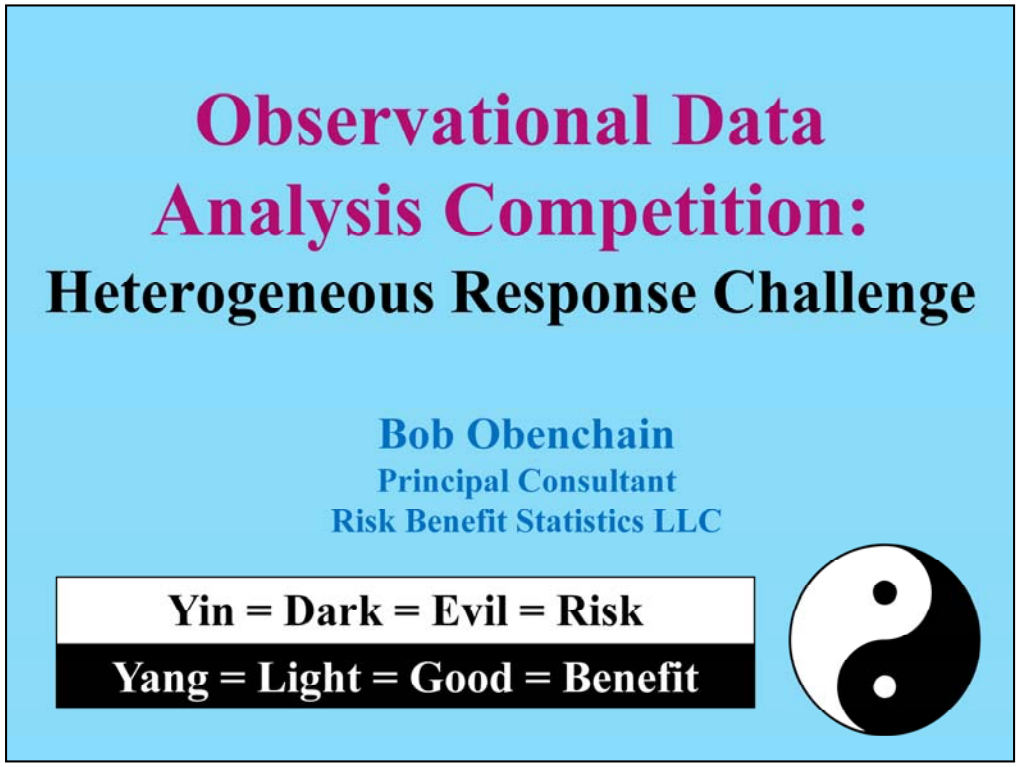


The observational data analysis competition discussed in this session is called the "Heterogeneous Response Challenge." The objective of the competition was to accurately quantify how different patients are expected to respond differently to a pair of treatments.

This presentation describes how the competition data were simulated to illustrate [1] patient differential (heterogeneous) response to treatment and [2] analytical complications due to unmeasured confounders, treatment selection bias, and measurement error.

When a observational data simulation algorithm developed earlier was used to generate a super-sample ten times larger than originally intended, considerable serendipity resulted. Large numbers of patients who are very well matched on all 8 of their observed baseline X-characteristics are now suddenly also matched on their unobserved confounders! Much more accurate estimates of heterogeneous responses thus become available.

Unfortunately, such serendipitous situations are not common; detection of heterogeneous patient response is typically a very difficult problem.

The primary take-away from this session is thus that health services researchers should seriously consider using simulation to develop much more realistic assessments of alternative approaches to analysis of observational data. The simulation tools and tactics described here are just one possible starting point for this sort of badly need health outcomes research.

Heterogeneous Response Challenge:

Given observational data on a single *y*-outcome and eight *x*-characteristics of ~250K patients who received either trtm = 1 or trtm = 0,

Accurately quantify how different, individual patients are expected to respond differently:

$$E(y \mid x, \text{trtm}=1) - E(y \mid x, \text{trtm}=0)$$

true local treatment difference

Competition One: Most accurately estimate all 249,958 of the LTD values above. I.E. Answer must have no missing values.

Competition Two: Most accurately estimate the Treatment Main-Effect …defined as the true mean of these 249,958 LTD values.

Runners-Up Competition: Most accurately estimate the LTD values with no more than 25,000 missing values. I.E. competitor chooses which 224,958 values to estimate.

Observational Data

- **Patients are not randomized to treatment in any *known* way …i.e. potential treatment selection bias.**
- **Perhaps many patients, but less is known about individual patients than is typical in clinical studies.**
- **Unknown confounders are present.**

Competition data were simulated in a way intended to be highly realistic …I.E. with implementations of all features typical of actual observational data, except (for simplicity) no missing values are present.

A Daunting Challenge!!!

Entrants are being asked to estimate ~250,000 parameters (expected values) from data on ~250,000 patients

…in a situation where the conditional expected outcome for each individual patient, $E(y_i \mid x_i, \text{trtm}_i)$, is not even one of the "counterfactual difference" parameters of interest!

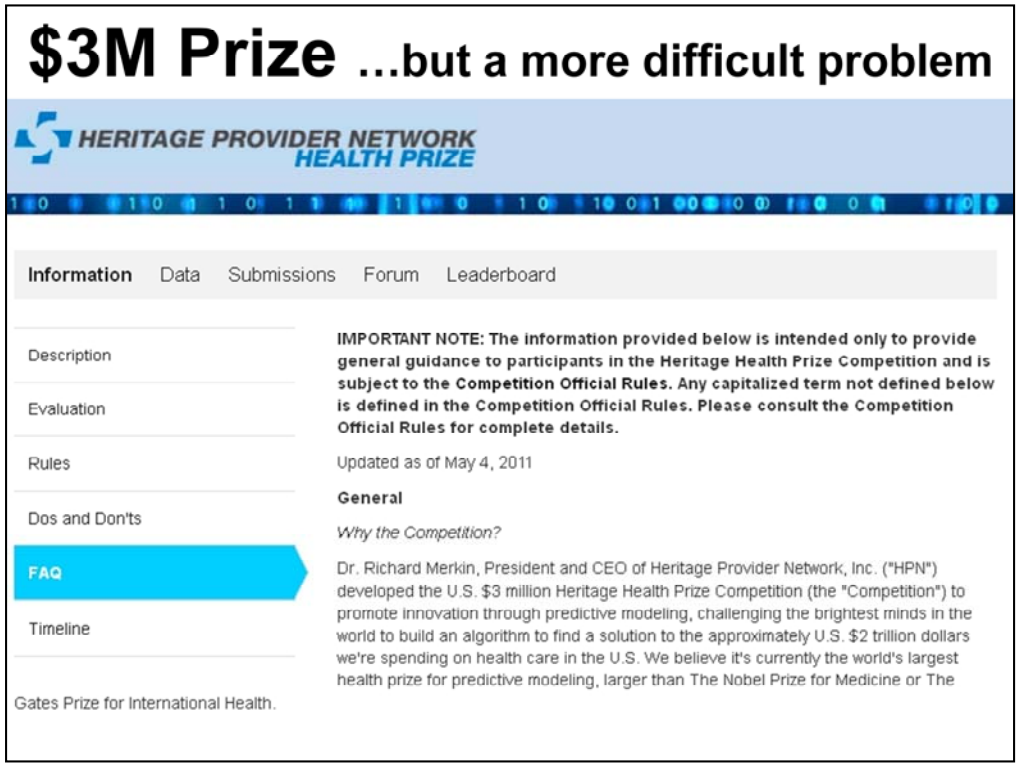
$$\text{LTD}_i = E(y_i \mid x_i, \text{trtm}=1) - E(y_i \mid x_i, \text{trtm}=0)$$

Clearly the analyst must make some assumptions and/or fit some model to generate his/her answers.

