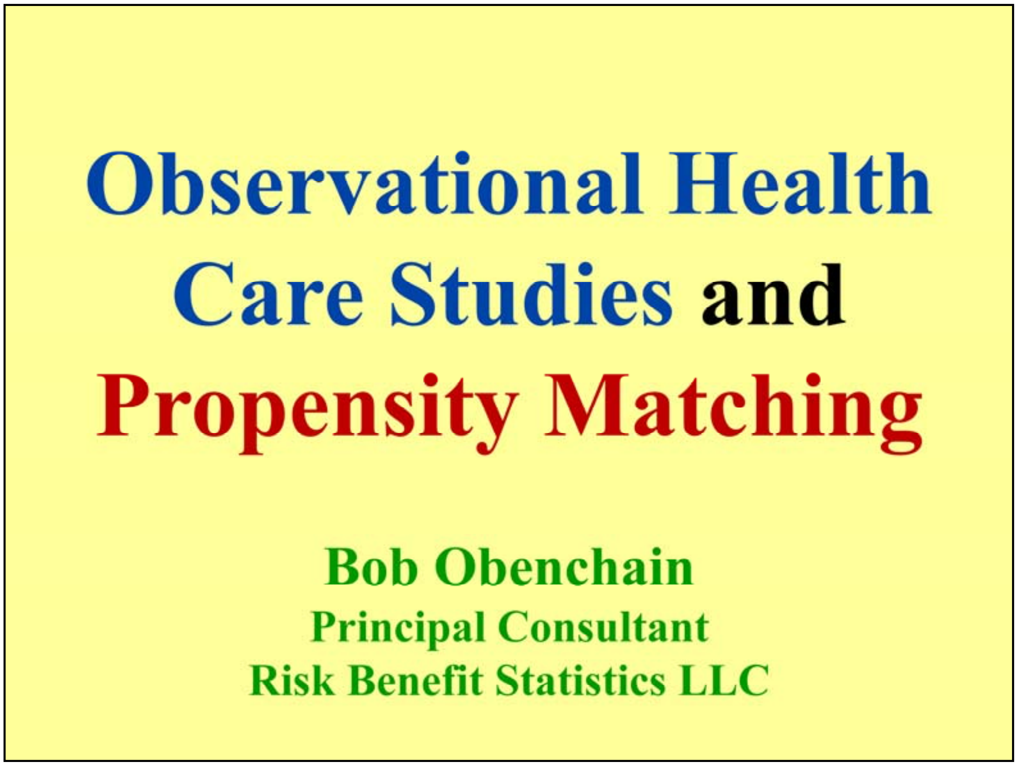


Observational Health Care Studies and Propensity Matching

Bob Obenchain
Principal Consultant
Risk Benefit Statistics LLC

We focus on some basic concepts and statistical methods that make me optimistic about the future of observational studies in health care. The majority of the information ultimately needed to improve the effectiveness of health care delivery in the US is finally being routinely captured and stored electronically. The key step is to avoid being fooled by well known sources of bias in observational data; traps that every outcomes researcher needs to be constantly on guard against. Thus we focus on use of propensity matching techniques and show, using a simple 2 x 2 table, that patient matching can address and resolve Simpson's Paradox.

New AHRQ Terminology:

