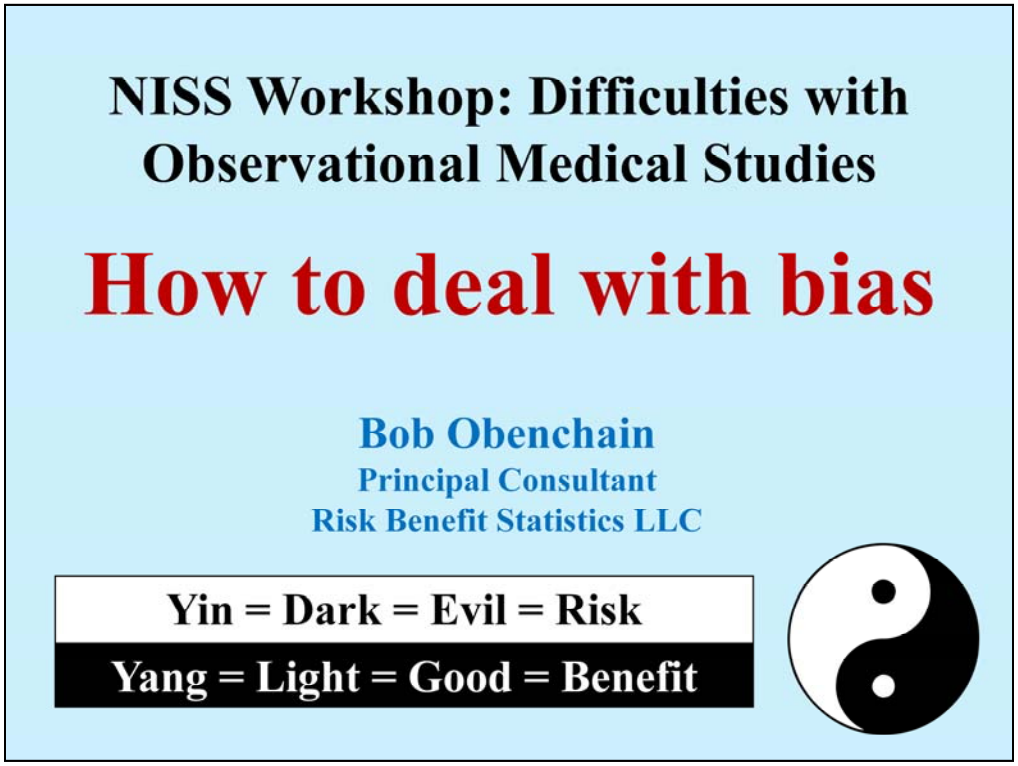


NISS Workshop: Difficulties with Observational Medical Studies

How to deal with bias

Bob Obenchain
Principal Consultant
Risk Benefit Statistics LLC

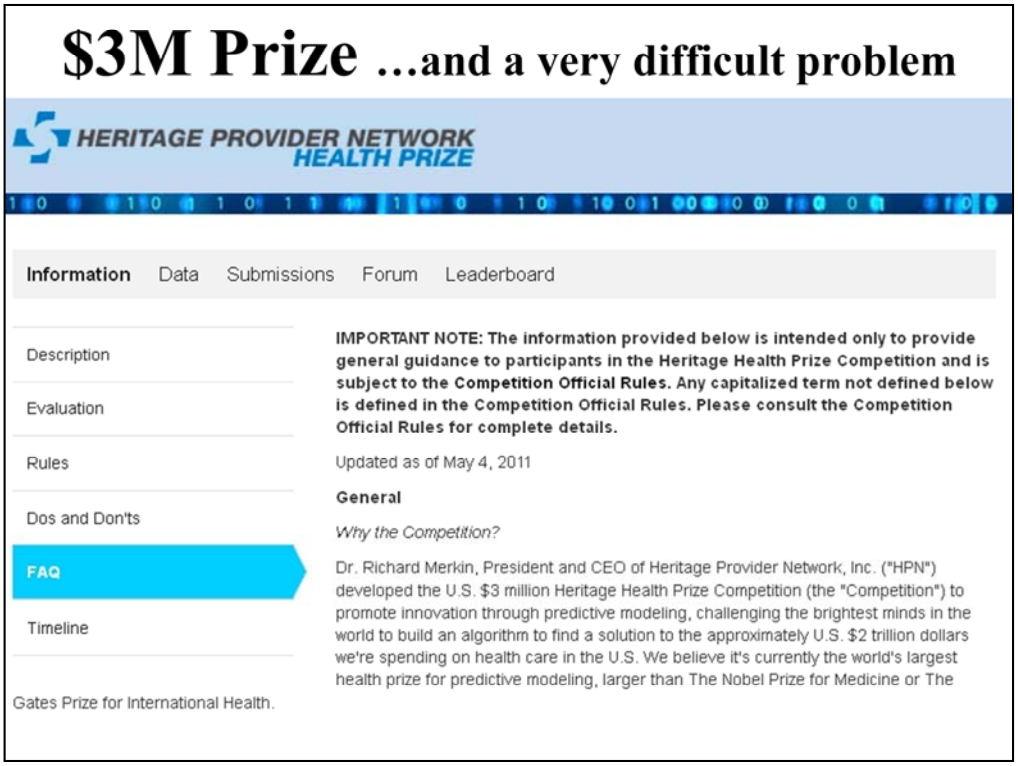
Yin = Dark = Evil = Risk

Yang = Light = Good = Benefit

This talk will propose **Three Practical** (but somewhat painful) ways to address the pervasive problem of BIAS in observational studies.

While I will be discussing some rather technical issues and referencing some technical jargon, I will try to keep the discussion general and avoid technical fine-points / details.

Another possible take-away from this session is that health services researchers should consider using data simulation techniques to develop much more realistic assessments of the reliability of findings from alternative methodologies for analysis of observational data. These more-realistic insights are badly needed when evaluating health outcomes research and formulating health policy. Here I will use a simulated dataset to illustrate the frailty of Instrumental Variables methods.



$3M Prize …and a very difficult problem

IMPORTANT NOTE: The information provided below is intended only to provide general guidance to participants in the Heritage Health Prize Competition and is subject to the Competition Official Rules. Any capitalized term not defined below is defined in the Competition Official Rules. Please consult the Competition Official Rules for complete details.

Updated as of May 4, 2011

General

Why the Competition?

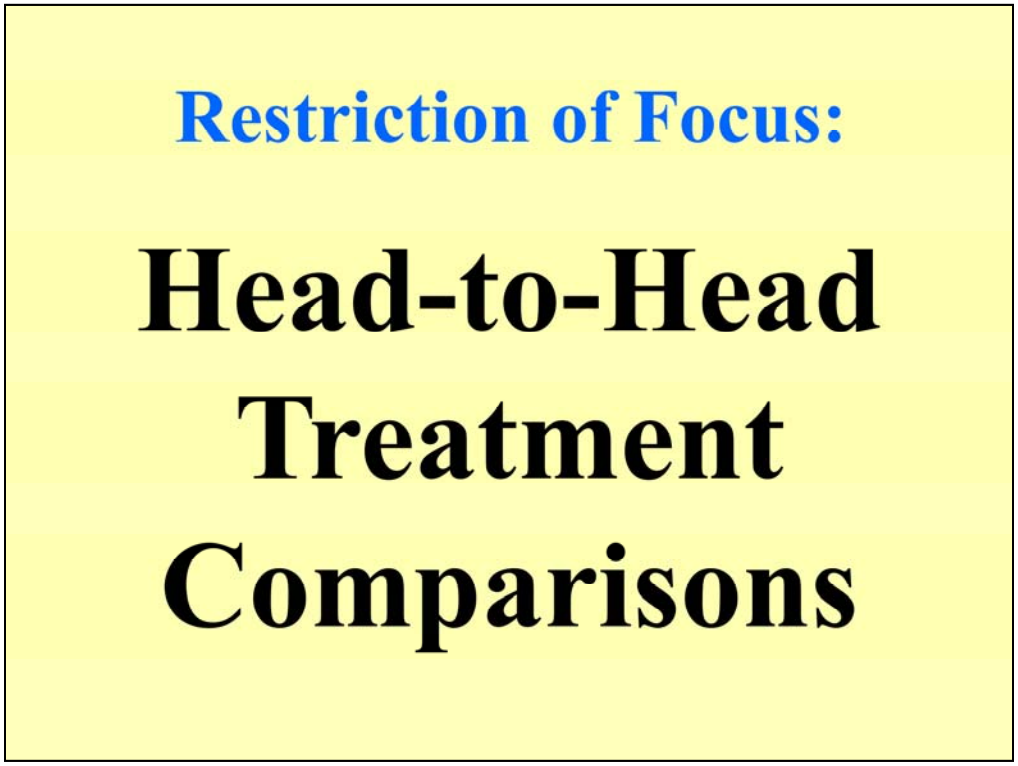
Dr. Richard Merkin, President and CEO of Heritage Provider Network, Inc. ("HPN") developed the U.S. $3 million Heritage Health Prize Competition (the "Competition") to promote innovation through predictive modeling, challenging the brightest minds in the world to build an algorithm to find a solution to the approximately U.S. $2 trillion dollars we're spending on health care in the U.S. We believe it's currently the world's largest health prize for predictive modeling, larger than The Nobel Prize for Medicine or The Gates Prize for International Health.

The HPN Competition focuses on prediction of hospital days for a wide variety of patients with multiple co-morbid conditions over successive years.

Notes for Slide 3 start on the next page. Some final notes for Slide 3 are here due to lack-of-space there.

For head-to-head comparisons, randomization is a reasonable “study design” tool only when patients are relatively homogeneous (interchangeable) …or have been pre-classified into distinct strata, each with this same property.

When patients are heterogeneous, observational data for head-to-head comparisons can be just as good (or better than) that from RCTs …as long as the same **great care** is exercised in collecting data on relevant patient baseline characteristics and resulting outcomes. Patient stratification (“local” analysis) then becomes a key **analysis tool** rather than an initial **design tool**.



Restriction of Focus:

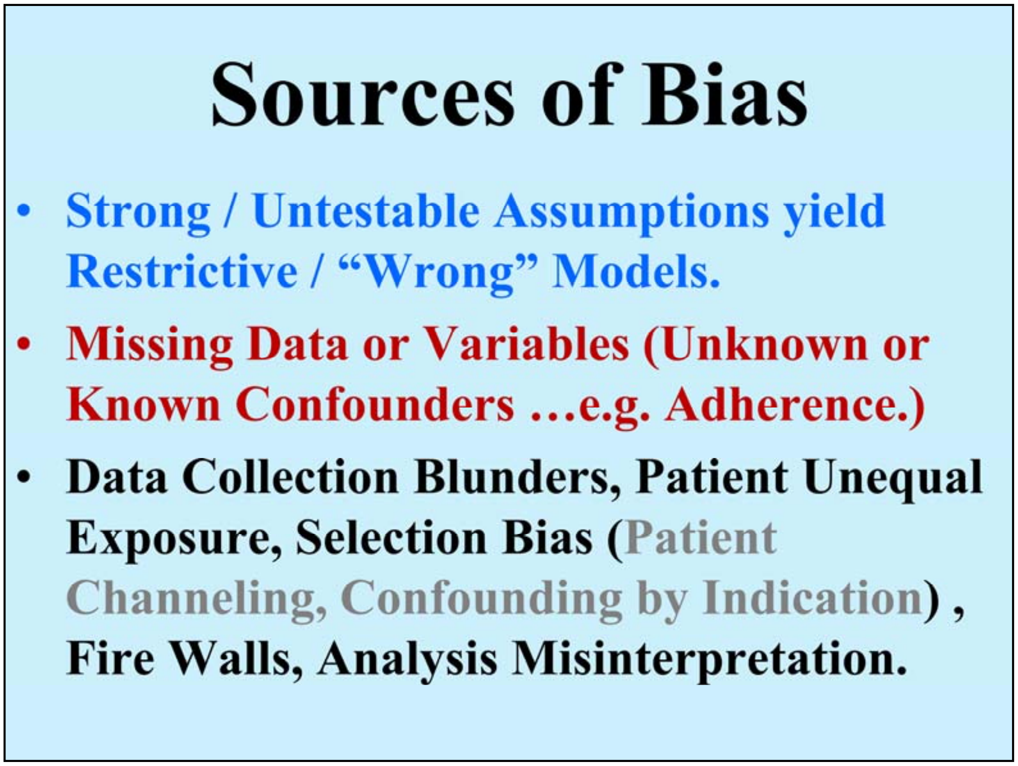
Head-to-Head Treatment Comparisons

All patients who qualify for study clearly have something in common: the disease or condition for which both treatments are prescribed. These studies can be done within **patient registries** where information on clinical outcomes is also gathered.