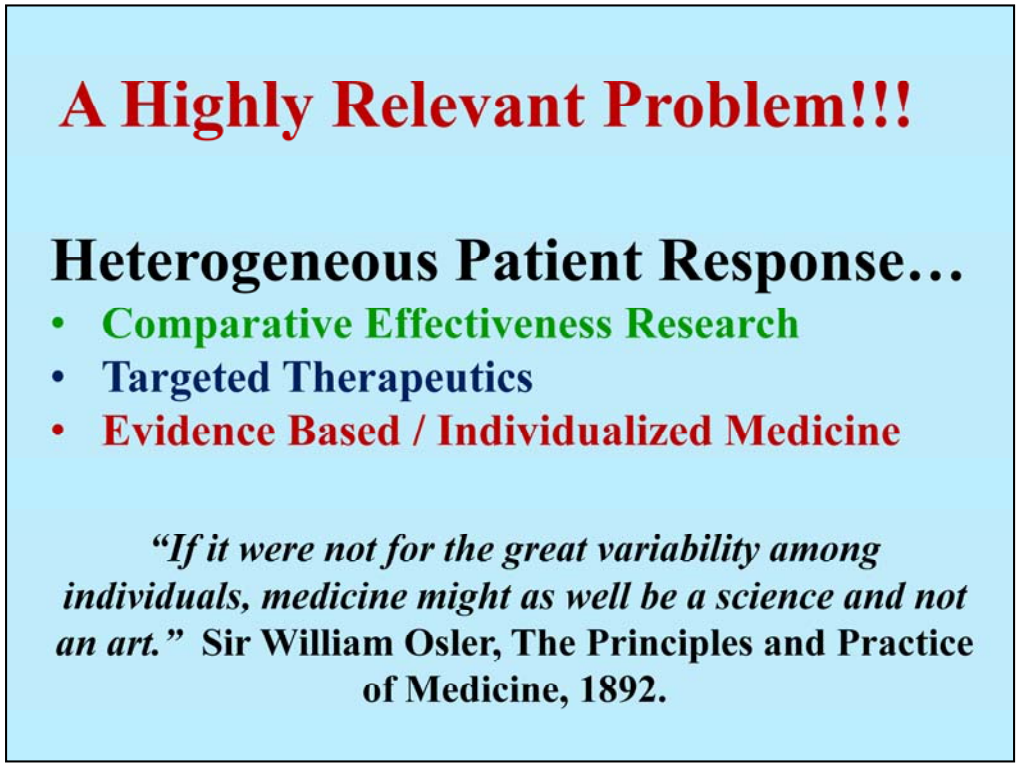
The "Good" News: Only point estimates are being requested …no confidence intervals, p-values, statistical tests of hypotheses, causal inferences, rational explanations or parsimony is required to win this competition.

The "Bad" News: Entrants cannot follow the traditional "Business-as-Usual" practices typical in analysis of RCT data. Specifically, to do "really well" in this competition, entrants will have to literally LOOK AT THE DATA before deciding how to analyze it!!! For example, since the number of observations (patients) is an even number (249,958), the data might possibly consist of 124,979 patient-pairs well matched on their x-vectors but with different trtm choices. Is this the case? (Certainly not exactly this. After all, there are 137,163 trtm=0 patients and only 112,795 trtm=1 patients.) Slide 10 describes other "insight" strategies.

If the competition dataset were "real" (rather than simulated), TRUE parameter values would be UNKNOWN and entrants might be able to earn "style points" in the process of scoring / ranking answer "quality." Here, the sole evaluation criterion is root Means Squared Error LOSS of 249,958 LTD estimates from their (unrevealed) true values.



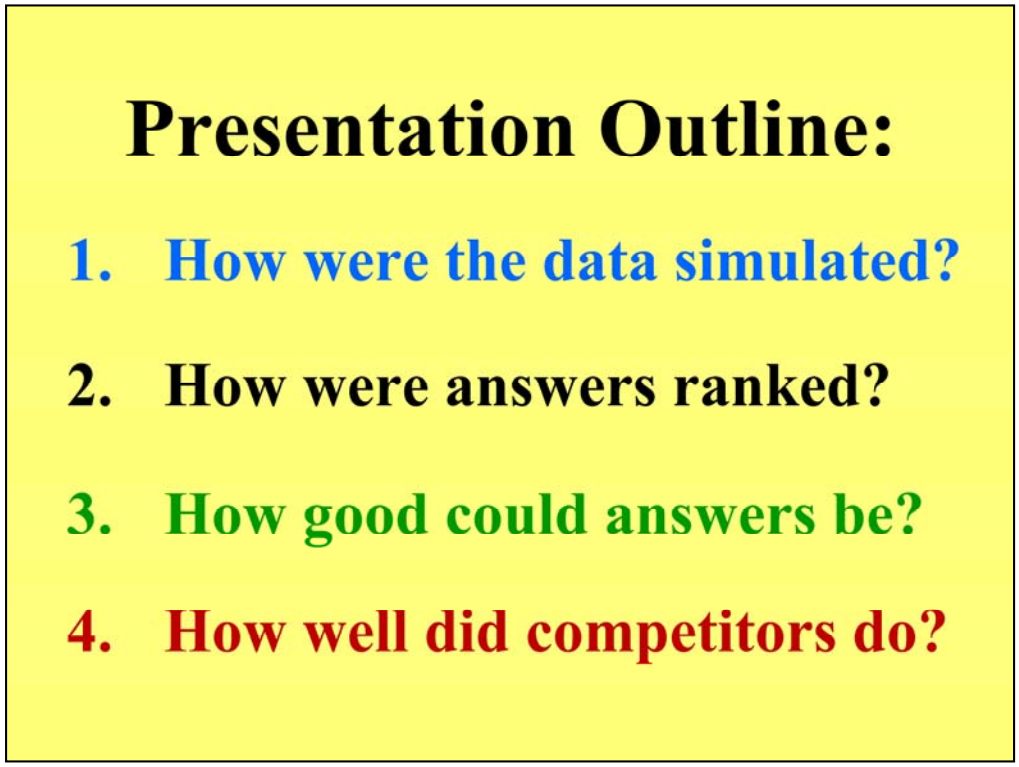
Perhaps, our competition failed to stimulate a flood of entrants because we didn't offer a substantial, monetary prize to stimulate participation!!!



In fact, why shouldn't our competition have attracted much more attention than it actually did?