Observational

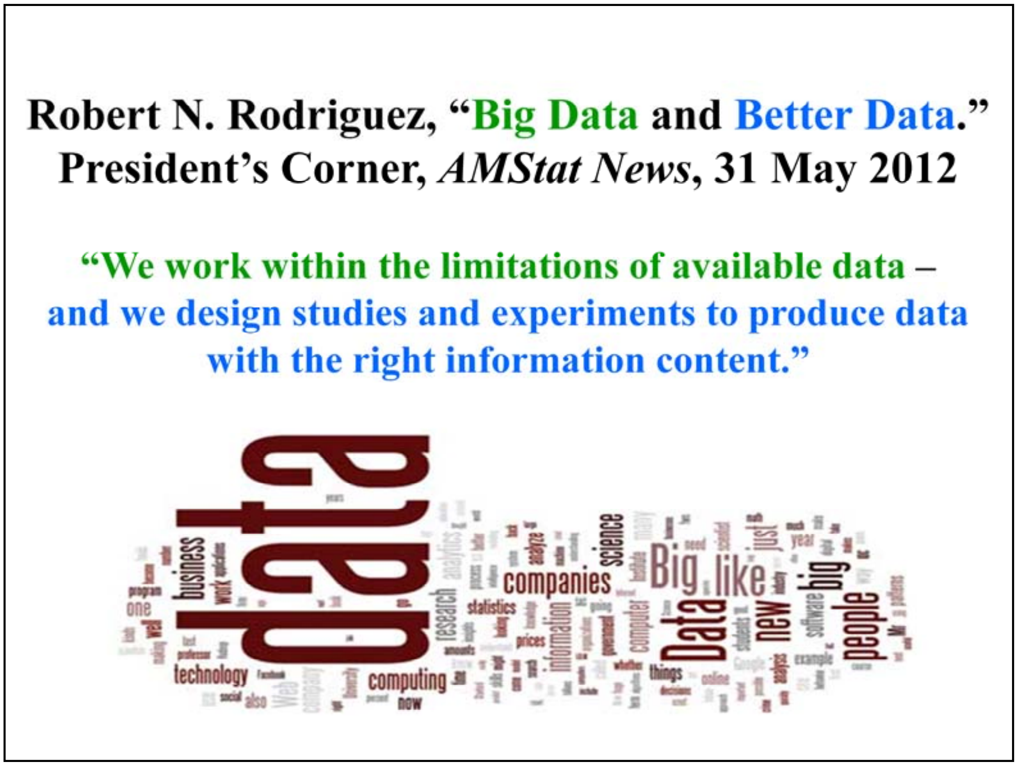
Comparative Effectiveness Research

(OCER)

OCER = AHRQ umbrella terminology for use of Observational data on actual Health Care practice to compare alternative treatments and approaches.

I will start by giving some background information intended to help motivate the need for specialized statistical analysis methodologies to address the unique challenges posed by this new and fundamentally different health outcomes research context.

AHRQ [Draft] Publication, 2012. Developing a Protocol for Observational Comparative Effectiveness Research (OCER): A User's Guide. Rockville, MD: Agency for Healthcare Research and Quality. in press.



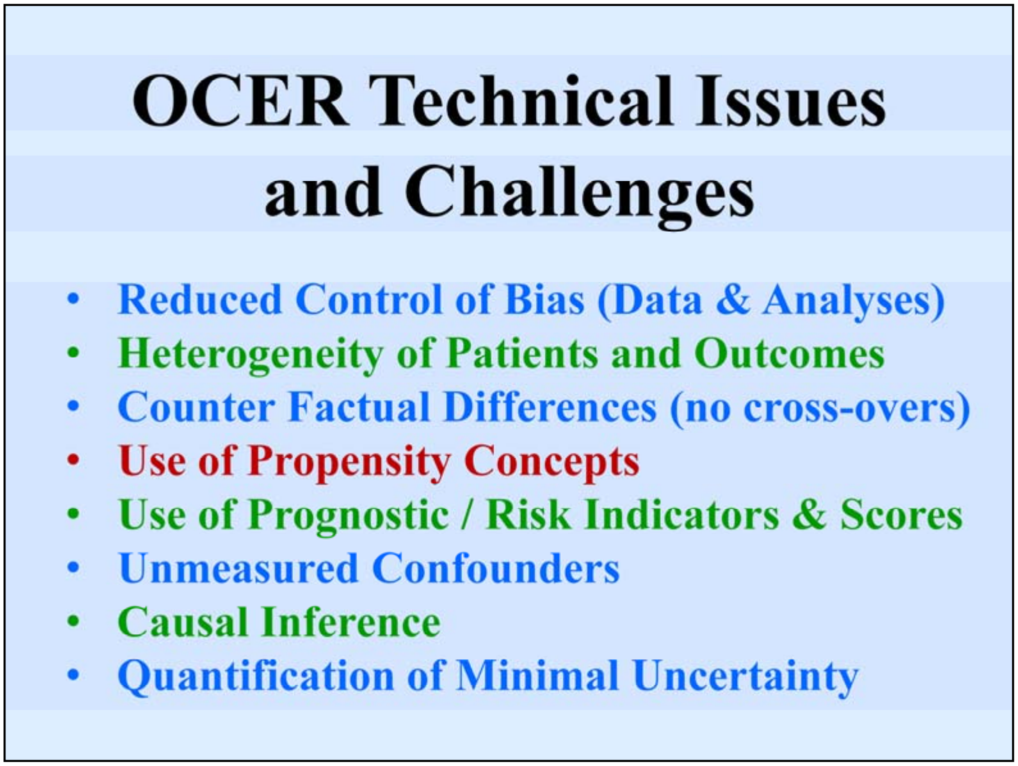
Robert N. Rodriguez, "Big Data and Better Data." President's Corner, *AMStat News*, 31 May 2012