When health care **claims data** are used, researchers attempt to reconstruct "episodes of care" and form rectangular "analytical files" with rows representing patients and columns containing variables that describe those disease episodes. Such claims-based studies may be viewed as being cross-sectional rather than longitudinal …but some explanatory X-variables may represent long-term summaries of treatment choices and/or patient responses.

Most importantly, because patients were not randomly / "fairly" assigned to treatment, it is essential for patient X-characteristics measuring disease severity and/or patient frailty to be present to **"recognize and adjust for" pre-treatment imbalances** …i.e. to make fair comparisons.

The OMOP initiative did not focus on studies of this relatively simple type. The main observational research "safety monitoring" issues that OMOP did address are <u>much more complex and difficult than this</u>. It is quite unfortunate if health policy makers believe that OMOP's findings about drug safety / surveillance methods apply to **all types** of observational studies.



Sources of Bias

- Strong / Untestable Assumptions yield Restrictive / “Wrong” Models.
- Missing Data or Variables (Unknown or Known Confounders …e.g. Adherence.)
- Data Collection Blunders, Patient Unequal Exposure, Selection Bias (Patient Channeling, Confounding by Indication) , Fire Walls, Analysis Misinterpretation.

BIAS: Nonrandom (long-term) difference between an estimate and the true value of the target parameter. Also termed Systematic Error or Invalidity of a Statistical Method (e.g. step-wise regression selection.)

Global, parametric models for observational data have particularly questionable validity due to the wide variety of patient subtypes included in the analysis. Furthermore, experience has suggested that law-like-relationships are rarely seen among health care variables …i.e. much lower model R-square statistics.

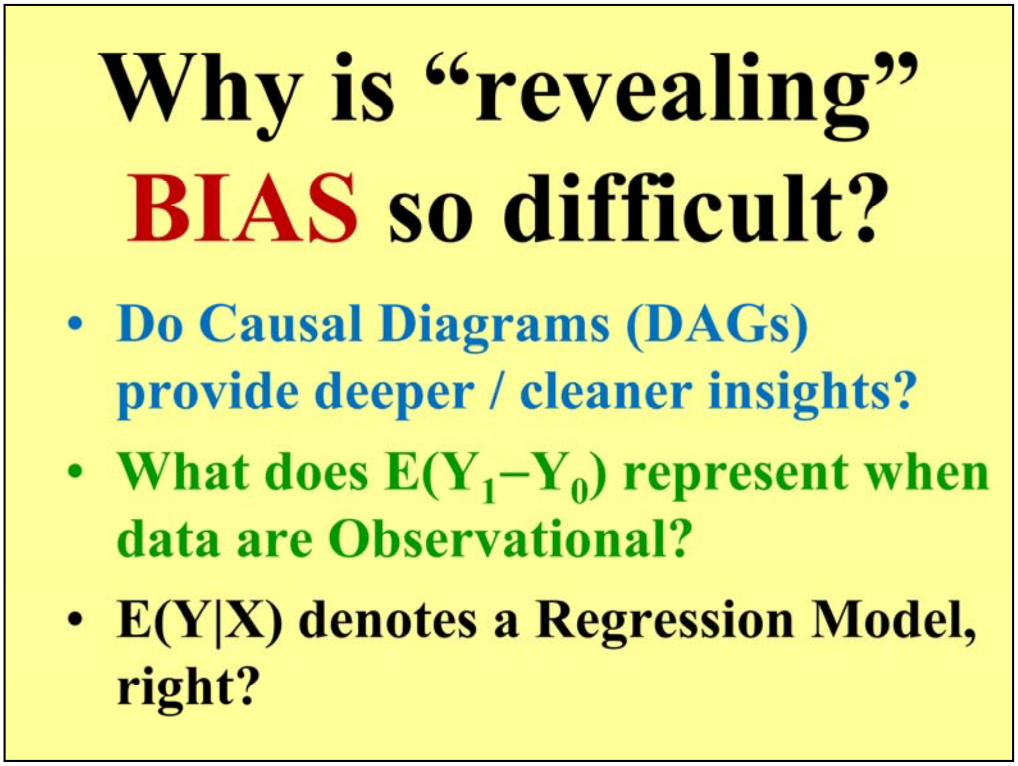
Example of an Untestable Assumption: Instrumental Variables ==> all effects are attributable to treatment.

Missing Variables ==> Genetic, Metabolic, Physiological, Sociological, Behavioral, Risk Perception

Confounding ==> Unequal Exposure of Patients (Comorbid Conditions and their Treatments, Competing Risks)

Selection ==> Unequal Sampling of Patient Subtypes

Given presence of all these biases, valid statistical analyses are, almost by definition, **sensitivity analyses**.



Why is “revealing” BIAS so difficult?

- Do Causal Diagrams (DAGs) provide deeper / cleaner insights?
- What does $E(Y_1 - Y_0)$ represent when data are Observational?
- $E(Y|X)$ denotes a Regression Model, right?

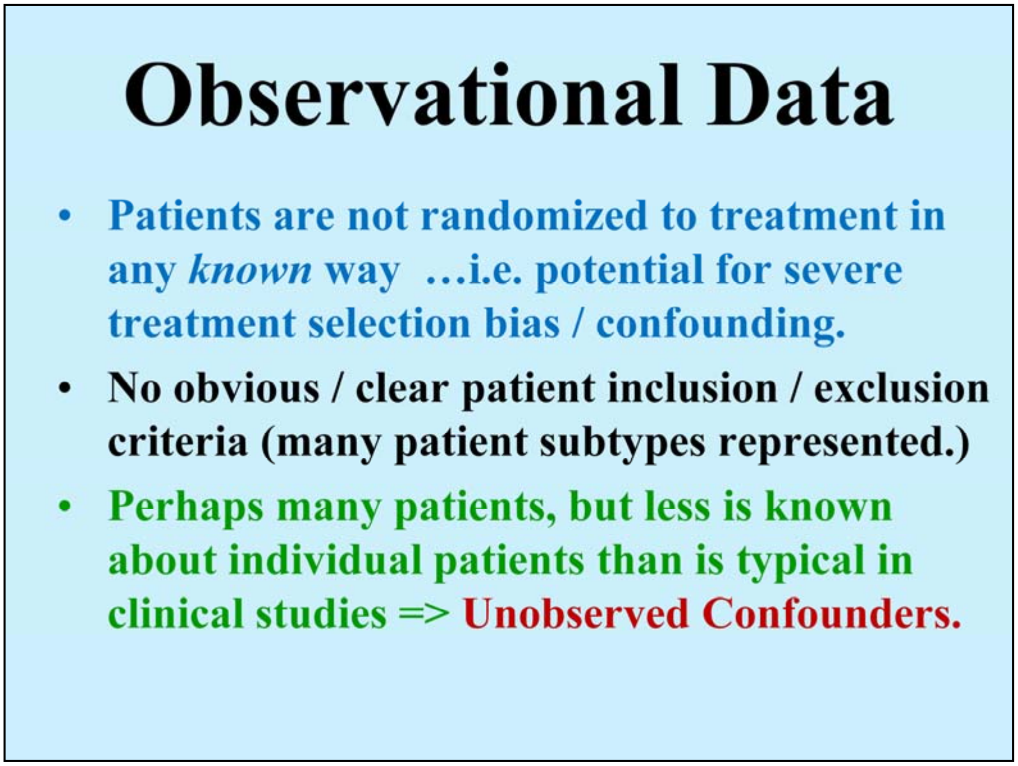
Directed Acyclic Graphs: no complete cycles allowed because **nothing can cause itself** …right? These pictures may be extremely helpful, but the terminology (from computer science, expert systems, robotics & artificial intelligence) is quite esoteric.

An Economist would say that the problem of estimating treatment effects for observational data is not “identified“ (is insoluble) and would feel compelled to make more-and-more assumptions until a solution becomes available.

Is an expected difference in outcomes over all patients taking one of the two treatments (union of patients) or only those patients with characteristics suggesting that they are willing to choose between them (intersection of patient populations, common support.)

The right-hand-side of a regression model equation is the “business end” where bias can be introduced via unrealistic assumptions. **Conditional inference based upon patient subgroups** can be done quite differently from traditional regression model fitting.

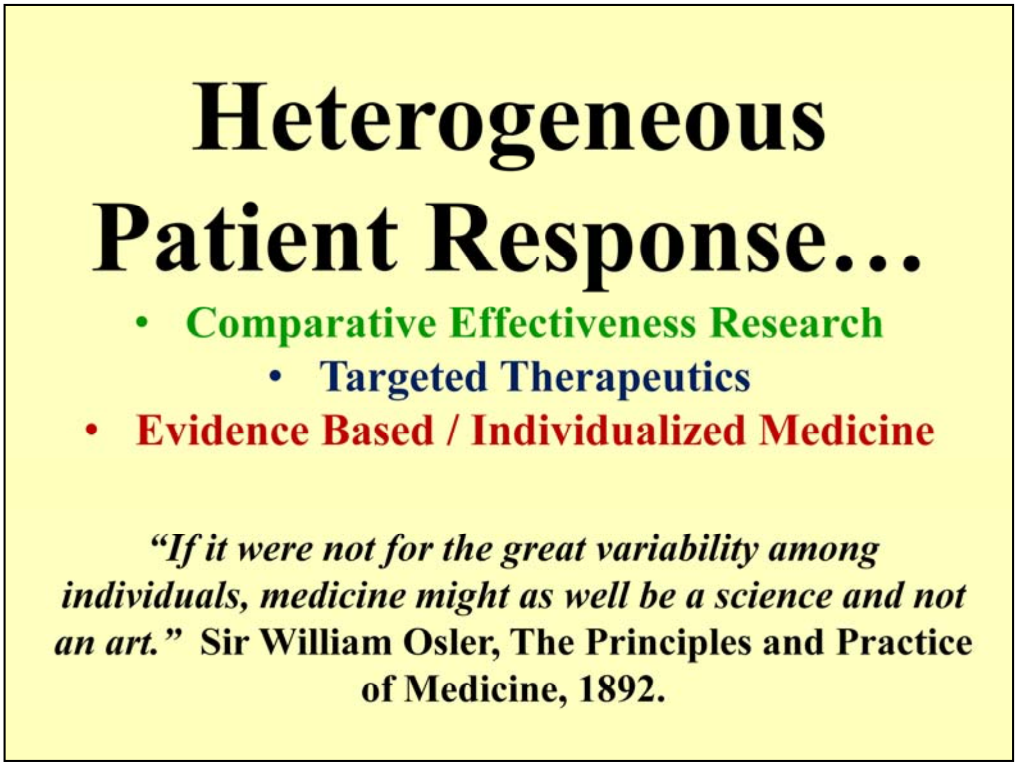
Most scientists don’t want to admit that they should be doing and reporting extensive **Sensitivity Analyses** to more realistically quantify the uncertainty in their findings.



