

An Illustration of the "Local Control" Approach to Adjustment for Treatment Selection Bias and Confounding in Observational Studies.

Bob Obenchain, PhD, FASA
Risk Benefit Statistics LLC

Yin = Dark = Evil = Risk

Yang = Light = Good = Benefit



We use simulated data on 10,000 patients to illustrate all four phases of Local Control (LC) Analysis. The LC approach stresses robustness by using nested ANOVA (treatment within block) techniques rather than smooth, global models. Although not really Bayesian as averse to classical, LC methods do stress comparison of empirical distributions of Local Treatment Differences (LTDs) instead of calculation of p-values. The four phases of LC are: (1) initial exploration of feasibility and interpretability, (2) confirmation that LTD distributions are meaningfully different from the distributions resulting from random patient clusterings, (3) systematic sensitivity analyzes, and (4) identification of baseline patient characteristics predictive of differential treatment response.

History of Local Control Methods for Human Studies

- **Epidemiology (case-control & cohort) studies**
- **Post-stratification and re-weighting in surveys**
- **Dynamic randomization within each Strata (block) to improve balance on predictors of outcome**
- **Matching or Sub-grouping using Propensity Scores**
- **Econometric Instrumental Variables (LATEs)**
- **Marginal Structural Models (IPW \propto 1/PS)**
- **Unsupervised Propensity Scoring: Nested Treatment-within-Cluster ANOVA model ...with LATE, LTD and Error sources of variation**

Why are “Human Studies” being singled out here? Primarily, because human subjects can refuse to participate in designed experiments, and some designs are unethical on human subjects.

Local → make only the clearly more relevant comparisons.

Nested ANOVA

Source	Degrees-of-Freedom	Interpretation
Clusters (Subgroups)	$K = \text{Number of Clusters}$	Cluster Means are Local Average Treatment Effects (LATEs) when X's are Instrumental Variables (IVs)
Treatment within Cluster	Number of "Informative" Clusters $\leq K$	Local Treatment Differences (LTDs) are of interest for All Types of X-variables
Error	$\geq \text{Number of Patients} - 2K$	Uncertainty

Although a NESTED model can be (technically) **WRONG**, it is sufficiently versatile to almost always be **USEFUL** as the number of "clusters" increases.

Unsupervised PS => treatment indicator is NOT used to help define clusters.

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. Cluster means are plotted vertically along a horizontal axis describing within-cluster fraction treated (propensity score.)

The full DISTRIBUTION of Local Treatment Differences (LTDs) quantify the likely range of EFFECTS of Treatment.

The MAIN Effect of Treatment is the MEAN of this LTD distribution.

The “LSIM10K” dataset contains 10 simulated measurements on 10,325 hypothetical patients.

- [1] mort6mo : Binary 6-month mortality indicator.
- [2] cardcost : Cumulative 6-month cardiac related charges.
- [3] trtm : Binary indicator (1 => treated, 0 => untreated).
- [4] stent : Binary indicator (1 => coronary stent deployment.)
- [5] height : Patient height rounded to the nearest centimeter.
- [6] female : Binary sex indicator (1 => yes, 0 => male.)
- [7] diabetic : Binary indicator (1 => diabetes mellitus, 0 => no.)
- [8] acutemi : Binary indicator (1 => acute myocardial infarction within the previous 7 days, 0 => no.)
- [9] ejecfrac : Left ejection fraction % rounded to integer.
- [10] ves1proc : Number of vessels involved in initial PCI.

Treatment is a hypothetical blood thinning agent.

Treatment groups are NOT balanced of baseline characteristics (green.)

Only 2 of the 7 patient baseline x-characteristics is continuous. Ves1proc is ordinal with 6 levels.

