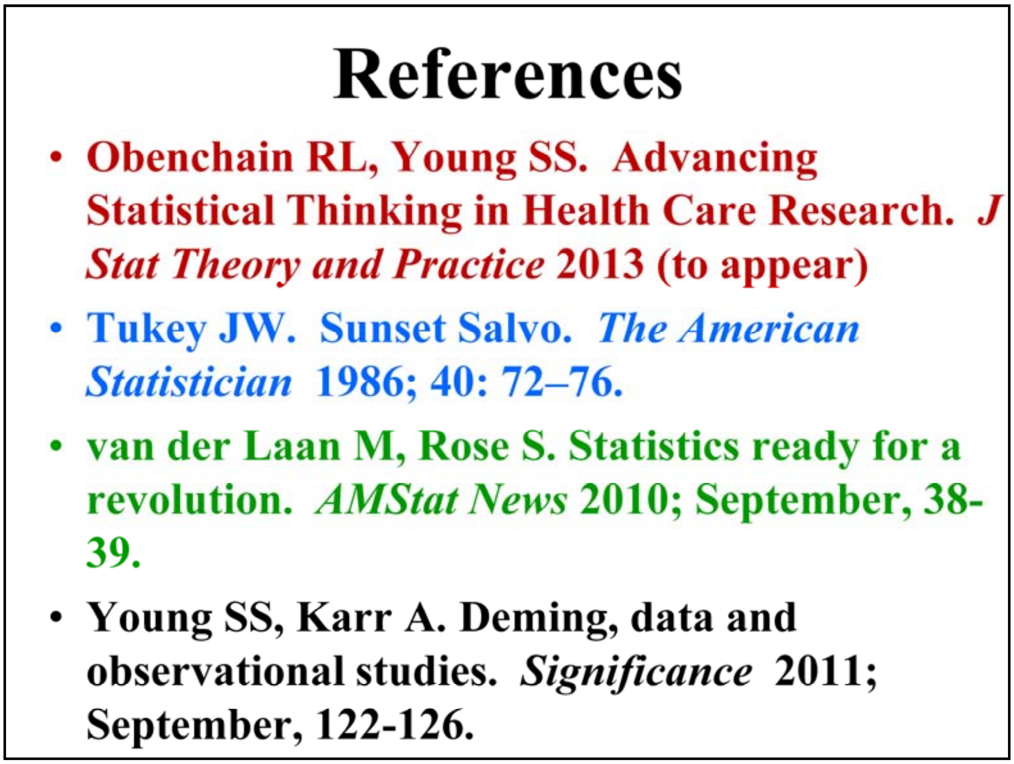


References

- 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.
- Obenchain RL. The local control approach using JMP, *Analysis of Observational Health Care Data Using SAS*, Faries DE, Leon AC, Haro JM, Obenchain RL, eds. January 2010. SAS Press: Cary, NC. 151–192.

Obenchain, RL, Hong, Q, Zagar, A, Faries, DE, 2011 Observational data analysis: MSE loss comparisons of local control versus parametric models. (under review)