Apparently, health outcomes researchers do not yet recognize how important an issue heterogeneity is 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.**



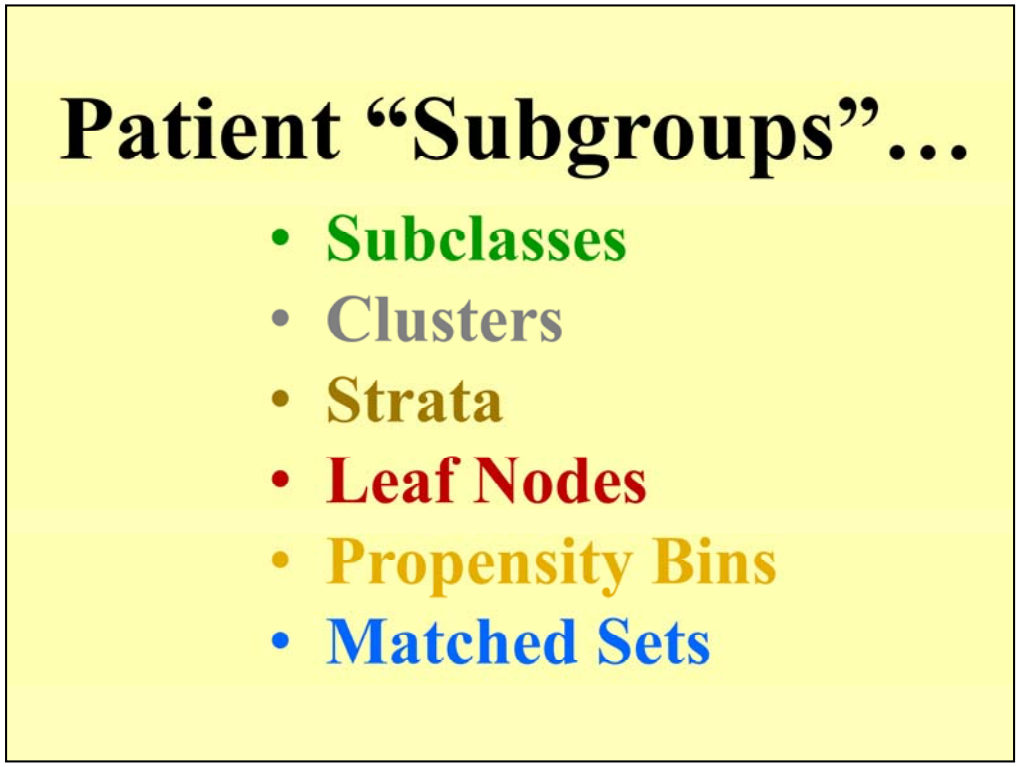
Much of this presentation (17 of 36 slides) concerns the first of these four topics.

Observational data simulation is an important, unexplored topic. I would like to encourage other health services researchers to consider using simulation techniques to get a much more realistic view of how difficult it is to arrive at good, data-based answers to important health care policy questions.

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.



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.

Method	50K Patients	250K Patients
Multivariable Regression	YES	YES
Hierarchical Clustering	YES	NO
K-means Clustering	YES	YES
Recursive Partitioning	YES	YES

Some approaches cannot be used in very large samples due to computational intensity and/or hardware memory restrictions.

Of the clustering approaches, only the K-means algorithms appear suitable in very large samples (like ours.)

This approach can be very fast when the researcher is looking primarily for **exact matches** in X-space. To do this, simply run the algorithm (such as SAS proc FASTCLUS) requesting **very, very many clusters**!

As we will see later in this presentation, requesting 40,000 or more clusters will reveal that there are only 39,788 distinct X-vector patterns in the competition dataset. On the other hand, an easier way to "discover" these guaranteed **balancing scores** (distinct X-vectors) is to run, say, SAS proc MEANS or use the aggregate( ) function in R.

Once subgroups have been formed, what does one typically do next?

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 seem to be most reasonable when clusters are small, but measurement error in Y-outcomes can make estimates from small clusters rather imprecise …so there is a trade-off 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: these 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 when the "model" is Nested ANOVA (treatment within cluster.)

Simulation Strategy/Tactics:

Use a MIXTURE of approaches so that no single approach can be "best" in the sense that all of its assumptions are most realistic.

…Continuing, joint work with my former Lilly colleagues: Doug Faries, Quan Hong and Tony Zagar.

This simulation work has not yet been submitted for publication. It focuses upon samples of 25,000 patients and shows that estimation of LTDs is very difficult when the true variation in LTDs (heterogeneity of response) is not large relative to the "noise" (measurement error) in the data.

Due to the competition dataset being 10-fold larger than those we originally simulated, new possibilities for matching patients on X-vectors resulted in considerable serendipity!!!

Variables in the Dataset

249,958 patients with Major Depressive Disorder (MDD).

[1] segno : Patients are numbered sequentially
[2] wyrcost : Windsorized (≤ $50K) After-Baseline-Year Cost.

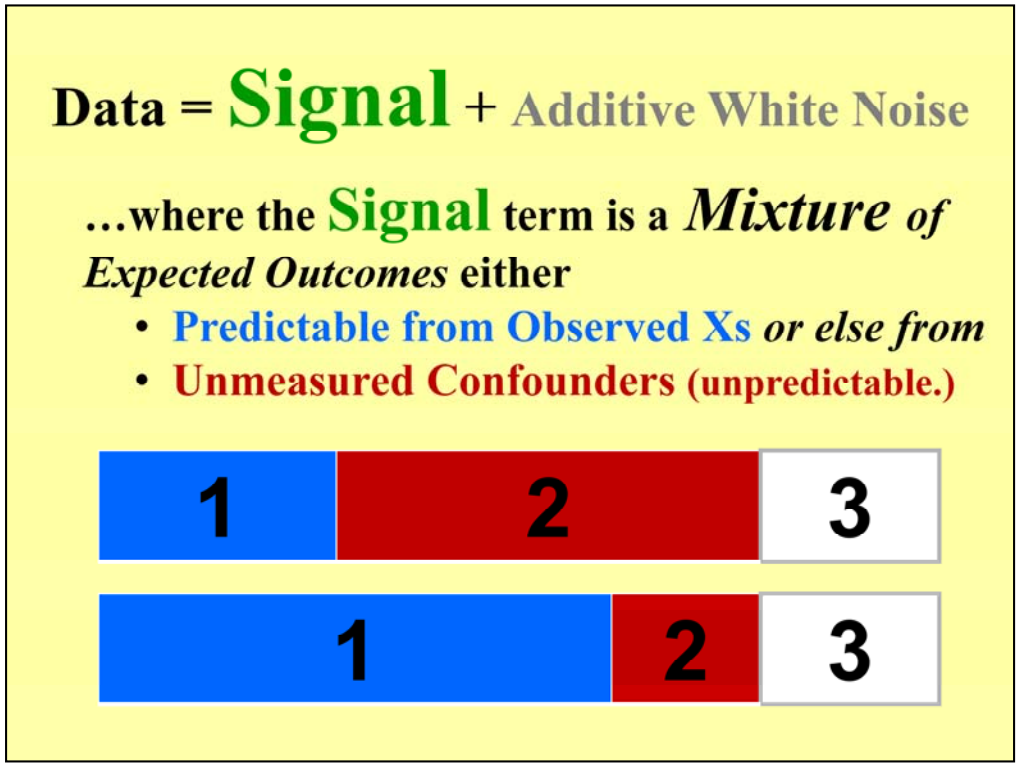
[3] trtm : Binary (0, 1) indicator for two hypothetical treatments.