Observational Data

- Patients are not randomized to treatment in any *known* way …i.e. potential for severe treatment selection bias / confounding.
- No obvious / clear patient inclusion / exclusion criteria (many patient subtypes represented.)
- Perhaps many patients, but less is known about individual patients than is typical in clinical studies => Unobserved Confounders.

So, here is the reality of Observational Data Analyses !!!



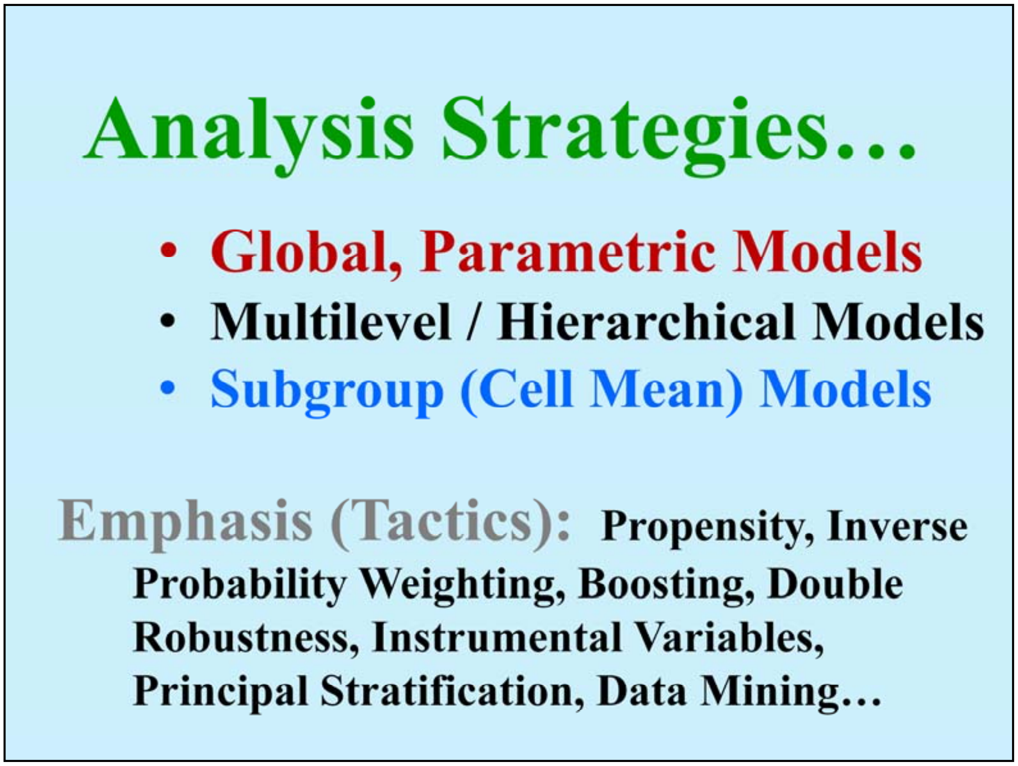
Heterogeneous Patient Response…

- Comparative Effectiveness Research
- Targeted Therapeutics
- Evidence Based / Individualized Medicine

"If it were not for the great variability among individuals, medicine might as well be a science and not an art." Sir William Osler, The Principles and Practice of Medicine, 1892.

Apparently, health outcomes researchers do not yet widely recognize how important an issue heterogeneity in patient response really is!!!

NOTE: Osler quote from Kaplan et al. (2010) "Who Can Respond to Treatment?" Medical Care • Volume 48, Number 6 Suppl 1, June …on **CER.**



Analysis Strategies…

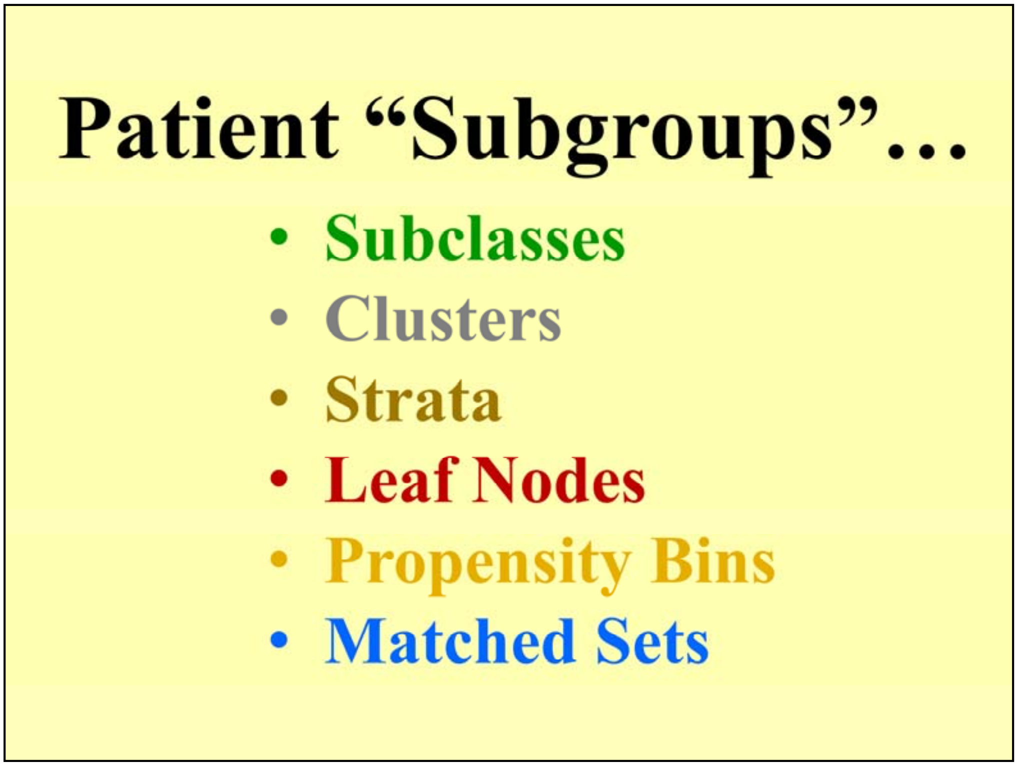
- Global, Parametric Models
- Multilevel / Hierarchical Models
- Subgroup (Cell Mean) Models

Emphasis (Tactics): **Propensity, Inverse Probability Weighting, Boosting, Double Robustness, Instrumental Variables, Principal Stratification, Data Mining…**

Global Parametric Models: Multivariable Regression Models (Covariate Adjustment), Heckman Selection Models (Inverse Mills Ratios), Simultaneous Equations Models (Causal Diagrams), …

Multilevel Models: Divide patients up in **pre-known ways** using their baseline X-characteristics.

Subgroup (Cell Mean) Models: Use all known patient X-characteristics (discrete or continuous) **only to determine which patients are most like which other patients**. Analyses within and across the resulting subgroups thus tend to be non-parametric, such as Nested ANOVAs. Such analyses tend to be robust in the narrow sense that they do not make any particularly strong or clearly unrealistic assumptions.



Patient “Subgroups”…

- Subclasses
- Clusters
- Strata
- Leaf Nodes
- Propensity Bins
- Matched Sets

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

Patents within a single subgroup are to either [1] have some common characteristic(s) or else [2] be as similar as possible.

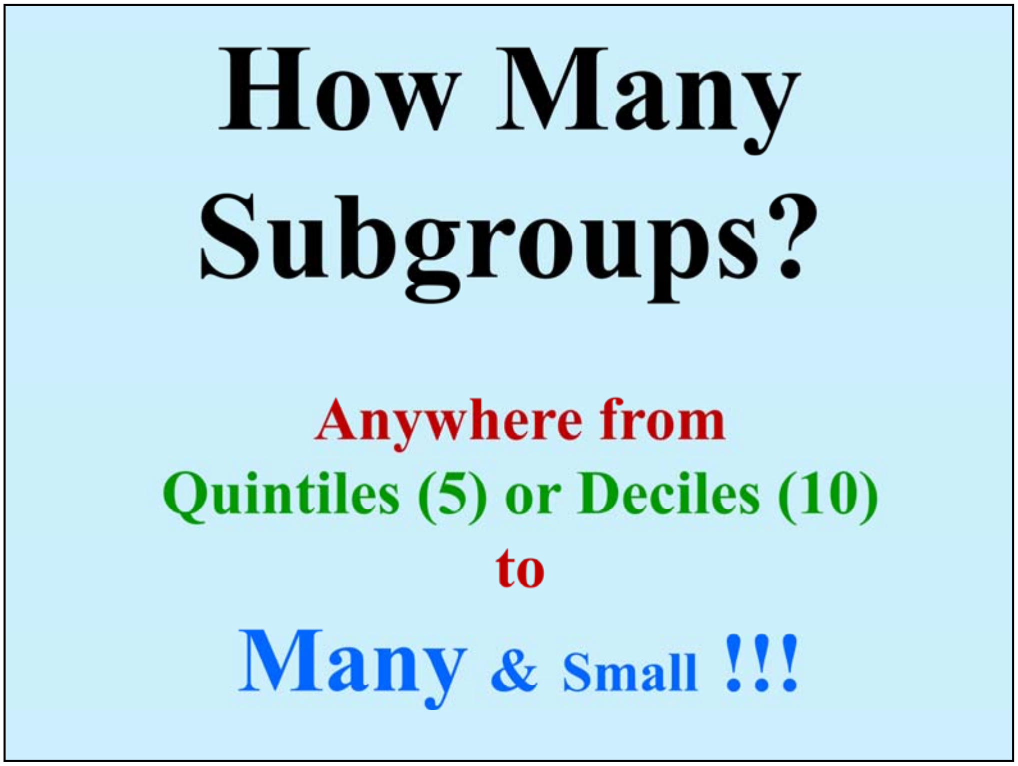
Patents in all other subgroups are to either [1] NOT have that/those characteristic(s) or else [2] be as dissimilar as possible from the patients in the given subgroup.

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

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.



How Many Subgroups?

Anywhere from

Quintiles (5) or Deciles (10)

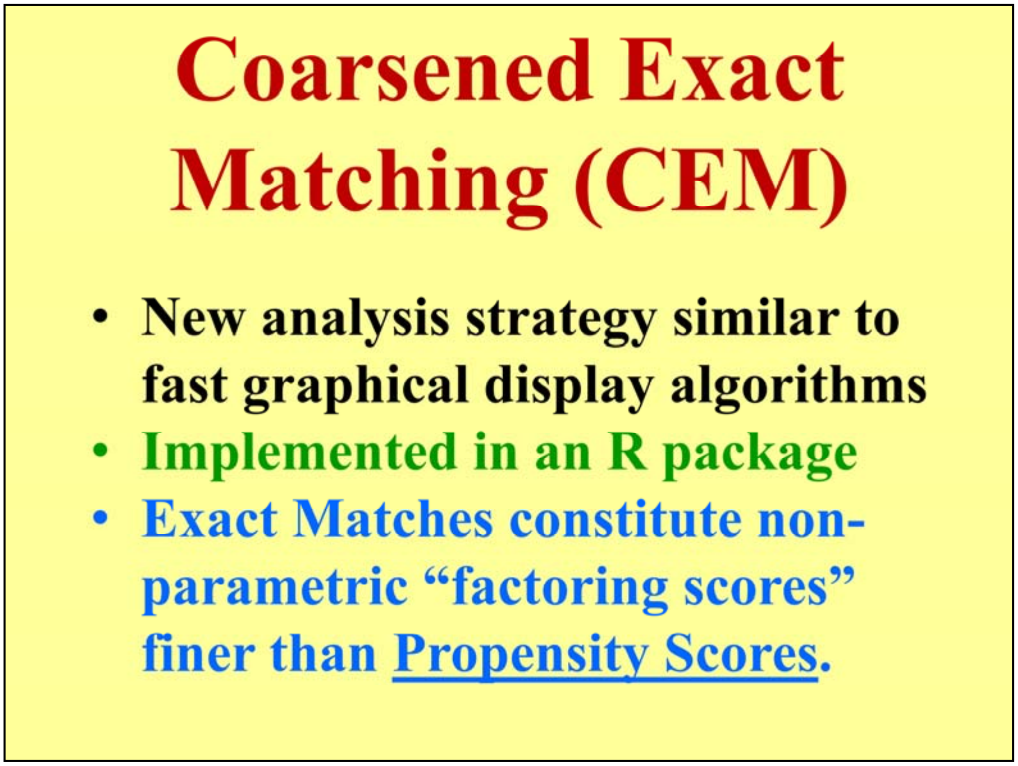
to

Many & Small !!!