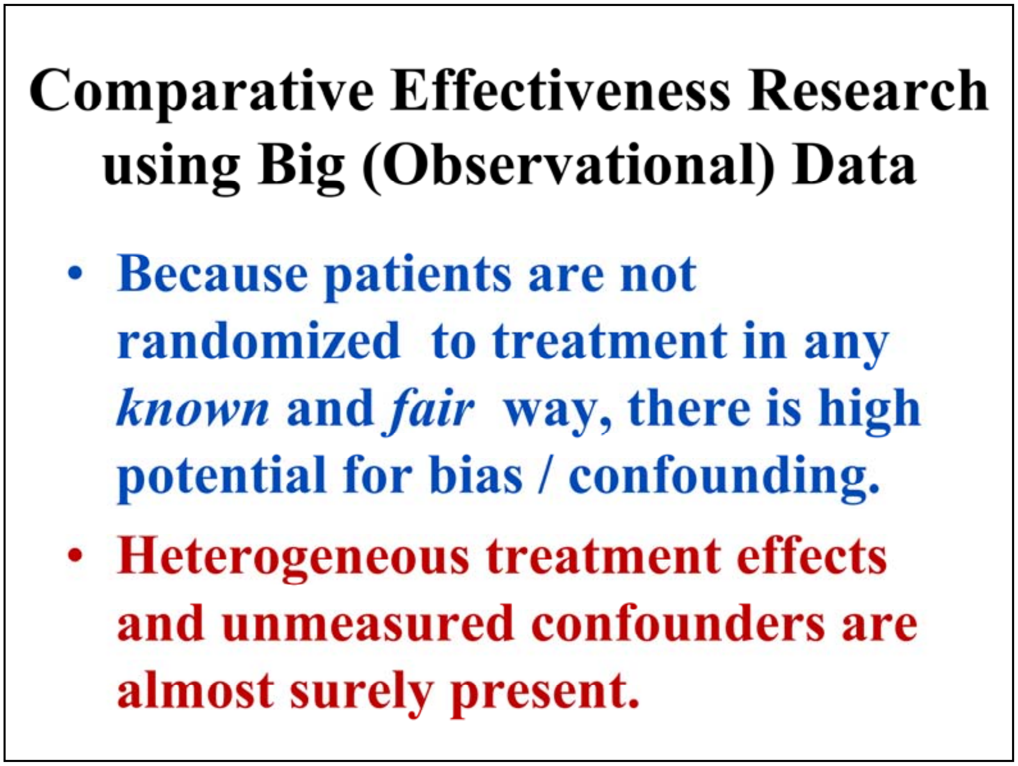
Rosenbaum PR, Rubin RB. The Central Role of the Propensity Score in Observational Studies for Causal Effects. ***Biometrika*** 1983; 70: 41-55.



References

- Obenchain RL, Young SS. Advancing Statistical Thinking in Health Care Research. *J Stat Theory and Practice* 2013 (to appear)
- Tukey JW. Sunset Salvo. *The American Statistician* 1986; 40: 72–76.
- van der Laan M, Rose S. Statistics ready for a revolution. *AMStat News* 2010; September, 38-39.
- Young SS, Karr A. Deming, data and observational studies. *Significance* 2011; September, 122-126.