[4] age : age in years (18 to 64)
[5] female : Binary gender indicator (1 => yes, 0 => male.)
[6] pain : 0, 1 or 2 (lower back and/or neuropathic.)
[7] hoscount : Hospitalizations (year before baseline)
[8] ercount : Emergency Room Visits (year before baseline)
[9] ofcount : Office Visits (year before baseline)
[10] psycpcnt : PSYC Visits (Percent, 0 =19%, ≥ 58 =99%)
[11] wprevcost : Windsorized (≤ $50K) Before-Baseline-Year Cost.

13

The 8 patient baseline X-characteristic variables (4 to 11) show highly realistic variation in the sense that they are quite similar to those computed for 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.



Using a **Mixture Proportion** (called “pmix” in the simulation R code) allows the average, total amount of **signal** in the data to be held approximately fixed while the relative importance of contributions from [1] observed patients X-characteristics (shown in BLUE) and [2] unmeasured confounders (shown in RED) can be deliberately varied.

The two sets of bars shown above depict **pmix** values of 0.33 (top) and 0.80 (bottom.)

The competition dataset was generated using equal proportions of predictable and unpredictable signal, **pmix = 0.50.**

The distinction between predictable and unpredictable signal components is revealed on the next slide.

- **The Predictable Signal for each trtm = 0 patient is the prediction from a fitted Multivariable Regression Model that is "Factorial to Degree Two" in the 8 given patient X-characteristics.**

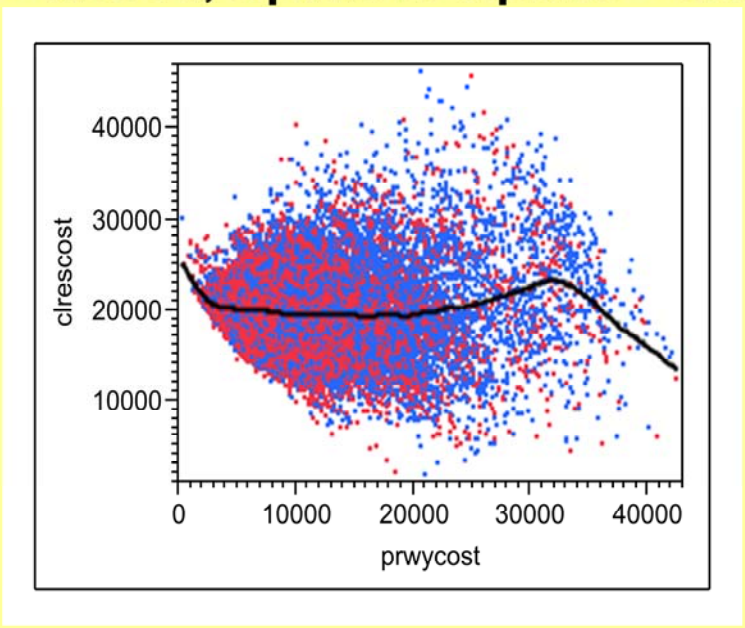
- **The Unpredictable Signal for each trtm = 0 patient is the fitted residual from a different Multivariable Regression Model that is also "Factorial to Degree Two" in the given Xs.**

- **The y-Responses used to define the above residuals were constant within 300 X-space clusters defined using data from 40,000 patients.**

These global. parametric models that are "Factorial-to-Degree-Two" contain 8 linear terms and 28 2-way interaction terms …but no squared terms.

This strategy for determining the full cost "signal" for trtm = 0 patients is clearly a MIXTURE of the global, parametric modeling and patient subgrouping approaches.

Corr = 0.0001; Spline R-square = 0.0224



RED points indicate the patients assigned to trtm = 0; their two cost components will remain as displayed above.

BLUE points indicate the patients assigned to trtm = 1. As explained on the next 3 slides, both coordinates for each BLUE point will ultimately be multiplied by factors (<1, =1 or >1) that vary with X, creating heterogeneous responses due to assignment to trtm = 1.

The purpose of the above graphic is to show that the predictable (horizontal) and unpredictable (vertical) components of the true cost signal started out essentially uncorrelated. Similarly, the (black) fitted spline curve shows that the unpredictable (vertical) component is also not non-linearly predictable from the horizontal true cost component.

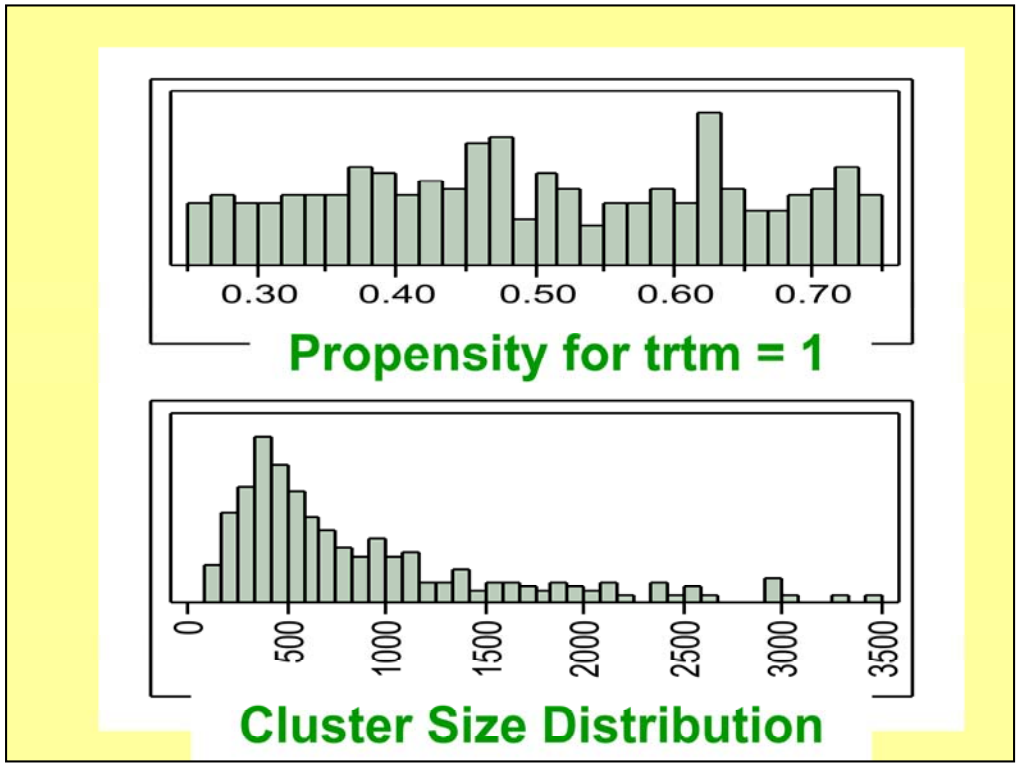
Each Individual Patient (as characterized by his/her *x*-vector) who is sampled is randomized to trtm (choices 0 or 1) via a single Bernoulli trial with given Propensity for trtm = 1.

- **Propensities are constant within 300 clusters.**
- **Propensities vary from 0.25 to 0.75.**
- **Propensities are assigned to clusters so as to be positively associated with high, predictable cost on trtm = 0 (correlation +0.803.)**

The Signal for each trtm = 1 patient is given by multiplying his/her trtm = 0 Signal by a FACTOR that depends only on this Propensity.