50 years ago, nobody wanted to do the same sorts of calculations by hand more than just a very few times.

Today’s software, hardware, and giant health-care databases make it easy to consider using literally thousands of patient subgroups …especially when within-subgroup effects are viewed as **random** (rather than fixed.)

Again, this **parameter** that helps define how “local” analyses really are needs to be varied over a range determined by **sensitivity analysis.**



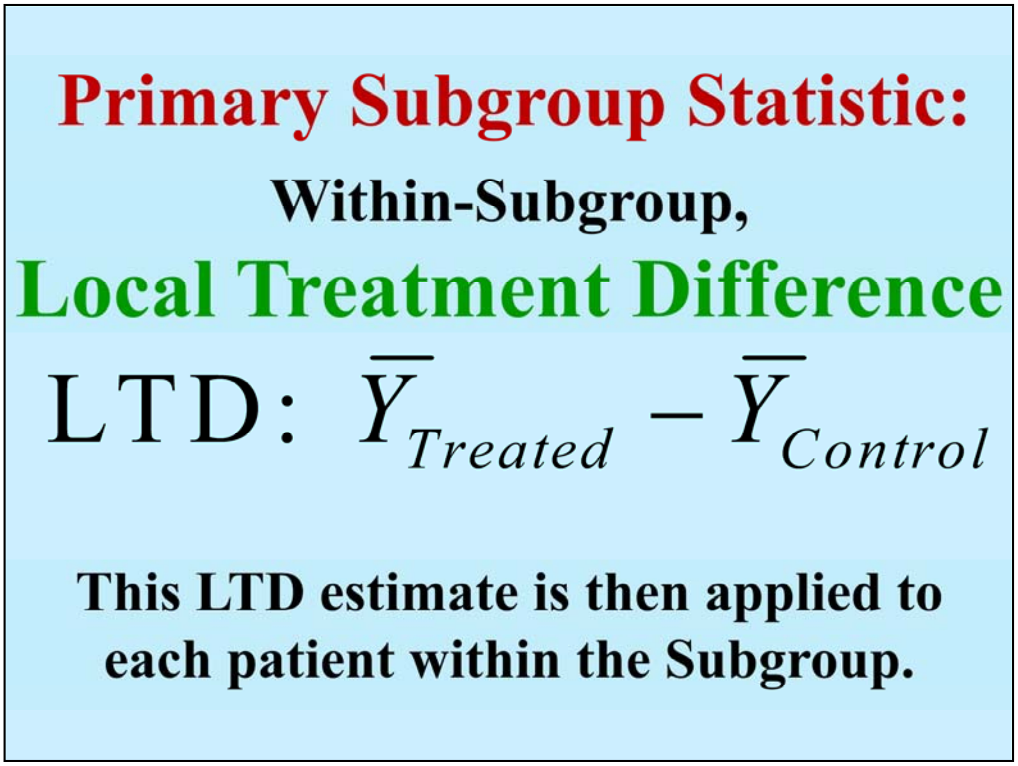
Coarsened Exact Matching (CEM)

- New analysis strategy similar to fast graphical display algorithms
- Implemented in an R package
- Exact Matches constitute non-parametric “factoring scores” finer than Propensity Scores.

Obviously, exact matches define the “upper limit” on Many (and Small) Clusters. When only a small proportion of patients can be exactly matched, “coarsening” becomes essential to avoid uninformative subgroups.

Algorithmically, **exact matches** are extremely easy to find in gigantic datasets. Conceptually, one essentially needs to sort the data on all X-variables so that exact matches will be adjacent to each other. In fact, SAS proc FASTCLUS yields exact matches in about 5 minutes on the numerical example used here (8 X-variables on ~250K patients) as long as the analyst **requests 40K or more clusters!**

NOTE: Although basic concepts quite similar to the CEM approach were used in the analyses discussed here, the R-functions implemented in the CEM package were not actually used / needed.



Primary Subgroup Statistic:

Within-Subgroup,

Local Treatment Difference

LTD: $\overline{Y}_{Treated} - \overline{Y}_{Control}$

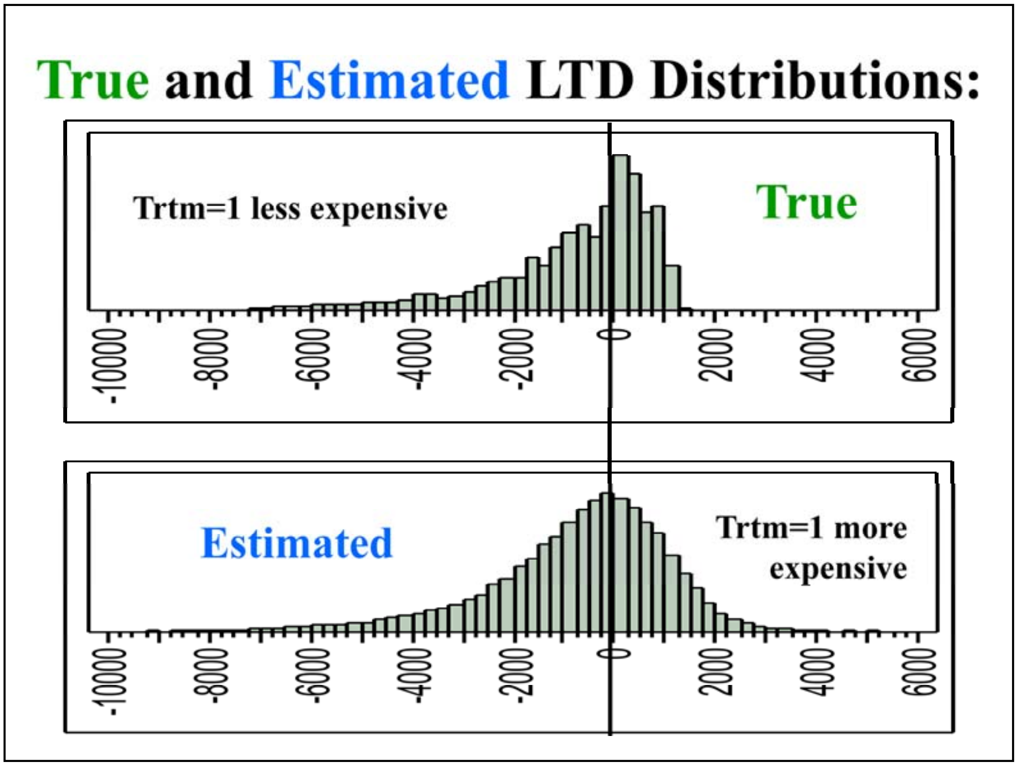
This LTD estimate is then applied to each patient within the Subgroup.

To estimate an LTD, a subgroup clearly needs to be large enough (>2 patients) to be “informative” …rather than contain only treated or only control patients.

This estimation tactic may well be most reasonable when clusters are small, but measurement error in Y-outcomes can make estimates from small clusters rather imprecise …so <u>there is a potential for trade-off </u>here.

NOTE: It’s quite clear intuitively that this “difference in mean values” statistic is both **unbiased** and fully **adjusted for confounders** when patients are well matched within subgroups.

FURTHERMORE: this pair of means consist of individual observed outcomes **weighted inversely proportional** to the **probability** of the treatment actually received. For example, the so-called “**doubly robust**” approach reduces to this simple statistic within each cluster when the “model” is **Nested ANOVA (treatment within cluster.)**

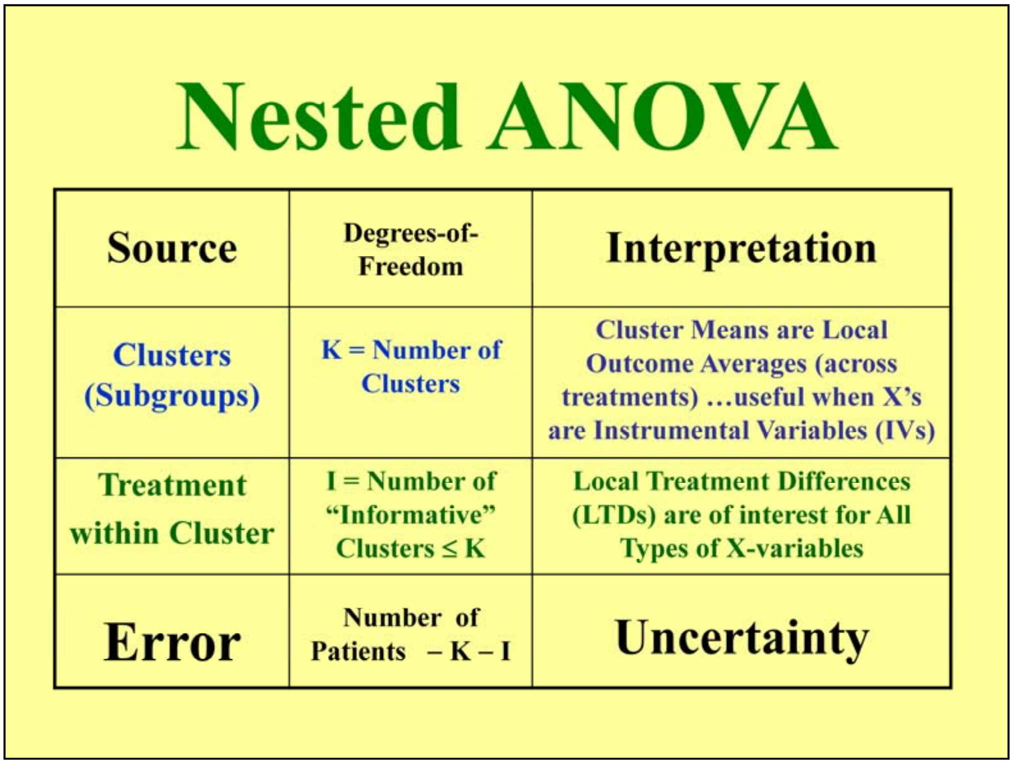


True and Estimated LTD Distributions:

An estimated LTD Distribution contains only classical, sample information and is essentially non-parametric …but can be interpreted much like a Bayesian posterior (or prior) distribution.

It is **essential to study the sensitivity** of the estimated LTD distribution to variation in number of subgroups and to choice of method used to form them …say, **unsupervised** clustering of patients in X-space or **supervised** recursive partitioning to construct a classification "tree" that yields "leaf nodes" to predict propensity for the two alternative treatment choices.

Here, the true and estimated means (treatment main effects) are almost equal (True = $-650, Estimated =$-652) in spite of the fact that the simulated yearly expenses of each patient contained **additive white noise** with a standard deviation of **$1,000.** However, the standard deviation of the full estimated LTD cost distribution was thereby increased from $1,447 for true LTDs to $1,667 for estimated LTDs.