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

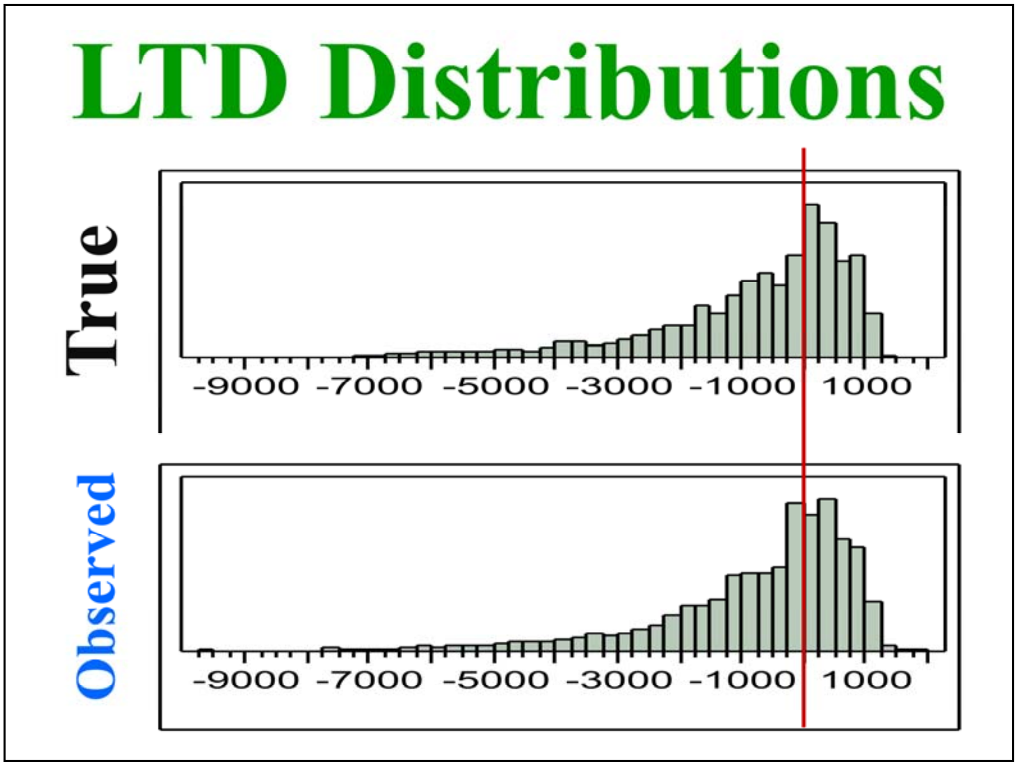


Comparative Effectiveness Research using Big (Observational) Data

- Because patients are not randomized to treatment in any *known* and *fair* way, there is high potential for bias / confounding.
- Heterogeneous treatment effects and unmeasured confounders are almost surely present.

Although information from observational studies is rather **clearly the data most relevant to current, actual health care practice**, it is subject to many forms of bias - ranging from treatment and model selection to heterogeneity in patient response and hidden / unmeasured confounders.

Although usually observational, **patient registry studies** can be an exception when “extra information” (perhaps even clinical evaluations) gets collected on each patient. For example, the Lilly “SCAP” study and the NIMH “CATIE” study of schizophrenia.



LTD Distributions

Of course, the data analyst usually cannot “see” the TRUE distribution of LTDs!!!

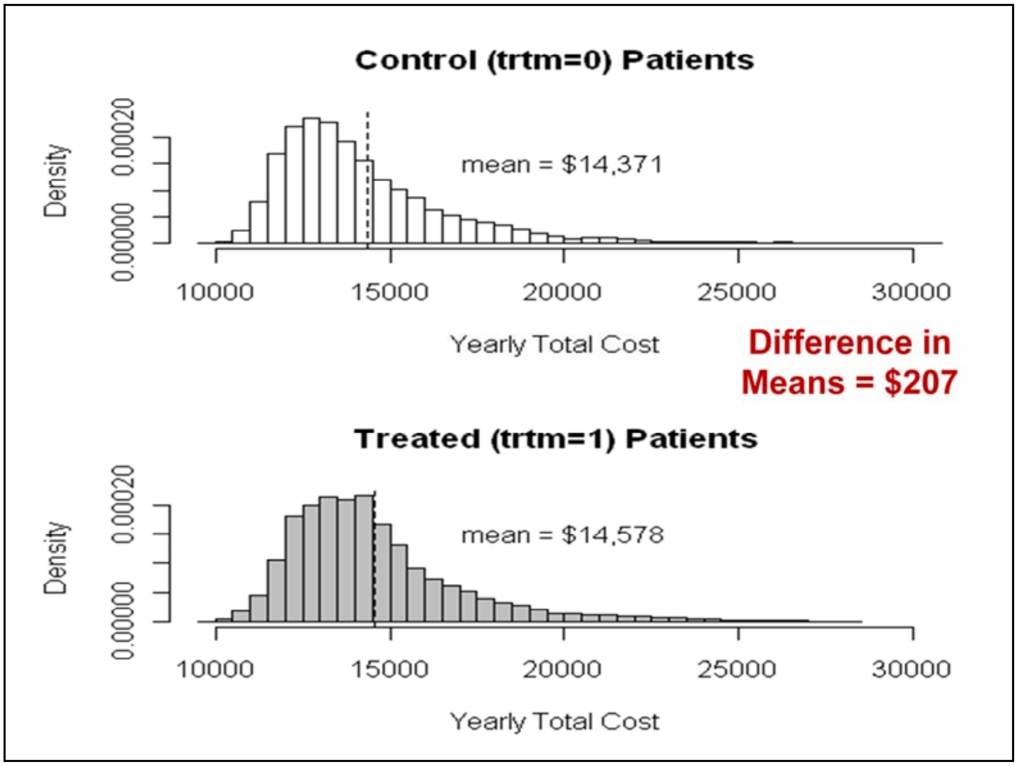
Correlation between True and Observed LTDs (from 1901 informative clusters out of 2K) is +0.982 here.