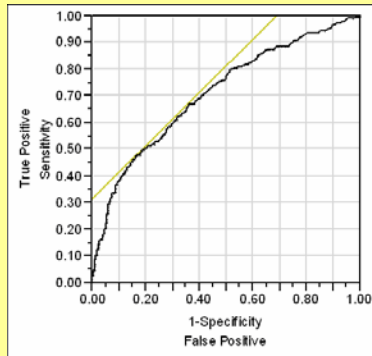
Logistic Regression

$y = \text{mort6mo}$

$t = \text{treatment}$

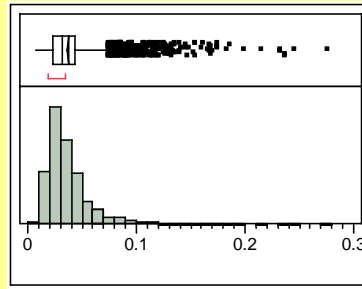
$X = \text{all 7 baseline vars}$

Area Under Curve = 0.709



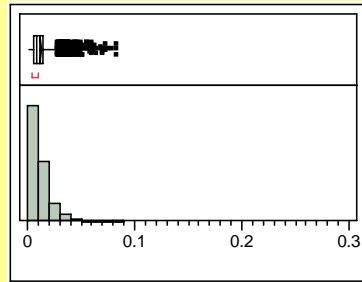
R-squared = 0.064

Untreated: Avg. Risk = .037



$n_0 =$
5,646

Treated: Avg. Risk = .012



$n_1 =$
4,679

66% reduction in risk of mortality due to treatment.

Equivalently, not being treated is 3-times riskier.

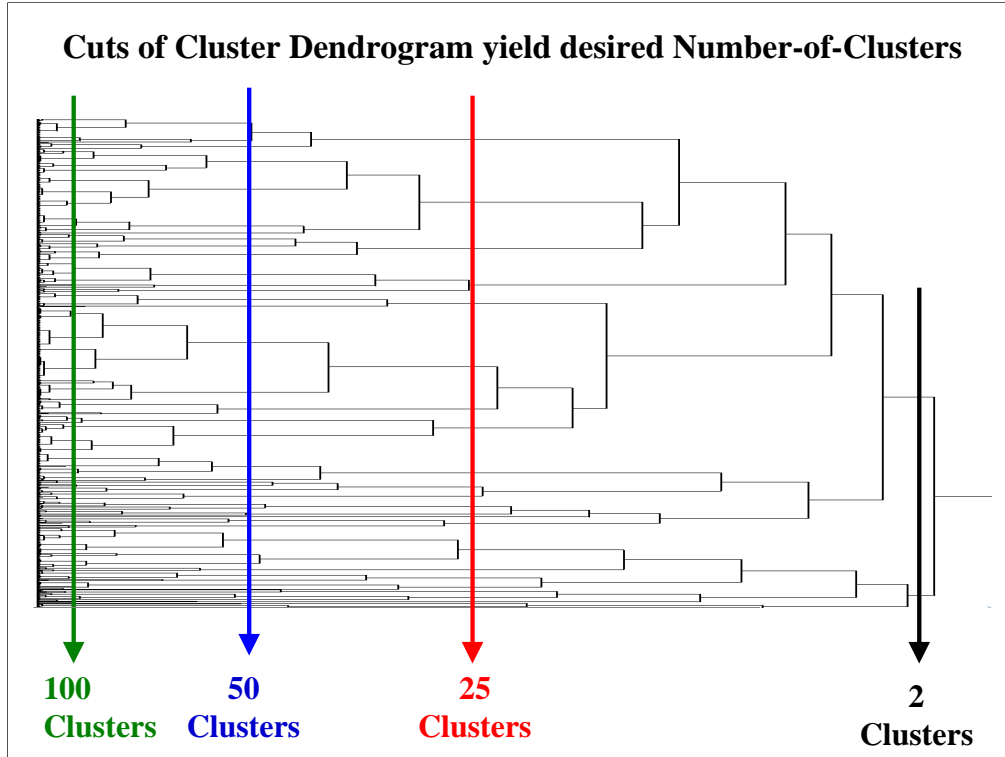
Main effect of treatment is highly significant, $p < .0001$.

(Treated – Untreated) Differences

	Pats	MortDiff (%)	Standard Error	CostDiff (\$)
	10,325	-2.50	0.30	255.08
ves1proc				
0	46	0.00	6.15	-2,582.06
1	7,265	-2.47	0.37	607.44
2	2,588	-2.70	0.53	-201.98
3	324	-4.64	1.51	2,841.52
4	91	0.00	0.00	1,345.69
5	11	0.00	0.00	-72.50

The Four “Tactical” Phases of Local Control Analysis

- 1. Does an Initial Exploration Demonstrate Feasibility and Interpretability?**
- 2. Is the Observed LTD Distribution Meaningfully Different from Random?**
- 3. Systematic Sensitivity Analyses !!!**
- 4. Which Patient Characteristics are Predictive of Differential Response?**



Today's presentation illustrates use of computing algorithms and graphical displays from JMP 7.0 from SAS Institute.