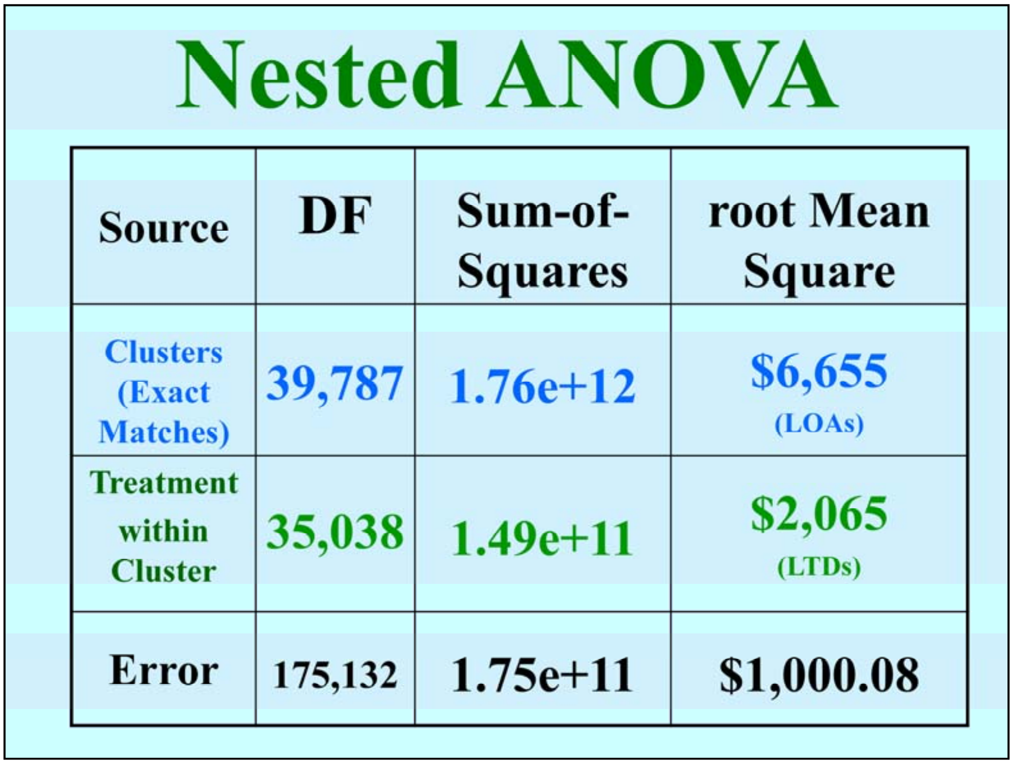
Nested ANOVA

Source	Degrees-of-Freedom	Interpretation
Clusters (Subgroups)	K = Number of Clusters	Cluster Means are Local Outcome Averages (across treatments) …useful when X's are Instrumental Variables (IVs)
Treatment within Cluster	I = Number of "Informative" Clusters ≤ K	Local Treatment Differences (LTDs) are of interest for All Types of X-variables
Error	Number of Patients – K – I	Uncertainty

When all patient X-characteristics are used ONLY to form subgroups, the resulting statistical model for treatment differences is essentially a "cell-means" model: Nested ANOVA (treatment within cluster of well-matched patients.)

McClellan et al. (1994) and many economists have championed "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. McClellan et al. (1994) proposed that cluster means be plotted vertically against a horizontal axis depicting within-cluster fraction treated ("observed" propensity score.) This approach uses information only from the "Clusters" row of the ANOVA table, and yields the display shown in Slide 16 for the numerical example. McClellan et al. (1994) contended that trends (up or down) in the displayed values from left-to-right across this plot are interpretable when all X-variables used to form patient clusters are **instrumental variables.**

The Local Control approach described above uses information only from the "Treatment within Cluster" row of the ANOVA table and yields the display shown in Slide 17. Interpretation of trends in this type of display is NOT based upon any un-testable assumptions.



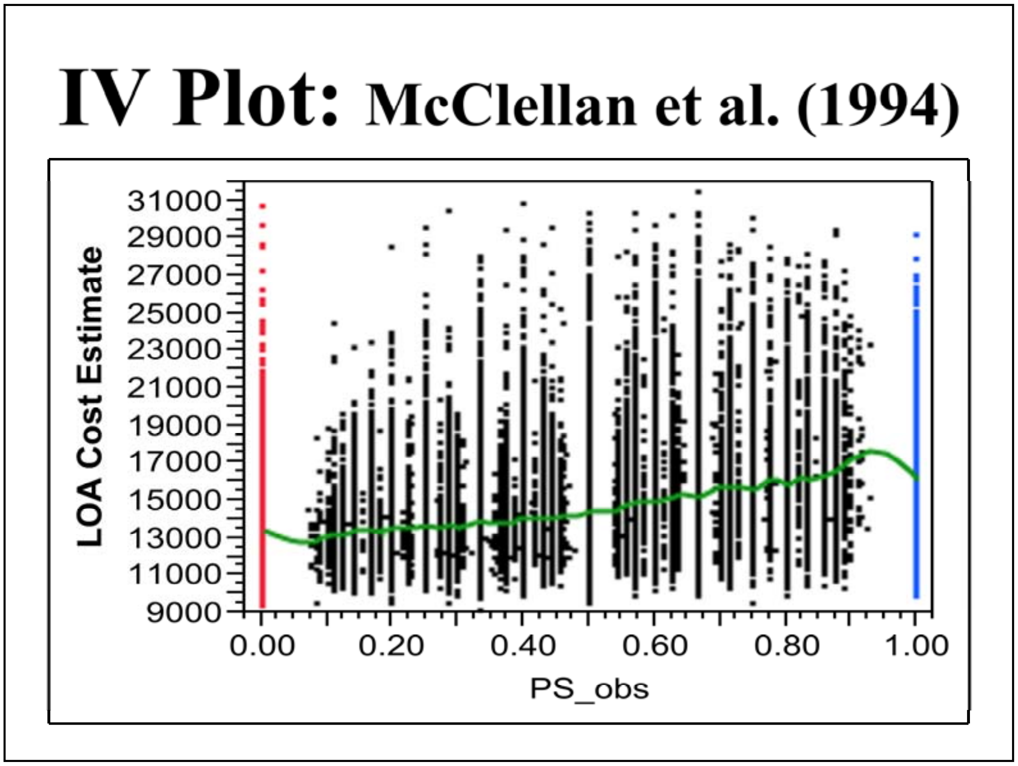
Nested ANOVA

Source	DF	Sum-of-Squares	root Mean Square
Clusters (Exact Matches)	39,787	1.76e+12	$6,655 (LOAs)
Treatment within Cluster	35,038	1.49e+11	$2,065 (LTDs)
Error	175,132	1.75e+11	$1,000.08

Nested ANOVA Table for simulated data on Yearly Costs ~250K MDD patients.

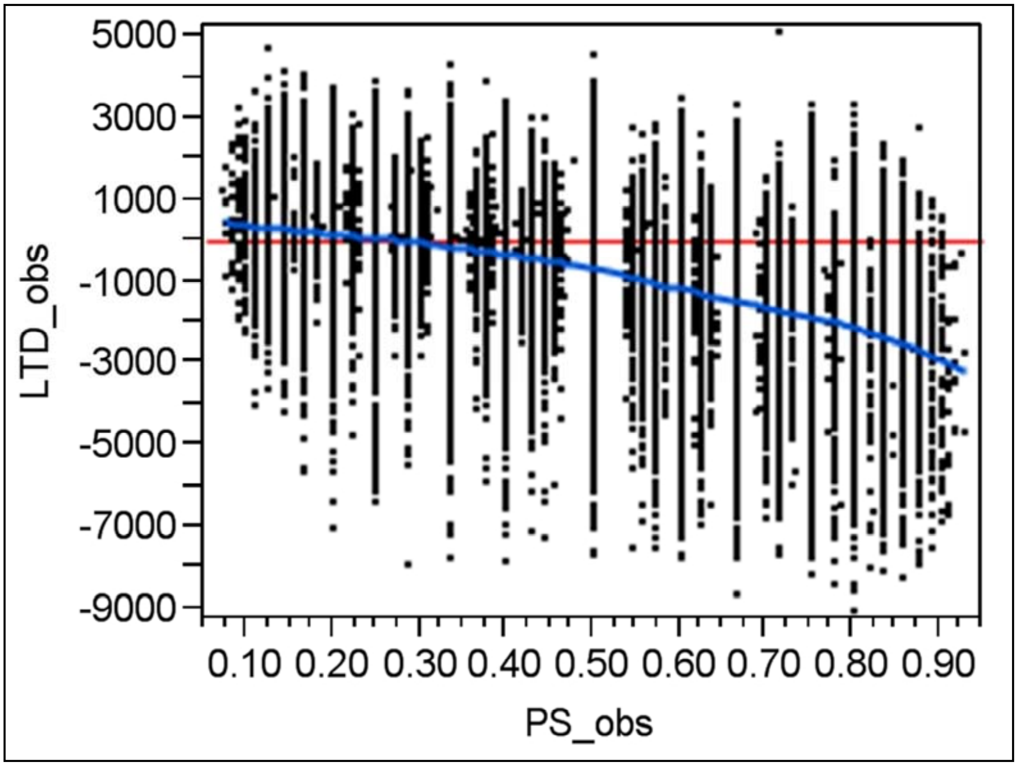
The R-square for this Nested ANOVA model is 91.6%

The computed **root Mean Square for Error** is incredibly close here to its actual, true value of $1,000 that was stated in the official Rules of our competition.



IV Plot: McClellan et al. (1994)

This plot is certainly not easily interpretable! It’s most straight-forward interpretation is almost surely that the 8 patient X-characteristics being used are NOT instrumental variables …instead of determining only treatment choice, they apparently also have direct effects on expected cost of treatment for MDD.

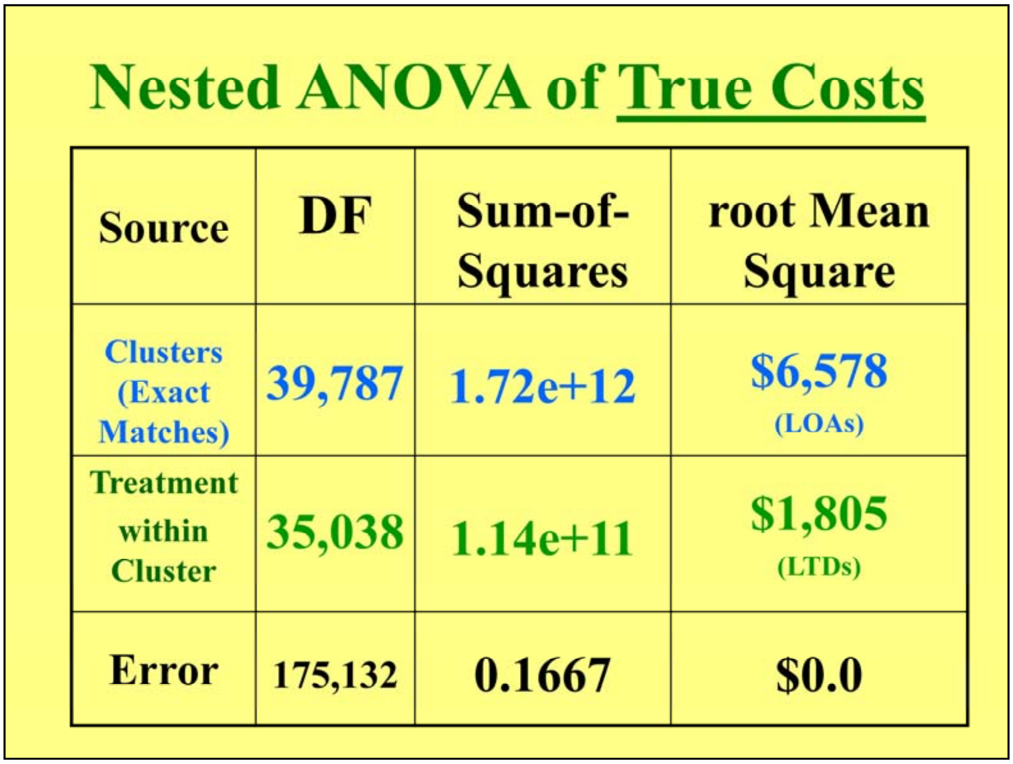


The 39,788 – 4,750 = 35,038 non-missing LTD estimates depicted in the above graphic are rather good predictors of their unknown, true values (correlation +0.855.) All of these estimates result from exact X-space matches of at least one trtm = 1 patient with at least one trtm = 0 patient.