Thus the corresponding **R-square goodness-of-fit is .964**

TECKNICAL BACKGROUND INFO: True Propensity for trtm = 1 was assigned using 300 patient subgroups arranged so that the correlation between propensity score and the cost “signal” predictable from X (expected cost on trtm = 0) is +0.803. As a result, the TRUE distribution of LTDs has a logical (but somewhat complex) “explanation” in terms of patient X-characteristics. For example, trtm = 0 can be considered to be the inexpensive (generic) first-line med for less-intensive MDD therapy. Patients selected for intensive MDD therapy, receive a more expensive (branded) second- or third-tier med (possibly augmented with psychotherapy) and, as a result, their “other” health care costs decrease sharply (i.e. more than offset the increased cost of effective MDD treatment.)

Additive white noise has somewhat “smoothed” and “smeared” the observed LTD distribution, but its skewness is still helpful in more accurately estimating the Main-Effect of treatment (overall “main effect” is $650 in savings with trtm = 1) and in estimating the proportion of patients experiencing lower cost on trtm = 1 than on trtm = 0 (true % = 58, observed % = 59.)

When models using patient X-factors fail to reliably predict Observed LTDs (e.g. low R-squares), the data analyst should become introspective. He/she should then

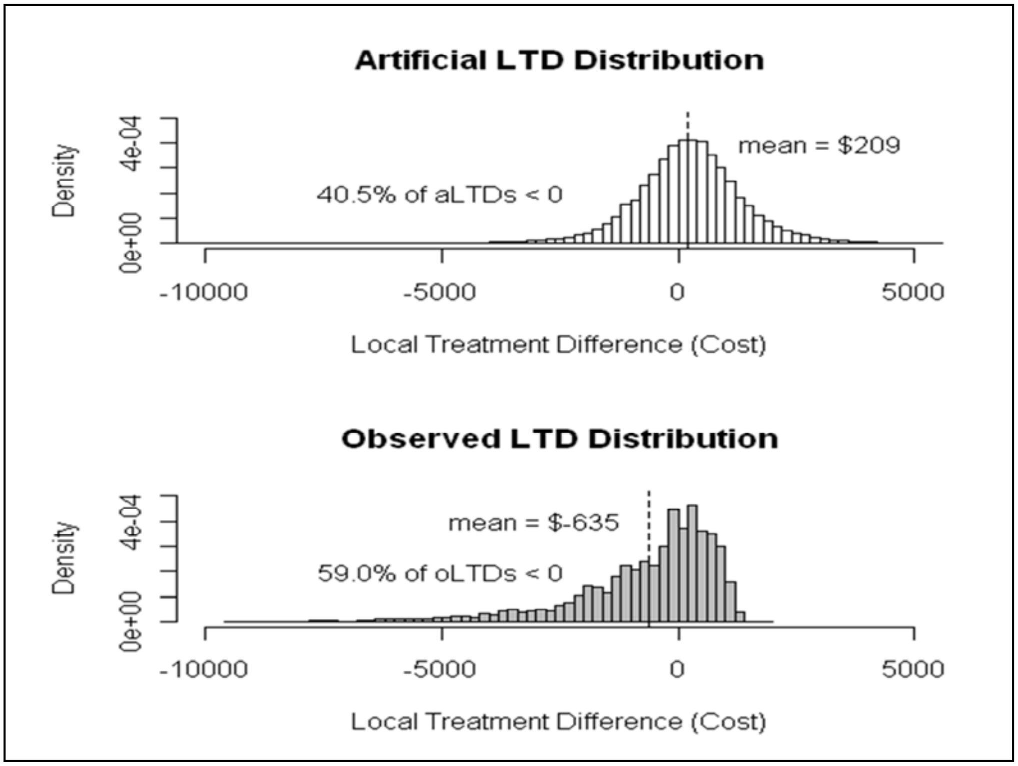


This pair of histograms would not strike most people as being clearly different.

The top distribution displays the *wyrcost* variable for 22,027 patients who chose *trtm*=0, while the corresponding distribution for 17,973 patients who chose *trtm*=1 is on the bottom.

While the difference in mean *wyrcost* is only $207, a conventional t-test nevertheless tags this difference as “highly significant” (p < 0.0001.)

However, note that **no consideration is given here for any differences in patient pre-treatment X-characteristics**.