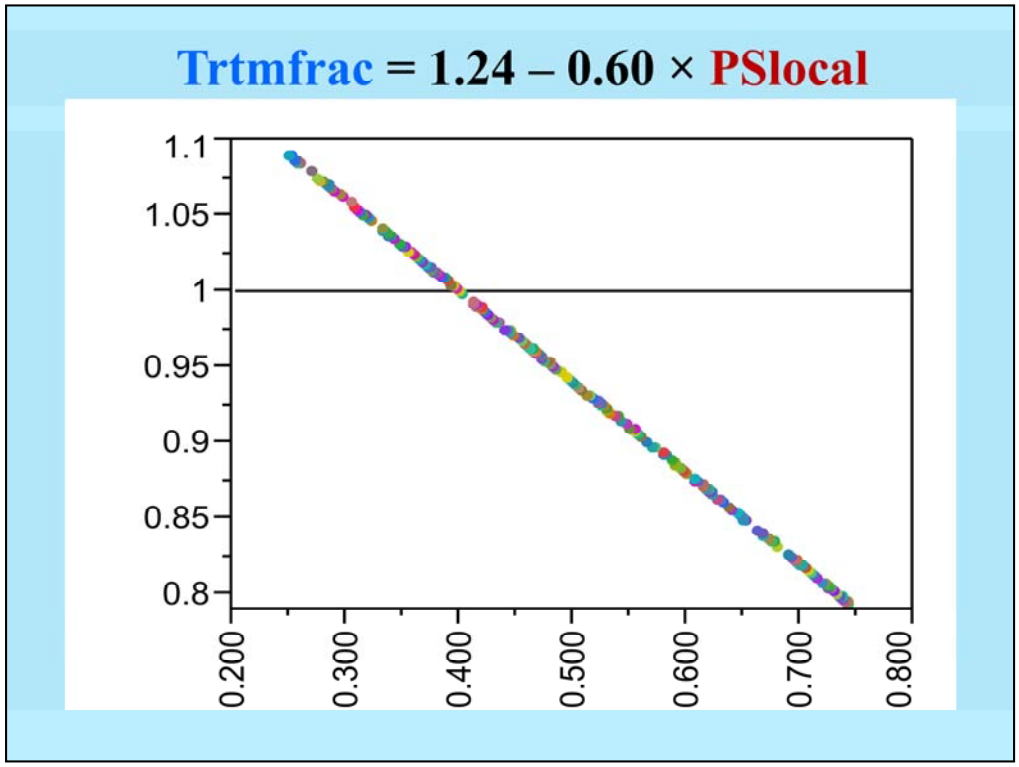
Hierarchical clustering was used to divide 40,000 original patient X-vectors into 300 mutually exclusive and exhaustive subgroups. Information about these subgroups was not provided to competitors entering the HR Challenge.

The next two slides illustrate these features of our observational data simulation algorithm.



The 300 within-cluster values of Propensity for trtm = 1 are approximately uniformly distributed on [0.25, 0.75].

Distribution of the original 40K patients into the 300 clusters was highly skewed, with a few very large clusters.



Note that 0.25 < or = PSlocal < or = 0.75 implies that 0.79 < or = Trtmfrac < or = 1.19 …with Trtmfrac decreasing linearly as PSlocal increases, as shown above.

Interpretation: **Patients have tended to gravitate towards the MDD treatment which is less expensive for them.** I.E. patients subject to relatively high costs because they desire aggressive treatment for MDD (such as psychotherapy and high acquisition-cost meds) tend to use trtm=1 because their ultimate net costs would be even higher if they chose trtm=0.

To generate the Competition Dataset, a sample of total size 250K patients was requested, with replacement, from 40K patients distributed across the 300 distinct clusters as in Slide 18.

- **Re-sampling was intended to be strictly proportional to cluster size.**
- **But an integer number of patients needed to be selected from each cluster.**
- **Due to the above restrictions, the resulting total sample size was 249,958 patients.**

Analysis of these data revealed that "only" 39,788 exact X-space matches were possible.

All of these almost 40K patient subgroups are not informative about their corresponding average LTD. In fact, 3,025 of these subgroups contain only trtm = 0 patients (12,230), while 1,725 subgroups contain only trtm = 1 patients (6,217.) As a result, a total of 18,447 patients have a missing value as their LTD estimate from the subgroups available via exact X-space matching.

On overall average, each of the original 40K patients was expected to be re-sampled more than 6 times. As it turned out, 66 patients were never selected, 493 were re-sampled only once, 6419 patients were re-sampled exactly 6 times (modal value), and 1 patient was re-sampled the observed maximum of 22 times.

Coarsened Exact Matching (CEM)

- New Approach similar to well-known graphical display tech?
- Competition patient X-variables are already "coarsened."
- Implemented in an R package.

It's somewhat curious that the competition y-outcome variable, **wyrcost**, varies only between $6,200 and $36,800. And no two of these values are equal …because they are expressed to the nearest penny! If they were rounded to the nearest dollar, there would be only 16,998 distinct values …slightly more than half of the observed range of $30,600.

Meanwhile the last X-variable, **wprevcost**, is actually the same variable as the y-outcome …except from the previous year. Its range is from $100 to $50,000 with many ties at both of these two extreme values (2008 and 2853, respectively) due to Windsorizing. All of these values are already rounded to the nearest dollar, and there are 16,768 distinct values within a much wider range of $49,900.

Algorithmically, exact matches are extremely easy to find in gigantic datasets …all one essentially needs to do is to first sort the data on all X-variables. In fact, SAS proc FASTCLUS runs in just a few minutes on the competition dataset as long as the analyst **requests 40K or more clusters.**

NOTE: The CEM functions implemented in R were not actually used in my analyses. However, very similar concepts were used.

Nested ANOVA

Source	Degrees-of-Freedom	Interpretation
Clusters (Subgroups)	K = Number of Clusters	Cluster Means are Local Outcome Averages (across treatments) 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

McClellan et al. (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. McClellan et al. (1994) proposed that cluster means be plotted vertically against a horizontal axis depicting within-cluster fraction treated (propensity score.) This approach uses information only from the "Clusters" row of the ANOVA table, and yields the display shown in Slide 24. 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 uses information only from the "Treatment within Cluster" row of the ANOVA table and yields the display shown in Slide 25. Interpretation of trends in this type of display is NOT based upon any un-testable assumptions.