LC “Unbiasing” TRACE

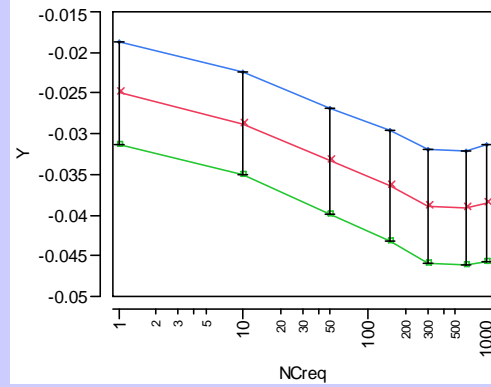
Y1LTD: Mortality within Six Months

Across Cluster Average LTD Outcome

Plus Two Sigma Upper Limit for Average LTD

Minus Two Sigma Lower Limit for Average LTD

NCreq = Number of Clusters Requested



NCreq	NCinfo	Y1 LTD	Local Std Err	Y1 Low Limit	Y1 Upr Limit
1	1	-0.0250	0.00313	-0.0313	-0.0188
10	10	-0.0288	0.00317	-0.0351	-0.0224
50	50	-0.0333	0.00326	-0.0399	-0.0268
150	149	-0.0364	0.00341	-0.0432	-0.0296
300	295	-0.0389	0.00350	-0.0459	-0.0319
600	557	-0.0391	0.00353	-0.0461	-0.0320
900	800	-0.0385	0.00358	-0.0457	-0.0314

LC “Unbiasing” TRACE

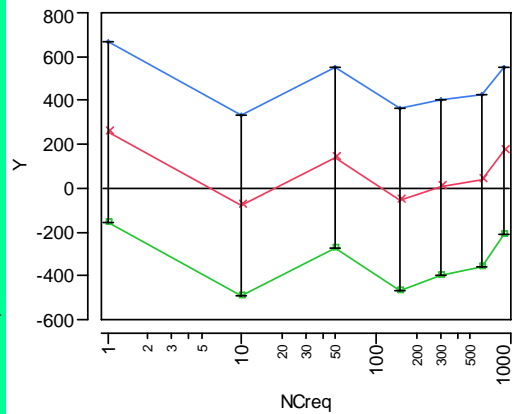
Y2LTD: Six Month Cardiac Costs

Across Cluster Average LTD Outcome

Plus Two Sigma Upper Limit for Average LTD

Minus Two Sigma Lower Limit for Average LTD

NCreq = Number of Clusters Requested

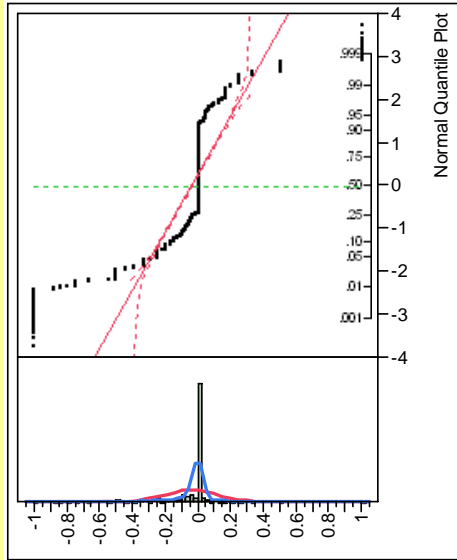


NCreq	NCinfo	Y2 LTD	Local Std Err	Y2 Low Limit	Y2 Upr Limit
1	1	255.08	206.21	-157.33	667.50
10	10	-79.17	207.63	-494.43	336.09
50	50	140.49	207.25	-274.01	554.99
150	149	-53.19	208.49	-470.18	363.79
300	295	3.42	199.27	-395.12	401.96
600	557	34.09	196.56	-359.04	427.21
900	800	168.74	190.61	-212.48	549.97

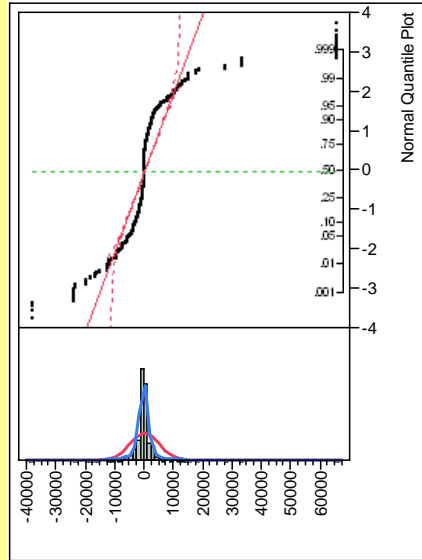
All results expressed in 1998 US Dollars (\$).