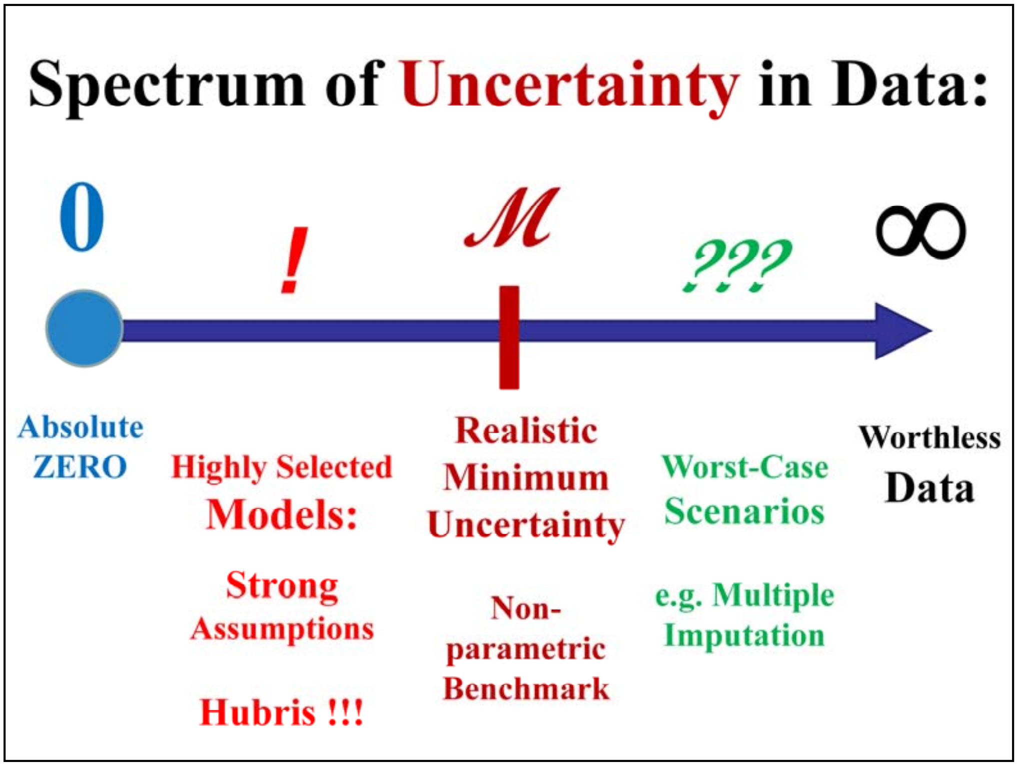


LC “Confirm” Step: Establish that the LTD Distribution is “Salient”

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

Both distributions displayed here are based upon dividing up 40K patients into 2K subgroups, of the same sizes and with the same fixed fractions of trtm=1 patients. As a result, exactly 1,901 subgroups will always contain both trtm=1 and trtm=0 patients …and thus be informative about counter-factual treatment differences.

The top “artificial” (completely random) distribtion can be computed with arbitrary precision via replication. Ten complete replications were used here.



Spectrum of Uncertainty in Data:

0

!

M

???

∞

Absolute ZERO

Highly Selected Models:

Strong Assumptions

Hubris !!!