Propensities can be predicted using a parametric model; here, a logistic regression fit with area under ROC curve = 0.606 could be used. However, the propensities used to make the above plot were simply observed as “fractions treated” within the subgroups of patients formed via exact X-space matching.

The observed propensities within these subgroups (matched sets) are binomial proportions and, thus, vary over a wider range than the true propensities, [0.25, 0.75]. Because the matched sets tend to be rather small (at most 43 patients), observed Binomial proportions again tend to look like vertical “bars” in this graphic.

Due to the inverse relationship of Trtmfrac to true Propensity in our simulation, the spline fit (blue curve) above depicts a general tendency for observed LTDs to become more negative as observed propensity for trtm = 1 increases.

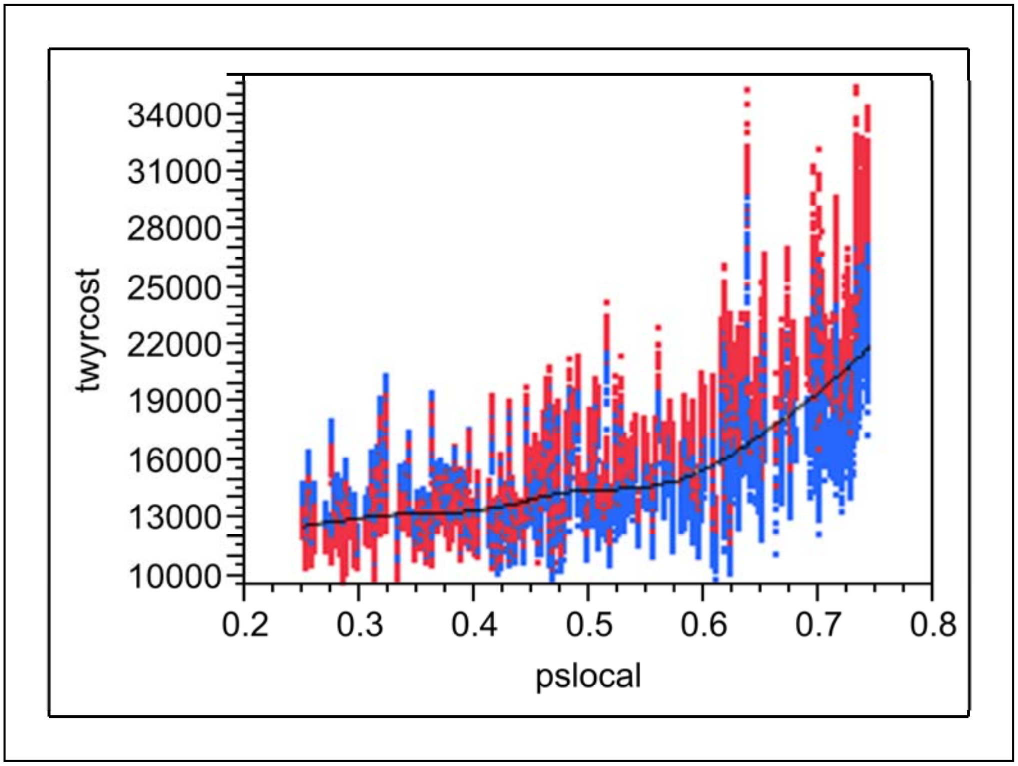


Nested ANOVA of True Costs

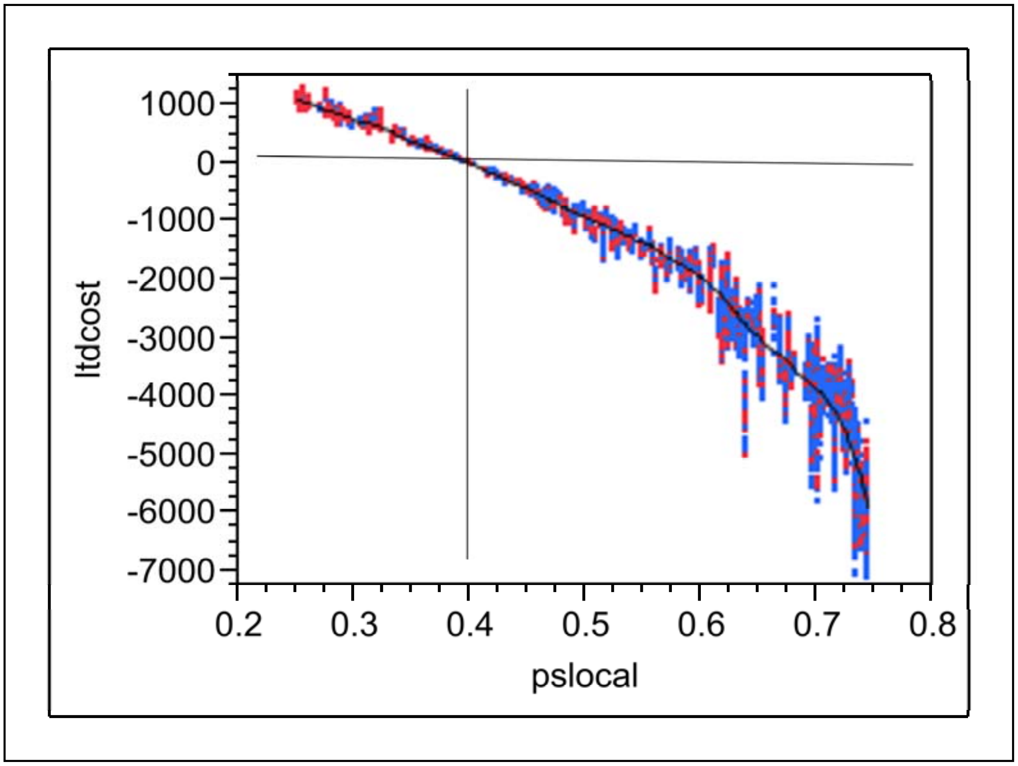
Source	DF	Sum-of-Squares	root Mean Square
Clusters (Exact Matches)	39,787	1.72e+12	$6,578 (LOAs)
Treatment within Cluster	35,038	1.14e+11	$1,805 (LTDs)
Error	175,132	0.1667	$0.0

With all additive white noise removed, the R-square for this Nested ANOVA model becomes essentially 100%.

The LTD estimates from 35,038 informative clusters formed via exact X-space matches are not exactly correct here. Specifically, the original 40,000 patients re-sampled included only 39,854 distinct X-vector patterns; 229 patients with 83 of these patterns had different true cost values.



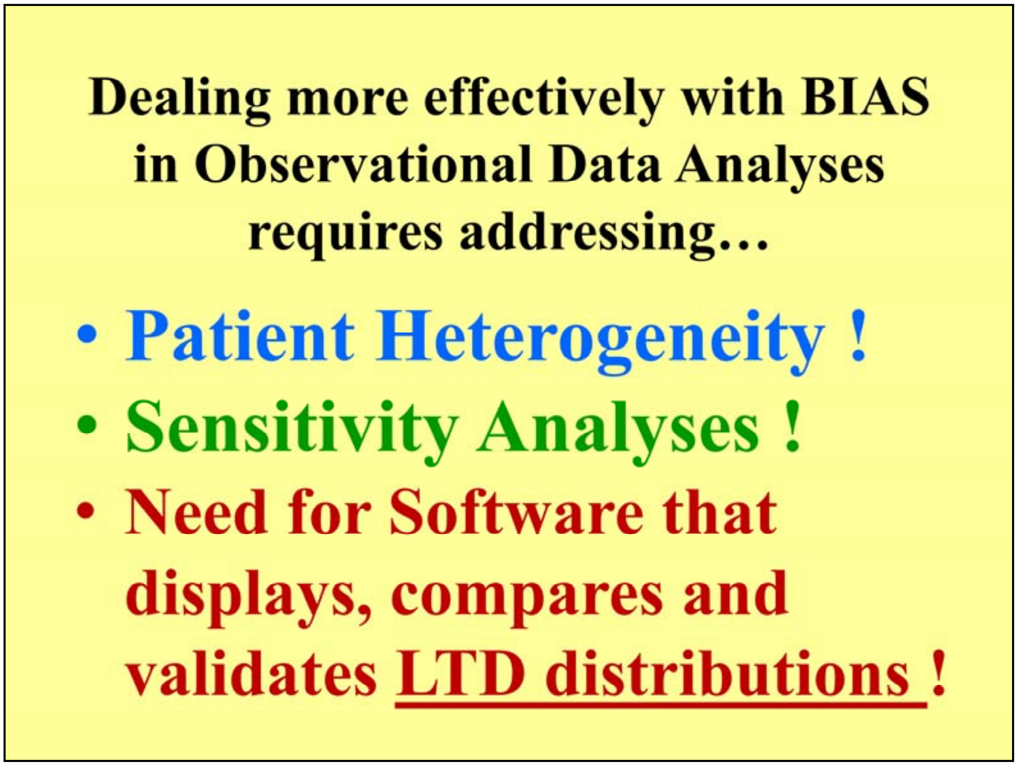
There is less over-striking here than in Slide 16. For large PS values, there are many fewer **red costs (denoting trtm=0 patients)**, but they tend to be LARGER than the **blue costs (denoting trtm=1 patients**).



Compare this “true values” plot with the plot of corresponding estimates of Slide 17.

Here only 300 different values of PSlocal are possible and the range is only 0.25 to 0.75.

There is much over-striking; **red denotes trtm=0** while **blue denotes trtm=1**.



Dealing more effectively with BIAS in Observational Data Analyses requires addressing…

- Patient Heterogeneity !
- Sensitivity Analyses !
- Need for Software that displays, compares and validates LTD distributions !

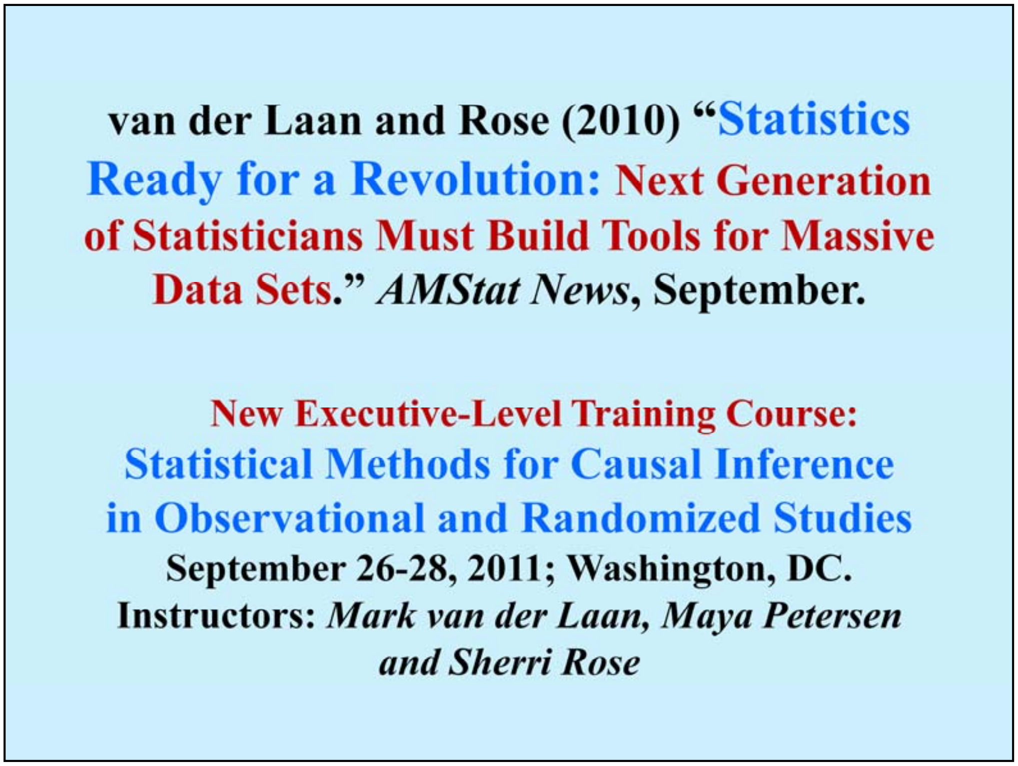
Different patients can and almost surely do respond **differently** to the same treatment.