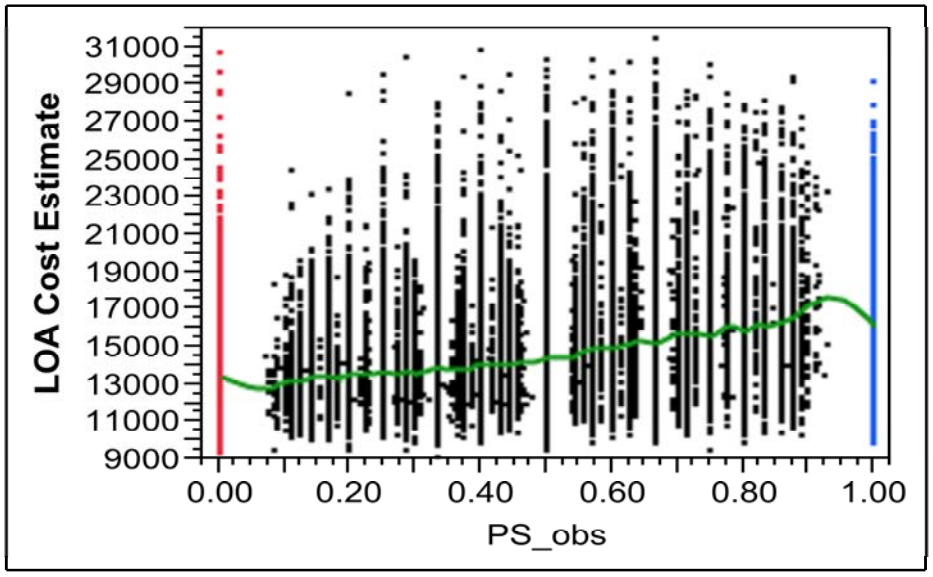
Nested ANOVA

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

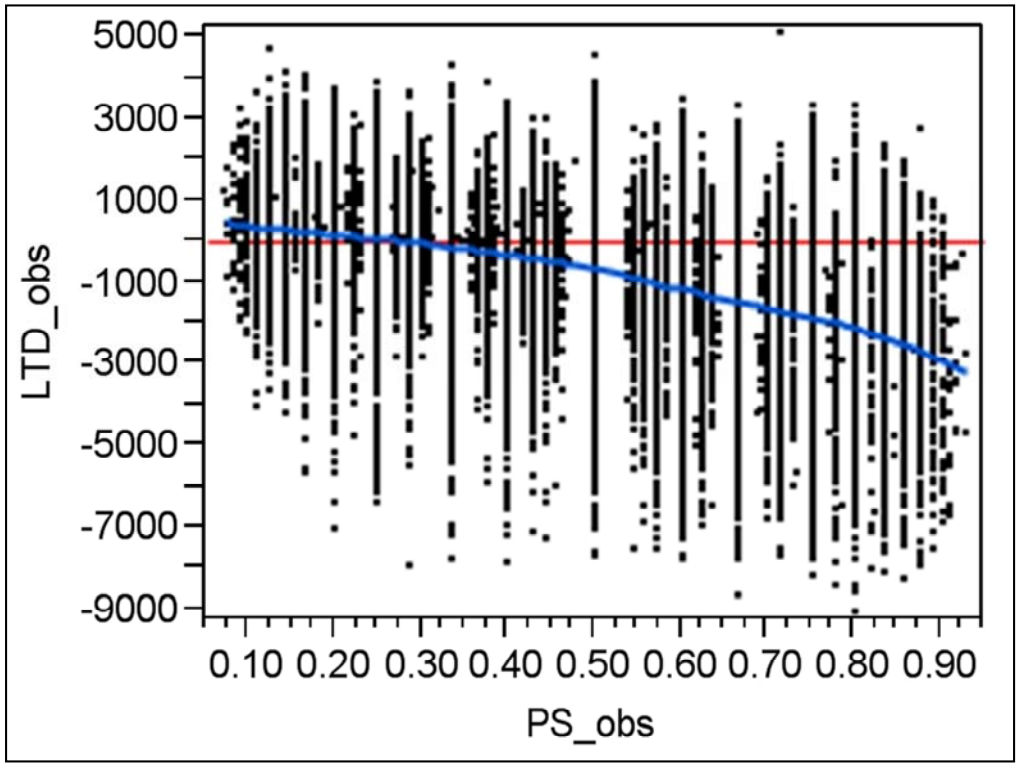
The R-square for this Nested ANOVA model is 91.6%

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.

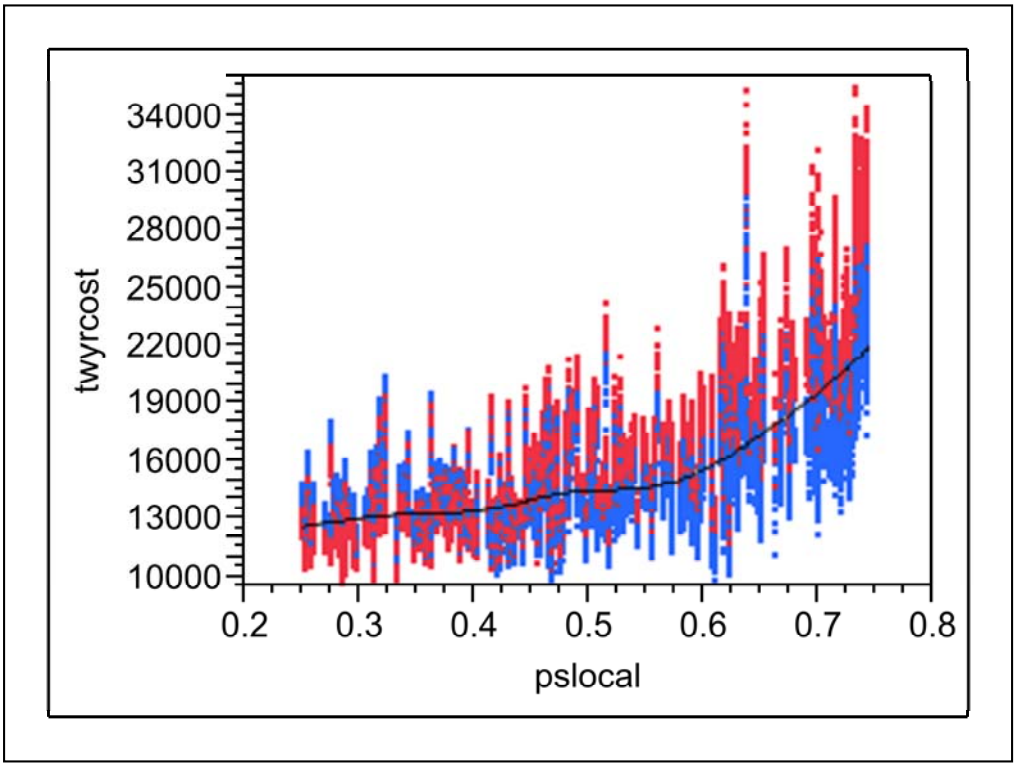
Due to the inverse relationship of Trtmfrac to true Propensity in our simulation, the spline fit (blue curve) above depicts a general tendency for observed LTDs to become more negative as observed propensity for trtm = 1 increases.