Observed LTD Distributions

mort6mo



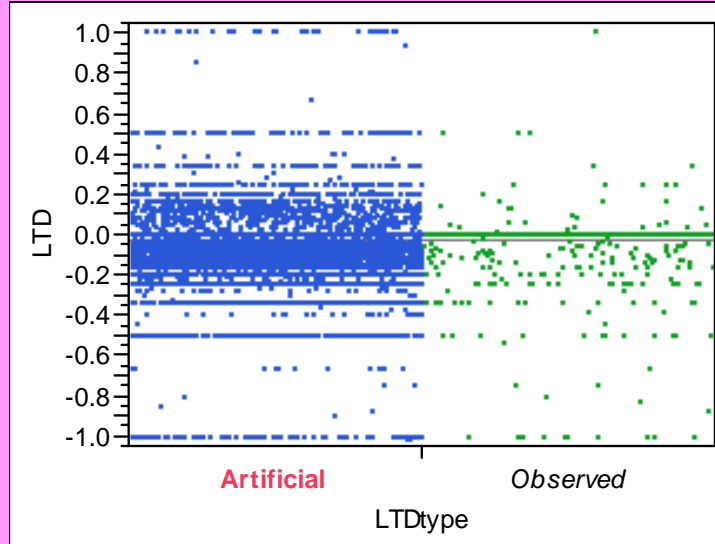
cardcost



LC Phase Two: Is the Observed LTD Distribution Meaningfully Different from Random?

- Simulate the “Artificial LTD distribution” by using 900 clusters of the same size with the same treatment fractions as the observed clusters; 100 clusters will be uninformative.
- Randomly distribute the observed outcomes (y) within these clusters, preserving Treated (t=1) vs Untreated (t=0) status but ignoring all other observed patient X-characteristics.
- Repeat this random process 25 times, yielding a sample of 25 x 800 = 20,000 Artificial LTDs.
- Compare the resulting Artificial LTD distribution with that from the 800 Observed LTDs.

JMP side-by-side comparison, using the “spread” option, of the artificial and observed LTD distributions for mort6mo.

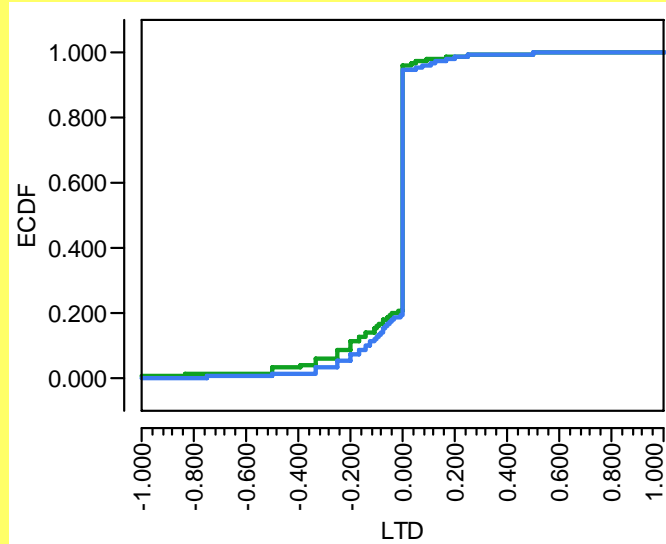


Comparisons of the Artificial (25 reps) and Observed LTD distributions for 800 informative clusters.

Unfortunately, only 304 distinct numerical values within 20,800 LTDs. Lots of zeros because mortalities are relatively rare.

Means: Artificial = -0.0247, Observed = -0.0469.

**Visual Comparison of “weighted” ECDFs
for the **observed** and **artificial** LTD
distributions of the mort6mo outcome.**



The observed LTD distribution (green) clearly has a “thinner” right-hand tail unfavorable to treatment than the artificial LTD distribution (blue.)

More importantly, the observed LTD distribution also clearly has a “thicker” left-hand tail favorable to treatment than the artificial LTD distribution.