Realistic Minimum Uncertainty

Non-parametric Benchmark

Worst-Case Scenarios

e.g. Multiple Imputation

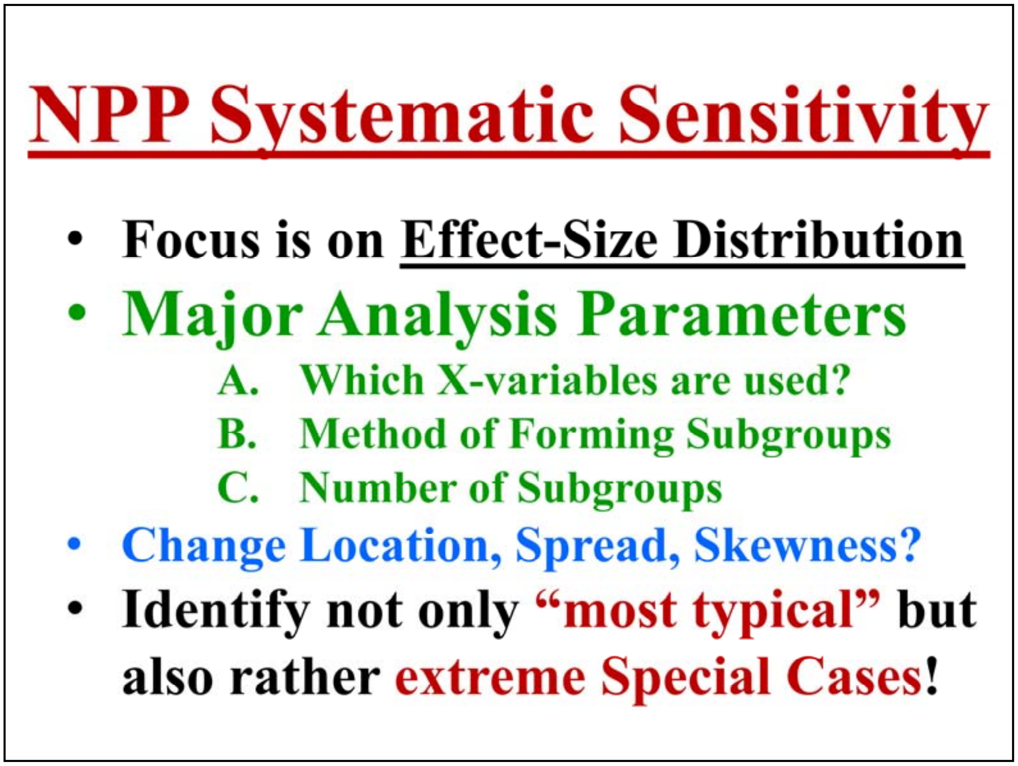
Worthless Data

Absolute ZERO: No statisticians live here !!! After all, **STATISTICS is never having to say you're CERTAIN.**

Hubris is pride considered punishable by the Greek Gods. As in: I'm the best @*#^ing stats/epi analyst around, I have done "my thing" to these data, and I say THE answer is Xxxx Yyyy Zzzz !!!

In his "Sunset Salvo" (American Statistician 1986), Tukey talks about ANTIHUBRISINES, minimum uncertainty, "pigeon-hole models" (possibly including nested ANOVA, treatment within patient subgroup), and the frustrating reality that most data sets don't contain an answer …let alone THE answer!

Emphasis on "Nonparametric Pre-Processing" (NPP) of observational data (via patient "matched sets") is being spear-headed in the social sciences by Gary King & Elizabeth Stuart of Harvard & J. Hopkins. Recent Obenchain work on Local Control takes NPP to a higher level …where NPP itself provides estimates of heterogeneous (local) effect-sizes and of their minimum uncertainty under changes of tactics for forming patient subgroups. And the Obenchain-Faries simulation studies show that NPP can produce more accurate estimates (smaller root MSE loss) than traditional parametric models that are "wrong" due to unmeasured confounders.



NPP Systematic Sensitivity

- Focus is on Effect-Size Distribution
- Major Analysis Parameters
  - A. Which X-variables are used?
  - B. Method of Forming Subgroups
  - C. Number of Subgroups
- Change Location, Spread, Skewness?
- Identify not only “most typical” but also rather extreme Special Cases!

When one’s observational data analysis strategy is based upon formation of many, relatively homogeneous subgroups of patients, using Systematic Sensitivity calculations to explore the stability of the resulting LTD distributions under choice of analysis parameter settings actually becomes relatively straight-forward.