Nested ANOVA of True Costs

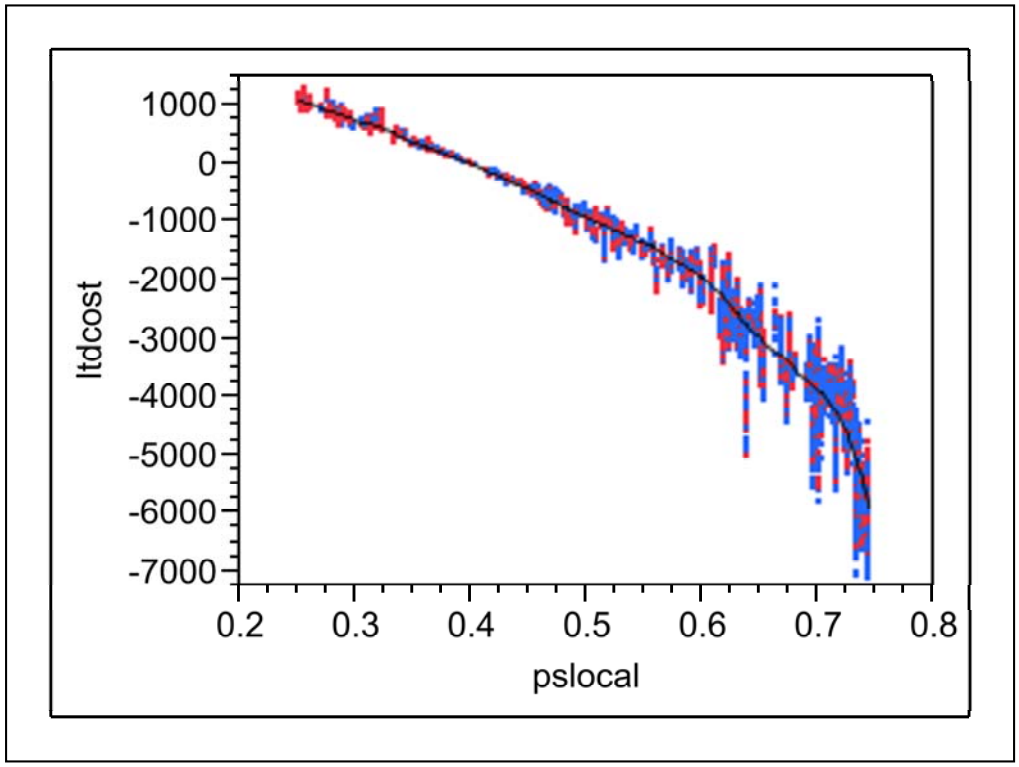
Source	DF	Sum-of-Squares	root Mean Square
Clusters (Exact Matches)	39,787	1.72e+12	$6,578 (LOAs)
Treatment within Cluster	35,038	1.14e+11	$1,805 (LTDs)
Error	175,132	0.1667	$0.0

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

So there is evidence of incredible SERENDIPITY here!!! And yet, the LTD estimates from 35,038 informative clusters formed via exact X-space matches are not exactly correct here. After all, the original 40,000 patients 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 24. 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 that of Slide 25.

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**.

Presentation Outline:

1. How were the data simulated?
2. **How were answers ranked?**
3. How good could answers be?
4. ~~How well did competitors do?~~

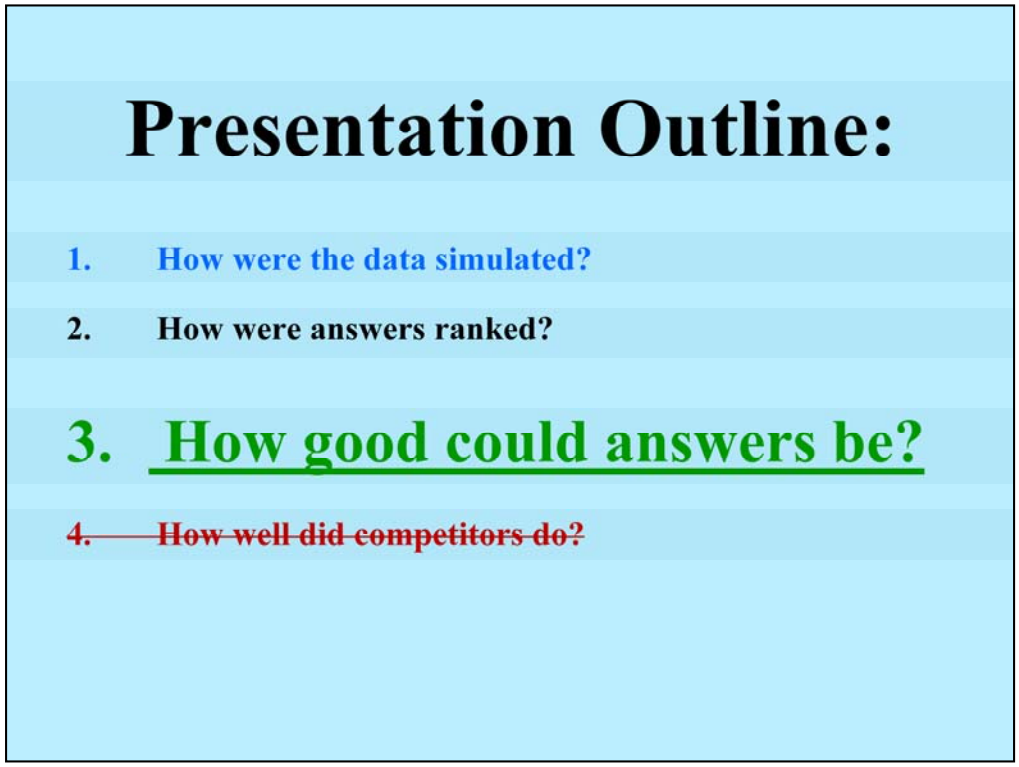
And now, PART TWO: The answer to this question is extremely simple!!!

root MSE Loss

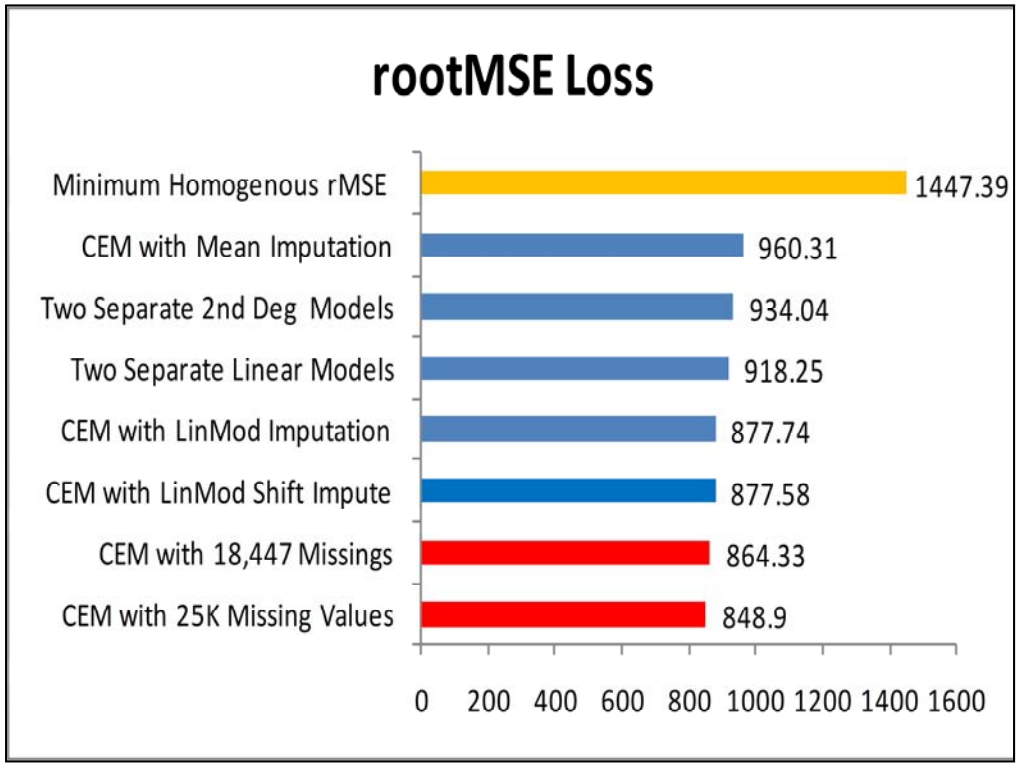
$$\sqrt{\text{Mean}[(\hat{\Delta}_i - \Delta_i)^2]}$$

Δ_i = True LTD for i^{th} Patient

= E(y | x, trtm=1) – E(y | x, trtm=0)



The presentation by Prof. Xiaochun Li of IU Biostatistics discusses some extremely simple ways to get relatively "good" estimates for all 249,958 patient-level LTD parameters. The next two slides provide a preview of the rMSE achieved by these methods.