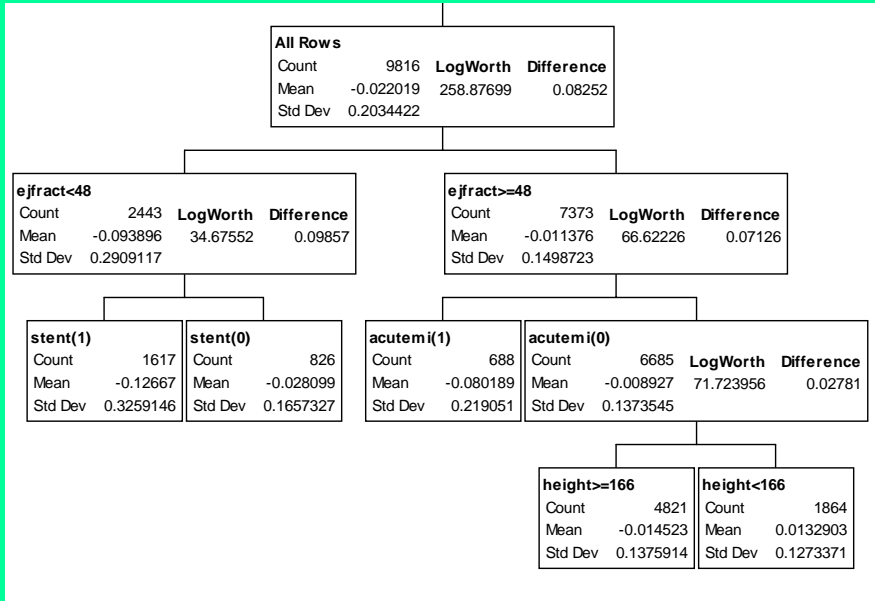
LC Phase Three: Systematic Sensitivity Analyses

- **No software currently available for this sort of “batch mode” processing**
- **One Interesting Alternative Clustering**

The three patient baseline x -characteristics that appear to be most predictive of t -treatment choice and the mortality y -outcome are stent, acutemi and ejfract. Thus, an LC analysis using only these three patient characteristics is parsimonious.

Interestingly, the resulting LC unbiasing trace for mort6mo is quite similar to that using all 7 x -characteristics, but the cardcost trace is much smoother!

JMP Partition Regression TREE for predicting mort6mo LTDs from Seven Baseline Patient x-Characteristics.



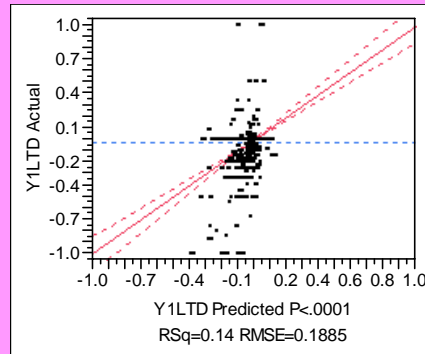
JMP Multivariable Model for predicting mort6mo LTDs from Seven Baseline Patient x -Characteristics (“effect screening” via Factorial to Degree Two)

Summary of Fit

RSquare	0.144396
RSquare Adj	0.141948
Root Mean Square Error	0.188451

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	28	58.65786	2.09492	58.9892
Error	9787	347.57256	0.03551	Prob > F
C. Total	9815	406.23042		<.0001



As usual, the above multivariable model for predicting mort6mo LTD described is rather complicated and, thus, not particularly easy to interpret or visualize.

Like the regression tree model above, this model has an R-squared statistic of only 14%, so it too has considerable lack-of-fit. For example, the mort6mo LTDs vary from -1.0 to $+1.0$, but the estimates from this model range only from -0.4 to $+0.14$ (see graphic.)

Furthermore, many terms are significant here primarily because the dataset contains so many observations (9,816 non-missing values of LTDs for patients within informative clusters.) Having such a large number of terms in the prediction equation greatly hampers use of such a model in practical applications, where an expected LTD needs to be computed for each individual patient.

SUMMARY

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

This example also dramatically illustrates that some of the most interesting effects of treatment are NOT main-effects. Evidence of patient differential response to treatment is always important ...even when it's not completely clear which patients will display which differences.

Back-Up Slides

What is LESS “coarse” than

$$\Pr(x, t | p) = \Pr(x | p) \Pr(t | p) ?$$

Conditioning upon *Cluster Membership* is intuitively somewhere between the two PS extremes in the limit as individual clusters become numerous, small and compact ...as long as *t* information is not used to form clusters

$$\Pr(x, t | C) \equiv \Pr(x | t, C) \Pr(t | C) \\ \approx \Pr(x | C) \Pr(t | C)$$