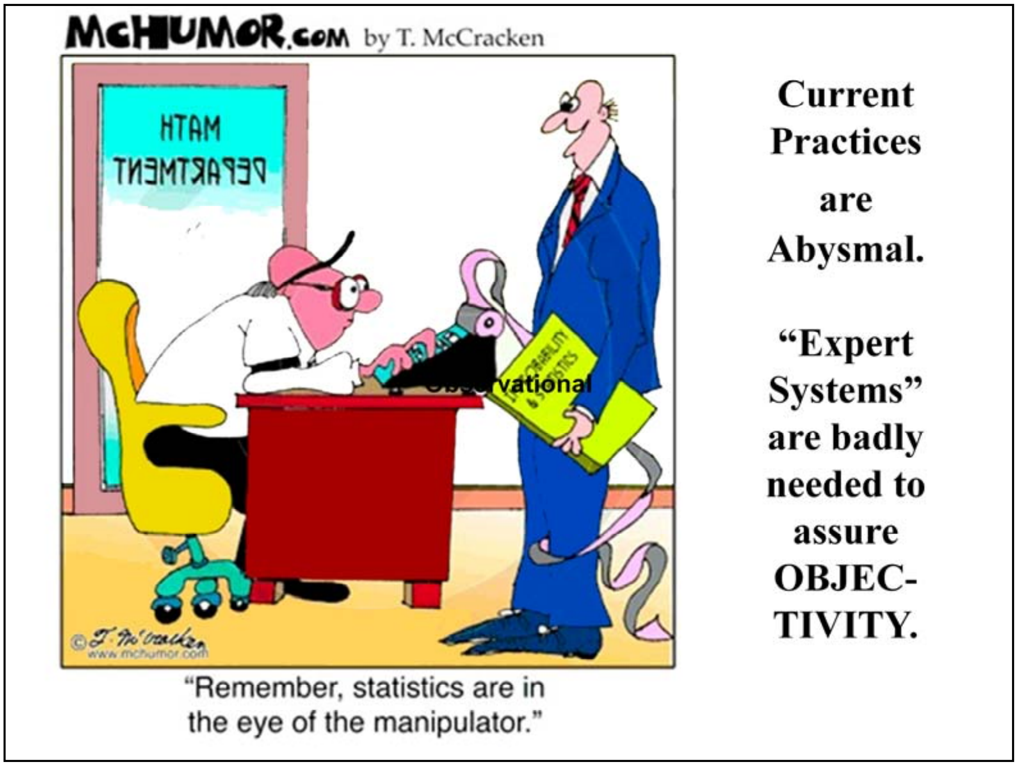
For example, one does not have to worry about usual “model specification and identification assumptions.” Specifically, one’s model is always simple Nested ANOVA (Treatments within essentially non-parametrically defined BLOCKS.)

Patients “matched” on two or more X-variables are essentially also matched on all transformations (like squares) and functions (like all forms of interactions) of those variables. For example, in the case of exact X-space matches, this is the basis for model lack-of-fit calculations.

Matching or Clustering algorithms can be hierarchical or not as well as either computer-intensive or quick-and-dirty. Do these sorts of choices make any real difference?

Use of many subgroups will reduce bias …unless too much information is then lost due to “uninformative (pure) subgroups” (containing only treated or only control patients.) In fact, what is the “common support” of the alternative treatment cohorts in X-space? Using fewer subgroups reduces variability and tends to impose some smoothing, so variance-bias trade-offs are involved here!!!

These Systematic Sensitivity calculations are simple enough to do automatically …using computerized learning algorithms. In fact, it may prove to be ESSENTIAL that these calculations are done COMPLETELY OBJECTIVELY, so automatic may



"Remember, statistics are in the eye of the manipulator."

Current Practices are Abysmal.

"Expert Systems" are badly needed to assure OBJECTIVITY.

Major stakeholders in the current rebate on comparative effectiveness appear to embrace at least seven distinct and sometimes conflicting perspectives. These diverse perspectives include those of (i) patients and their families, (ii) health care providers, (iii) health care payers (observational data owners), (iv) government funding agencies, (v) health care regulators / policy makers, (vi) academics and consultants seeking income and/or professional recognition and (vii) the pharmaceutical and device manufacturing industry. While all seven of these perspectives claim to support exchange of scientific (objective) information, **each may sponsor only analyses tailored to their unique perspective**. When the corresponding (de-identified) analytical files created from observational data are not also released, the magnitude of any induced bias and ignored confounding remains unknown.

Researchers may validly claim that patient specific and/or proprietary information should not be shared …let alone make public. However, to advance the science of OCER, there is no valid reason for authors to not provide both summary statistics for and graphical visualizations of the typical and extreme LTD distributions that quantify the purely objective uncertainty about the implications of their data …before any subjective injection of Bayesian prior or "additional" information / opinion. **In fact, editors of professional journals and health policy makers should DEMAND provision of this evidence of objectivity and credibility!**