No **single analysis** of a complicated dataset is unquestionably best or truly objective (unbiased) from the diverse perspectives of all health care stake-holders.

I claim the first two points are quite obvious. To continue **ignoring these glaring problem areas** in health care data analysis is unthinkable and irresponsible.

Some might argue that development of software to explore, graphically display, compare and validate **LTD Distributions** can wait until statistical research has shown that this approach produces estimates that are truly better (or at least different) from those provided by traditional approaches. I say that development of such software is badly needed now to specifically nurture that very research.

At some future time, statistical methods using patient subgroup (local) analysis strategy and tactics may be considered the “gold standard” approach in health care research. Until then, they can quite effectively augment or supplement traditional analysis methods.

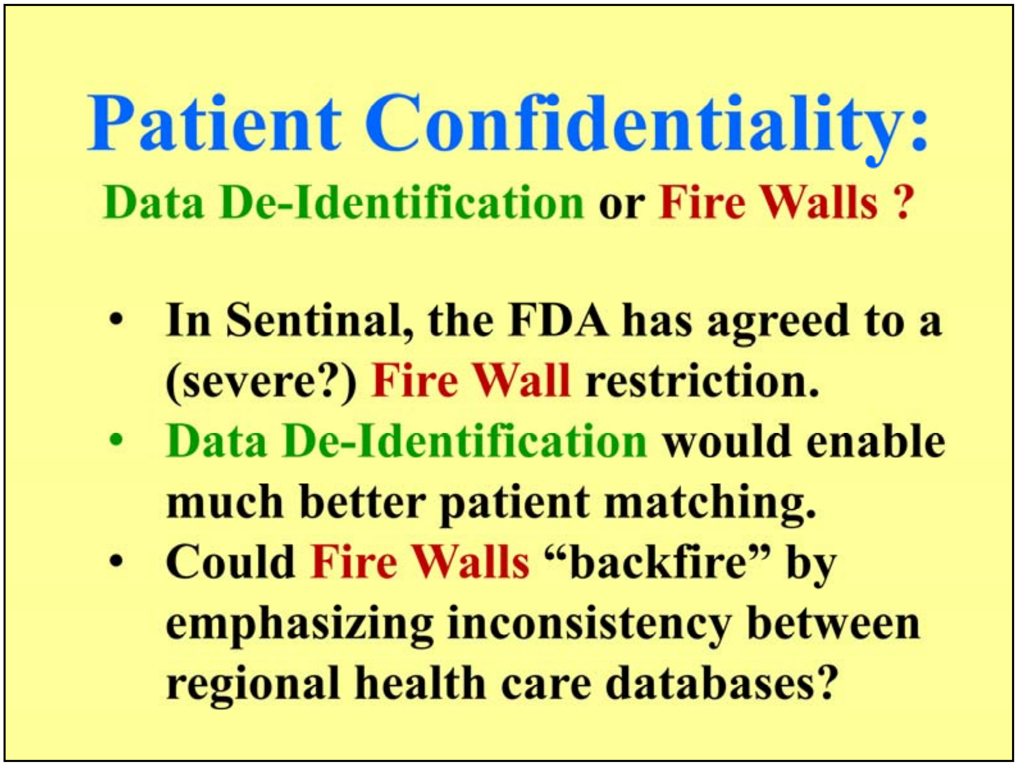


van der Laan and Rose (2010) "Statistics Ready for a Revolution: Next Generation of Statisticians Must Build Tools for Massive Data Sets." *AMStat News*, September.

New Executive-Level Training Course:
Statistical Methods for Causal Inference in Observational and Randomized Studies
September 26-28, 2011; Washington, DC.
Instructors: *Mark van der Laan, Maya Petersen and Sherri Rose*

The top reference is to a very interesting "Popular Press" article.

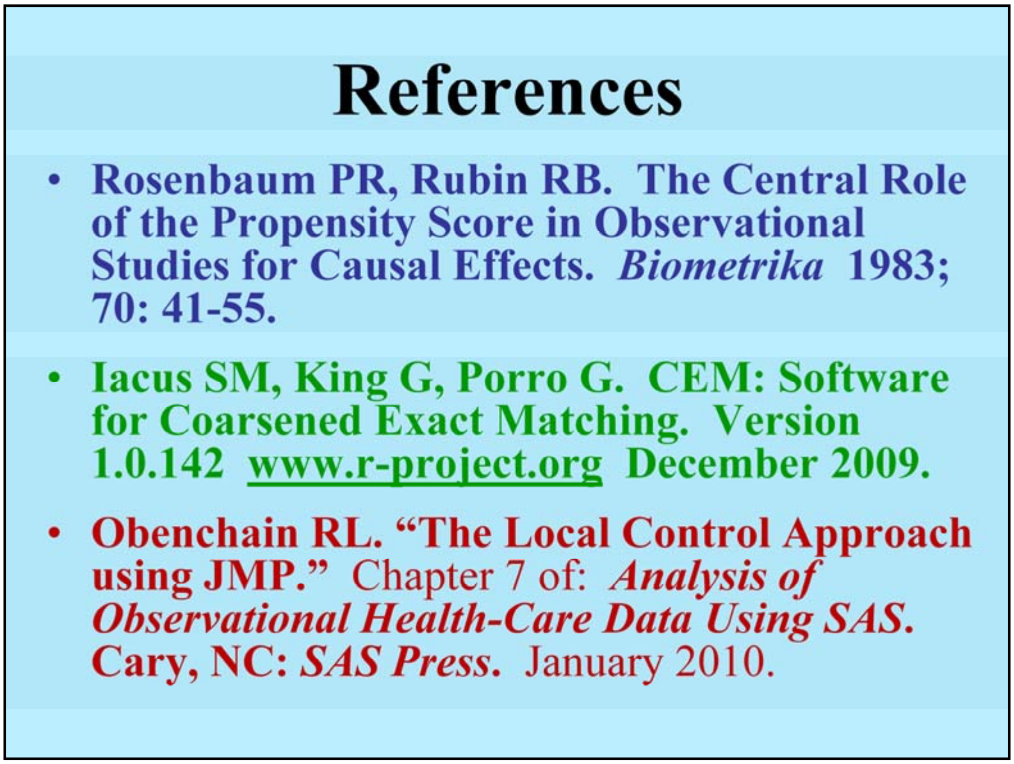
The **Forum for Collaborative HIV Research** is sponsoring a workshop this fall that will present material from their Book: *Targeted Learning: Causal Inference for Observational and Experimental Data*, van der Laan, Rose (2011), Springer: New York. Copies of the PowerPoint presentations and computer lab material using functions from the R-packages *tmle* and *SuperLearner* will be provided.



Patient Confidentiality:

Data De-Identification or Fire Walls ?

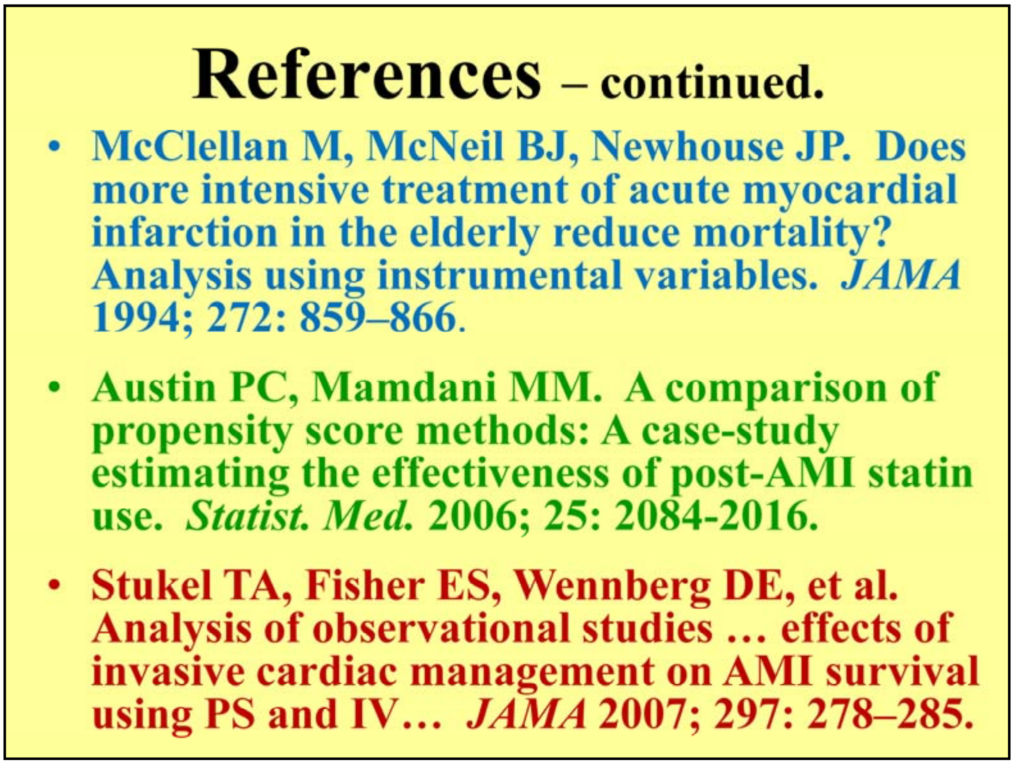
- In Sentinal, the FDA has agreed to a (severe?) Fire Wall restriction.
- Data De-Identification would enable much better patient matching.
- Could Fire Walls “backfire” by emphasizing inconsistency between regional health care databases?