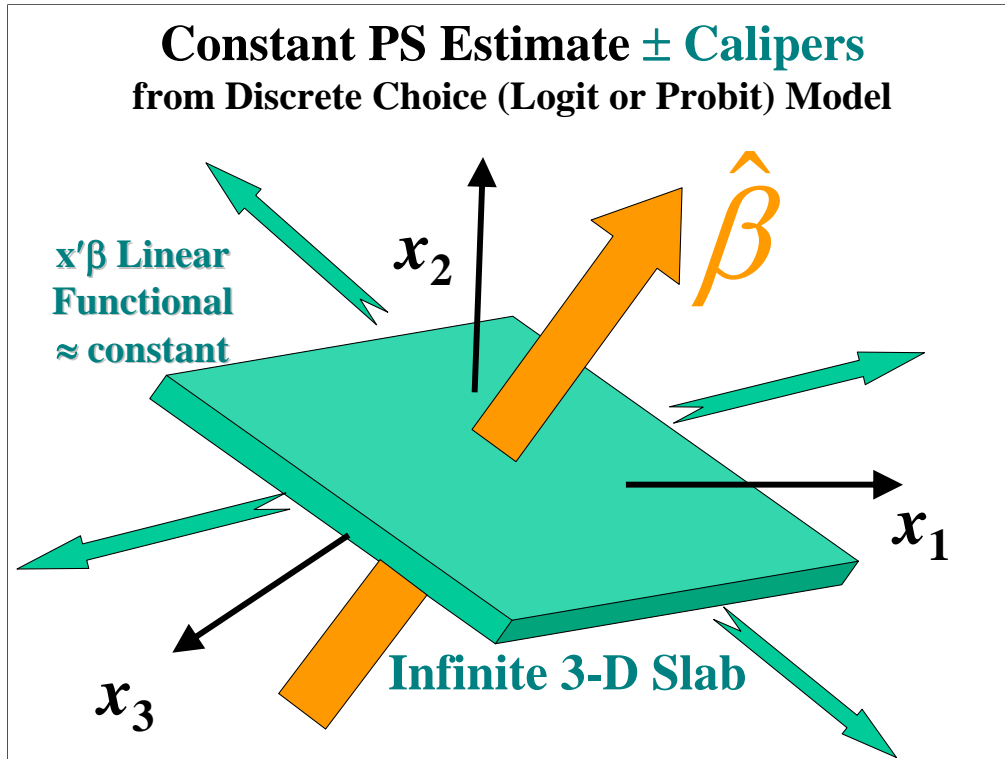
But LESS “detailed” than

$$\Pr(x, t) = \Pr(x) \Pr(t | x) ?$$

Here, we propose using (hierarchical) clustering to form numerous and compact (complete linkage) patient sub-groups.

The middle approximation is very poor when clusters are large; otherwise, PSs could not be the MOST COARSE balancing scores.

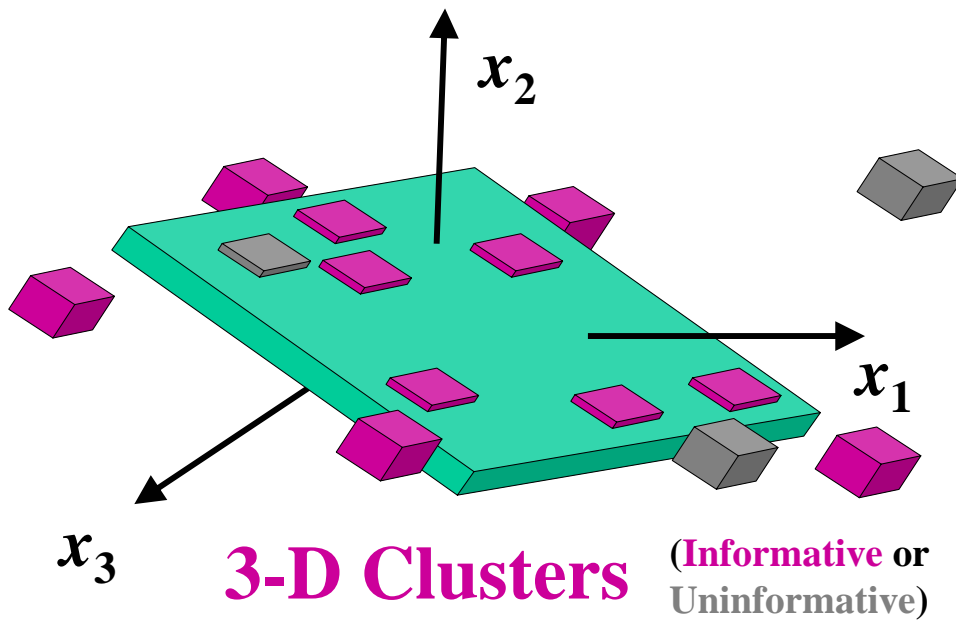
Clusters will still vary in SIZE relative to both [1] number of patients within the cluster and [2] X-space volume of the cluster.