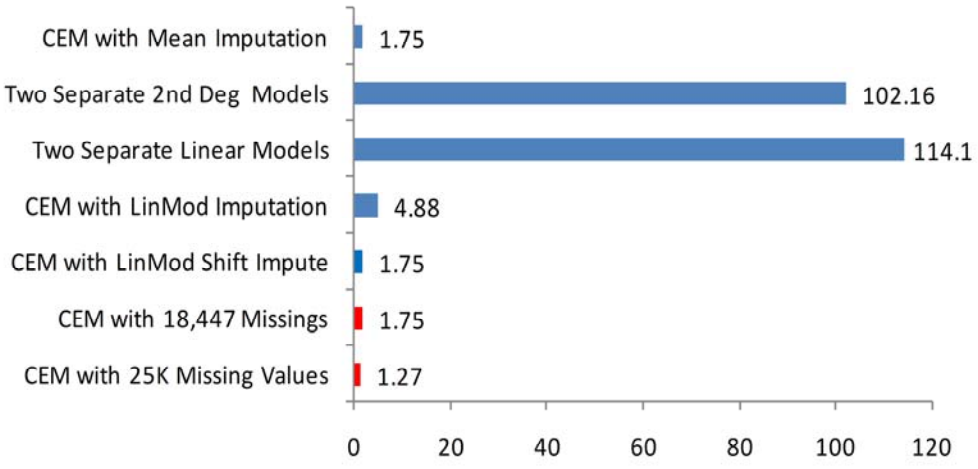


rMSE values are expressed in dollars ($)

The bottom seven rMSE values are remarkably good in the sense that the standard deviation of the additive white noise added to the "signal" from each patient was $1K.

In fact, the implied standard deviation of the difference in response between any two patients (e.g. observed y for a trtm = 1 patient minus observed y for a trtm = 0 patient) is sqrt(2) times $1K = $1,414.21.

True Main-Effect = $ –650.42



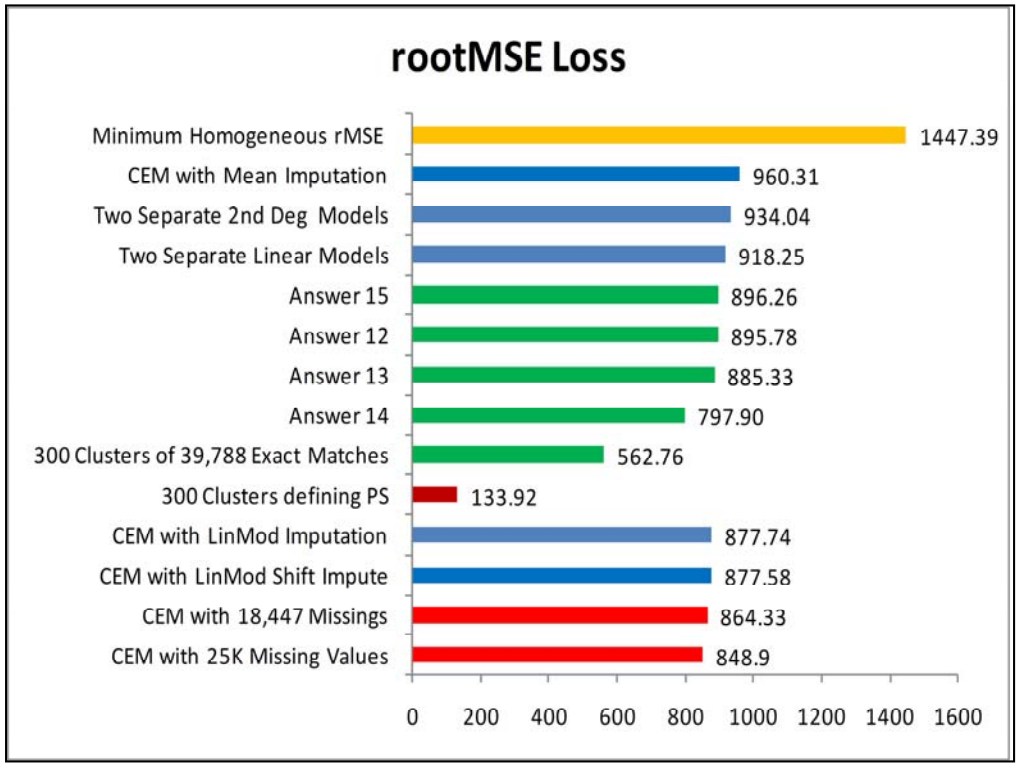
Presentation Outline:

1. How were the data simulated?
2. How were answers ranked?
3. How good could answers be?
4. ~~How well did competitors do?~~

Some Good Alternatives

All entries received had rMSE > $1,400, and these competitors remain anonymous.

However, I asked some colleagues how they would have analyzed the competition data if they had found time to enter. This generated the four answers (shown in GREEN and numbered 12 to 15) on the next slide. Needless to say, Answer 14 (rMSE = $797.90) was a true SHOCKER!



Answer 14 formed a **Trtm Classification Tree** via Recursive Partitioning with a minimum node size of 250 patients, resulting in 739 final leaf nodes. Subgroups formed this way achieve low rMSE = $797.90 loss by exploiting the strong (linear, exact) relationship between true Propensity Scores and the multiplicative "Trtmfrac" factor in the simulation (see Slide 19).

After thinking about how this could have happened, I realized that the "key" reasons are:

[a] the simulation used only 300 clusters to define the **linear PSlocal vs Trtmfrac relationship**, and

[b] while larger clusters are more biased than the small clusters use in the CEM approach, they would also be **much less variable** (subject to relatively high measurement error.)

Using the 300 "exactly correct" and LARGE clusters from the simulation yields rMSE = $133.92. However, this "original" clustering information was **not provided to competitors.**

The closest competitors could have come to using these "optimal subgroups" would apparently have been for them to hierarchically cluster only the 39,778 unique X-vectors (weighted equally) observed in the competition dataset. Using 300 subgroups would then yield rMSE = $562.76 . But, how could they have (correctly) guessed that roughly 300 was the right number of subgroups to use???

References

- **Rosenbaum PR, Rubin RB. The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika* 1983; 70: 41-55.**
- **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.
- **Iacus SM, King G, Porro G. CEM: Software for Coarsened Exact Matching. Version 1.0.142 www.r-project.org December 2009.**

36

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