References

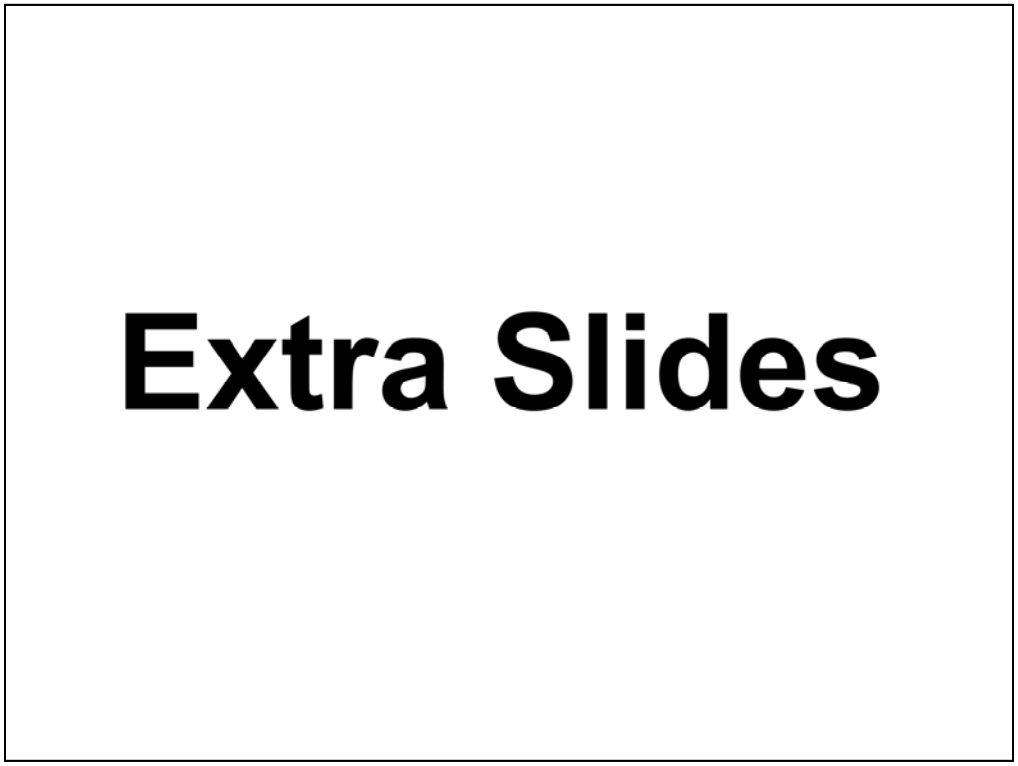
- **Rosenbaum PR, Rubin RB. The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika* 1983; 70: 41-55.**
- **Iacus SM, King G, Porro G. CEM: Software for Coarsened Exact Matching. Version 1.0.142 www.r-project.org December 2009.**
- **Obenchain RL. "The Local Control Approach using JMP."** Chapter 7 of: ***Analysis of Observational Health-Care Data Using SAS*. Cary, NC: *SAS Press*.** January 2010.

Stefano Iacus, Gary King, Giuseppe Porro, "Matching for Casual Inference Without Balance Checking: Coarsened Exact Matching," http://gking.harvard.edu/files/abs/cem-abs.shtml



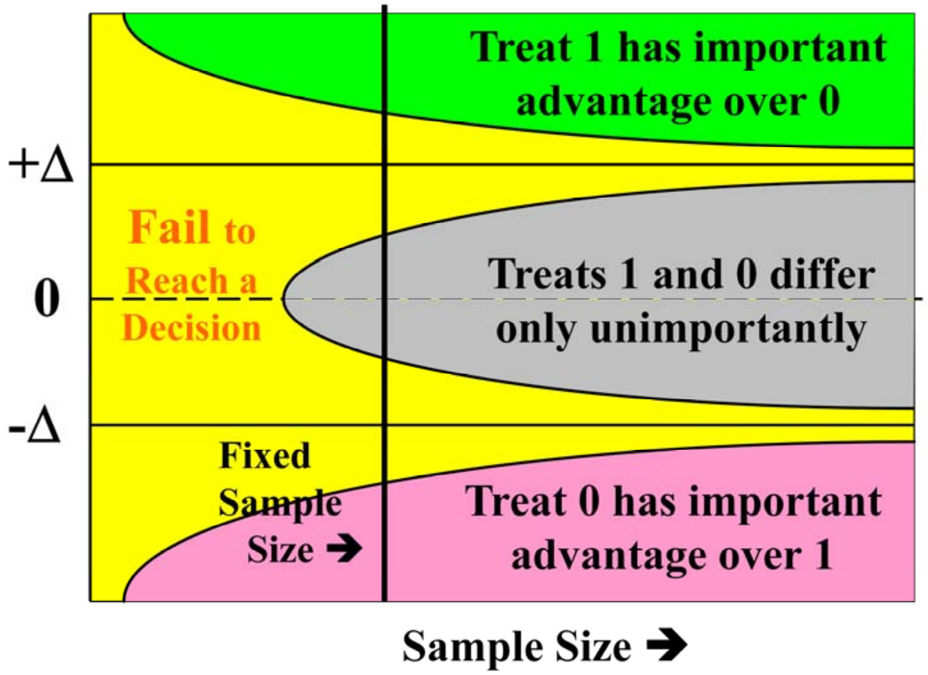
References – continued.

- 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.
- Austin PC, Mamdani MM. A comparison of propensity score methods: A case-study estimating the effectiveness of post-AMI statin use. *Statist. Med.* 2006; 25: 2084-2016.
- Stukel TA, Fisher ES, Wennberg DE, et al. Analysis of observational studies … effects of invasive cardiac management on AMI survival using PS and IV… *JAMA* 2007; 297: 278–285.



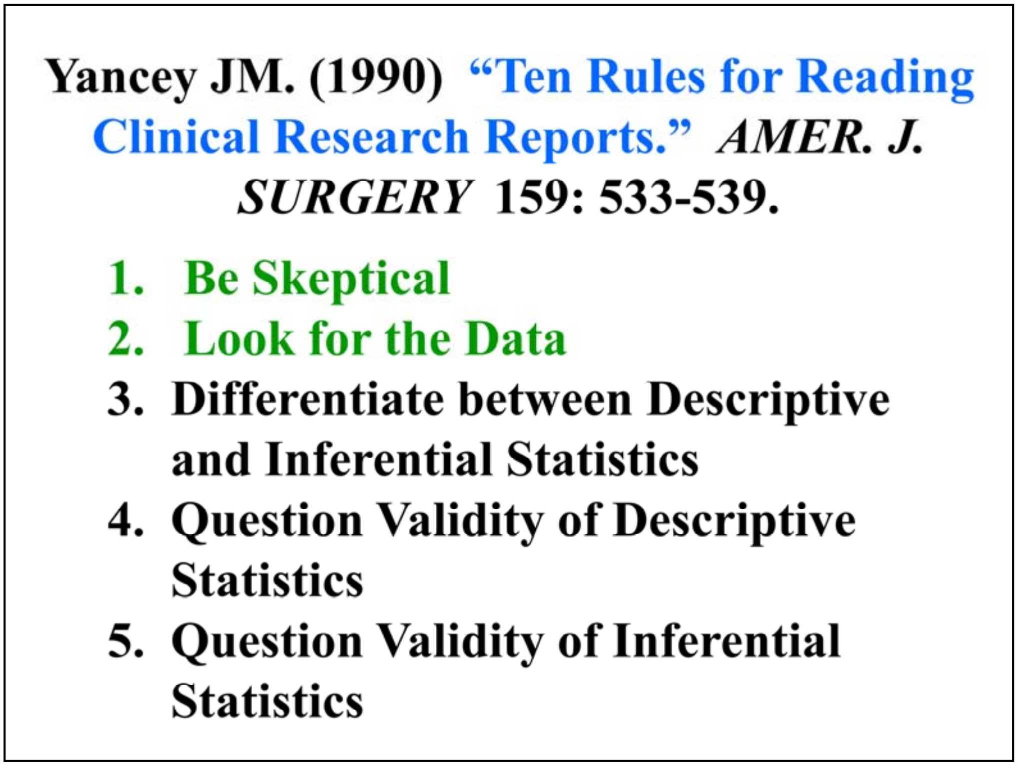
Extra Slides

Even-Handed Test of “Composite” Hypotheses


Treat 1 has important advantage over 0
+Δ
Fail to Reach a Decision
0
Treats 1 and 0 differ only unimportantly
-Δ
Fixed Sample Size ➔
Treat 0 has important advantage over 1
Sample Size ➔

p-Values should be considered unimportant …primarily because (1) they are computed relative to a “point” Null Hypothesis that can never be “accepted” as true, and (2) they are essentially irrelevant for judging “effect size.”

New terminology is needed for even-handed testing of composite hypotheses, each of which can be “accepted” or “rejected.” The DELTA above is the minimum effect size considered medically important. Ordinary confidence intervals for either Local Treatment Differences or the Main Effect of Treatment are all that is needed to create these new kinds of (unadjusted) composite hypothesis tests.



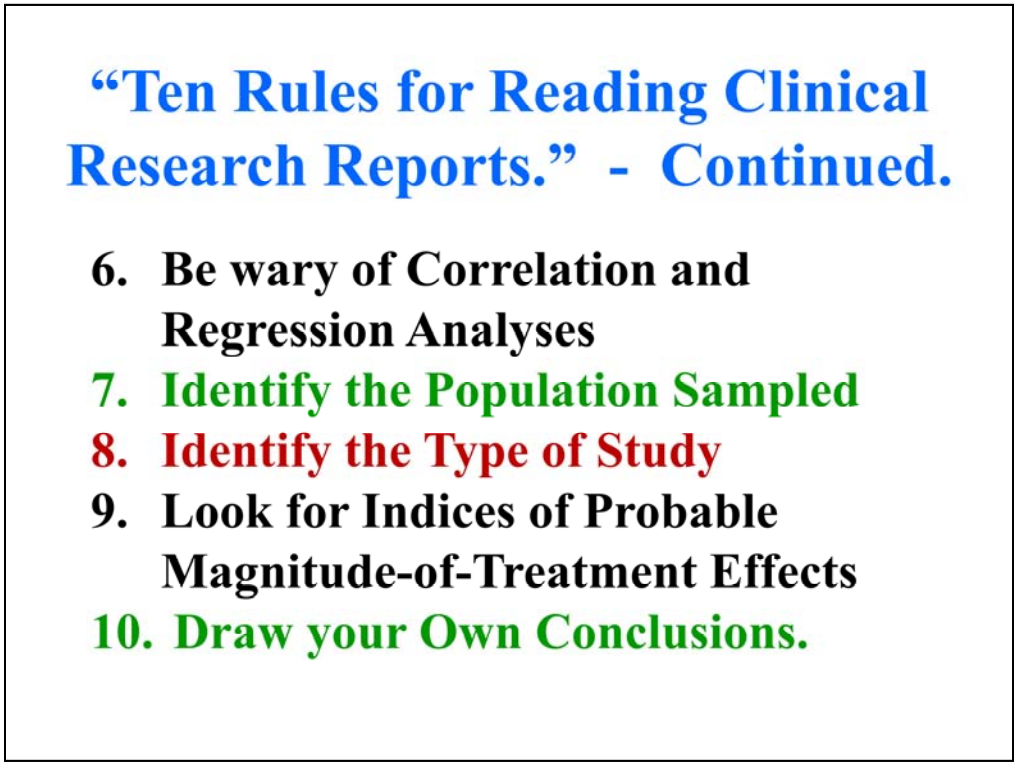
Yancey JM. (1990) "Ten Rules for Reading Clinical Research Reports." *AMER. J. SURGERY* 159: 533-539.

1. Be Skeptical
2. Look for the Data
3. Differentiate between Descriptive and Inferential Statistics
4. Question Validity of Descriptive Statistics
5. Question Validity of Inferential Statistics

3: Describe available data vs. Attempt to answer questions about data not yet collected (& perhaps never will be.)

4: Descriptive statistics too often are simple measures of central tendency. Which patients are AVERAGE?

5: Inferential Statistics should ask if effects are **big enough to be medically important** …not simply "non-zero." Statistical tests where all hypotheses (the null and the alternatives) are "composite" (consisting of a range of values rather than a single point) do not yield p-values …because there are no unique, obvious places to compute the statistical size or power of the test.



"Ten Rules for Reading Clinical Research Reports." - Continued.

6. Be wary of Correlation and Regression Analyses
7. Identify the Population Sampled
8. Identify the Type of Study
9. Look for Indices of Probable Magnitude-of-Treatment Effects
10. Draw your Own Conclusions.

6: As in point 4, these methods typically address whole populations rather than individual patients. Much wider bands than those being reported apply to predictions for individual patients.

8: Yancey's Opinion: Only Randomized, Tightly Controlled and Prospective studies provide clear evidence of cause-effect relationships.

9: Again, statements about the direction of differences and p-values do NOT do this. Credibility is increased by providing confidence intervals or prediction limits for individual patients.