Slab extends to plus/minus infinity in all directions orthogonal to beta-hat (2 dimensional space here.) Note that the slab has finite depth = (PS plus/minus Calipers) but has infinite volume.

Patients within this X-space slab could certainly have very different x_1 , x_2 and x_3 coordinates. Thus no balance on x-factors is automatic.

Unsupervised → No PS Estimates Needed



A cluster is “Informative” when it contains at least one patient from each treatment group.

Local Treatment Differences (LTDs) in outcomes can then be computed.

Observed Treatment Fractions within Clusters are Local, non-parametric PS estimates.

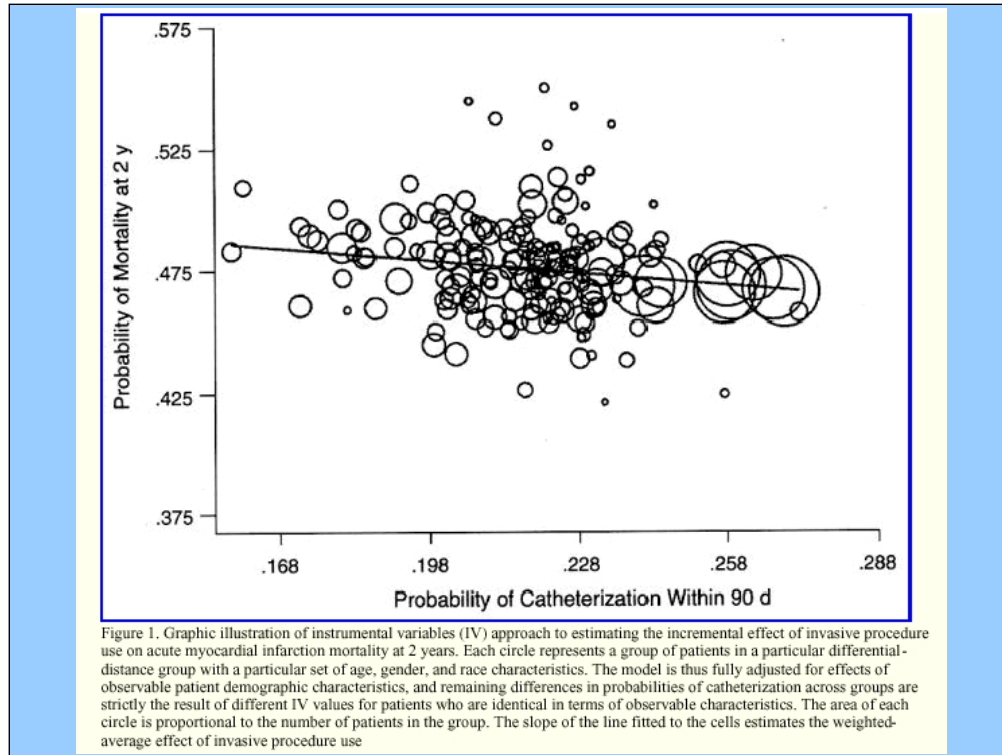
**Conditioning on
Propensity Scores implies both...**

**Blocking: Same Local X-covariate
distributions for both treatments**

and

**Balancing: Equal local treatment
fractions when $\Pr(t | p) = p = 0.5$**

Because $\Pr(t|x)$ is unknown in most cases, not only does $\Pr(t|p)$ need to be estimated but also balance needs to be checked / verified.



NOTE: Mark McClellan was recently commissioner of the FDA and in charge of CMS. Ergo, this is now a “classic” paper!!!

N = 205K 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 proved to be likely to receive an (expensive) invasive procedure (whether they really needed 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.

Nested ANOVA

Source	Degrees-of-Freedom	Interpretation
Clusters (Subgroups)	K = Number of Clusters	Cluster Means are Local Average Treatment Effects (LATEs) when X's are Instrumental Variables (IVs)
Treatment within Cluster	Number of "Informative" Clusters $\leq K$	Local Treatment Differences (LTDs) are of interest for All Types of X-variables
Error	\geq Number of Patients $- 2K$	Uncertainty

Although a NESTED model can be (technically)
WRONG, it is sufficiently versatile to almost always be
USEFUL as the number of "clusters" increases.

When X-covariates measure disease severity and/or patient frailty, they are usually predictive of both treatment selection (especially when expensive) and ultimate outcome. In this case, cluster means from a nested model are totally confounded and "K" degrees-of-freedom are immediately lost. But within cluster treatment differences are ALWAYS relevant and become more-and-more relevant as number of clusters increases and, thus, sizes of clusters decrease.

Notation for Variables

- y = observed outcome variable(s)
- x = observed baseline covariate(s)
- t = observed treatment assignment
(usually non-random)
- z = unobserved explanatory variable(s)

Z variables (unmeasured confounders) provide unknown, causal effects on outcomes, Y. In statistical/econometric models, existence of Z variables (as well as uncertainty in measuring Y and X variables) necessitate inclusion of "error terms."

Today, the patient's genome is mostly a Z variable; some day soon, more and more of this sort of information may become routine X variables.

Food and Drug Administration Meeting 9/21/2006, Afternoon Session, Page 155

DR. DeMETS: **We're going to have to rely on observational data ...to get some further information. We will not have randomized trials of long duration for rare events.** But today's discussion has demonstrated just **how big a challenge that is.**

And we've heard over and over again (that) the challenge (is) in the analysis. I mean, **"It's in the analysis stupid,"** is sort of the bottom line. **It's very tricky stuff, and it's very hard to do.**

...As we look at these trials or these kind of data in the future, **we're going to have to really drill down on the analysis details a lot more than we do in, say, randomized trials.**

Nick Freemantle and Alar Irs

Observational evidence for determining drug safety

Is no substitute for evidence from randomised controlled trials

BMJ 2008; 336: 627-628.

Multivariable methods may help but inclusion of (variables) does not fully account for their effects. Identified risk factors and their inter-relations are often complex and statistical models deal with these only crudely and partially. Thus a statistical model may include hypertension as just a “yes or no” dichotomy... In addition, the effects of raised blood pressure may differ in people who do, or do not, have other risk factors in ways that cannot be readily dealt with in (these models.) In any case, the multivariable analysis simply estimates the effects of known risk factors from the available data, and the best estimate available from the data may not be a very good one.

Highly Influential ???

D'Agostino RB Jr. Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. [TEACHER'S CORNER] Stat Med 1998; 17: 2265–2281.

Claimed that 3rd form of PS Adjustment (after matching and sub-grouping) was to simply use some function of PS estimates as an additional X in Covariate Adjustment.

Other researchers were doing this (as a short cut) before this tutorial was published. Afterwards, almost everybody was doing it.

Don Rubin says the 70% to 85% of publisher PS Applications are garbage!!!

With titles like these, does one really need to read the paper?

Shah BR, Laupacis A, Hux JE, Austin PC. Propensity score methods gave similar results to traditional regression modeling in observational studies: a systematic review. J Clin Epidemiol 2005; 58: 550–559.

Stürmer T, Joshi M, Glynn RJ, Avorn J, Rothman KJ, Schneeweiss S. A review of the application of propensity score methods yielded increasing use, advantages in specific settings, but not substantially different estimates compared with conventional multivariable methods. [REVIEW ARTICLE] J Clin Epidemiol 2006; 59: 437–447.

Two recent review articles on analysis of observational data.

References

Bang H, Robins JM. Doubly Robust Estimation in Missing Data and Causal Inference Models. *Biometrics* 2005; 61: 962-972.

Fraley C, Raftery AE. Model based clustering, discriminant analysis and density estimation. *JASA* 2002; 97: 611-631.

Imbens GW, Angrist JD. Identification and Estimation of Local Average Treatment Effects. *Econometrica* 1994; 62: 467-475.

McClellan M, McNeil BJ, Newhouse JP. Does More Intensive Treatment of Myocardial Infarction in the Elderly Reduce Mortality?: Analysis Using Instrumental Variables. *JAMA* 1994; 272: 859-866.

McEntegart D. “The Pursuit of Balance Using Stratified and Dynamic Randomization Techniques: An Overview.” *Drug Information Journal* 2003; 37: 293-308.

References ...concluded

Obenchain RL. USPS package: Unsupervised and Supervised Propensity Scoring in R. Version 1.1-0. www.r-project.org August 2007.

Obenchain RL. The “Local Control” Approach to Adjustment for Treatment Selection Bias and Confounding (illustrated with JMP Scripts). *Observational Studies*. Cary, NC: SAS Press. 2009.

Robins JM, Hernan MA, Brumback B. Marginal Structural Models and Causal Inference in Epidemiology. *Epidemiology* 2000; 11: 550-560.

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

Rosenbaum PR. *Observational Studies, Second Edition*. New York: Springer-Verlag 2002.