"We work within the limitations of available data – and we design studies and experiments to produce data with the right information content."

data
Big
Data
like
companies
people
new
big
business
technology
computing
science
research
information
computer
statistics
software
analytics
things
now
also
one
just
year
prices
example
online
social
Web
company

Big OCER datasets are widely considered to be "inferior" sources of information.

"Designing" and/or adopting "Analysis Protocols" for Observational Research could go a long way towards preventing currently "abusive" practices (not adjusting for multiple testing, unreported multiple modeling, etc.) but cannot change fundamental realities.



OCER Technical Issues and Challenges

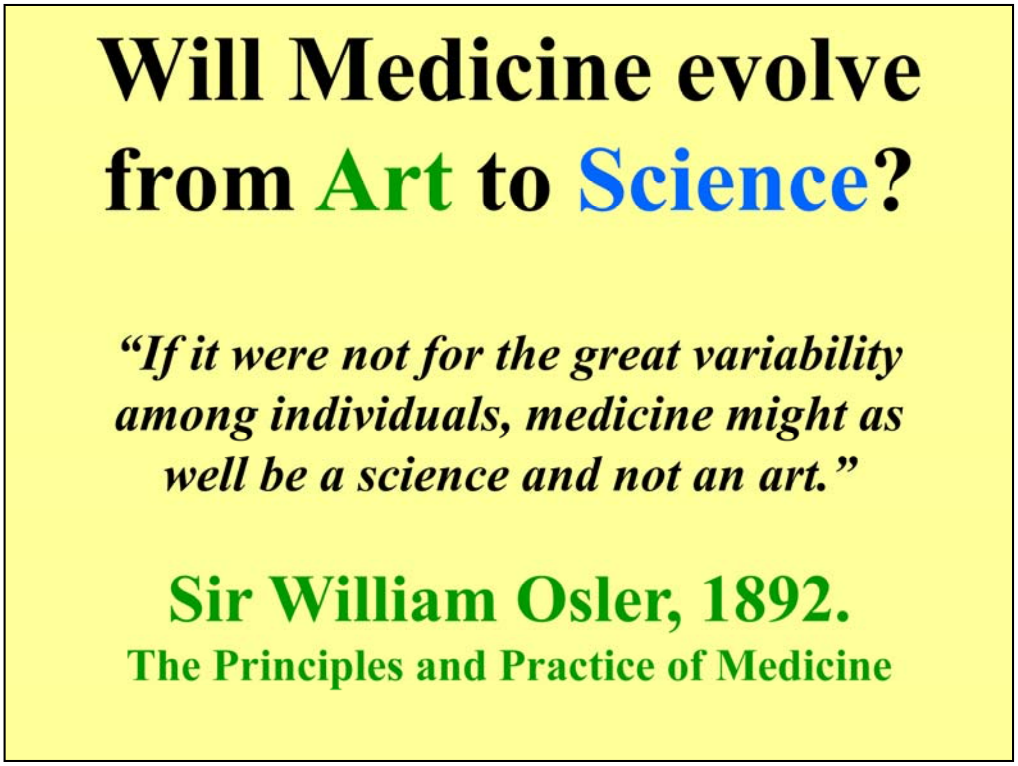
- Reduced Control of Bias (Data & Analyses)
- Heterogeneity of Patients and Outcomes
- Counter Factual Differences (no cross-overs)
- Use of Propensity Concepts
- Use of Prognostic / Risk Indicators & Scores
- Unmeasured Confounders
- Causal Inference
- Quantification of Minimal Uncertainty

Detecting Heterogeneity is much more Science than Art.

CFD => We can’t simply take differences at the individual patient level …as in a paired t-test. However, starting off by forming differences at the lowest available level will be our key message today.

Much of the talk will be about Propensity and/or Prognostic (Risk) concepts.

The final three bullets concern topics we will not have time to cover in any detail today.



Will Medicine evolve from Art to Science?

"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

Heterogeneous patients respond heterogeneously to treatment. One size (one choice) definitely does not fit all.

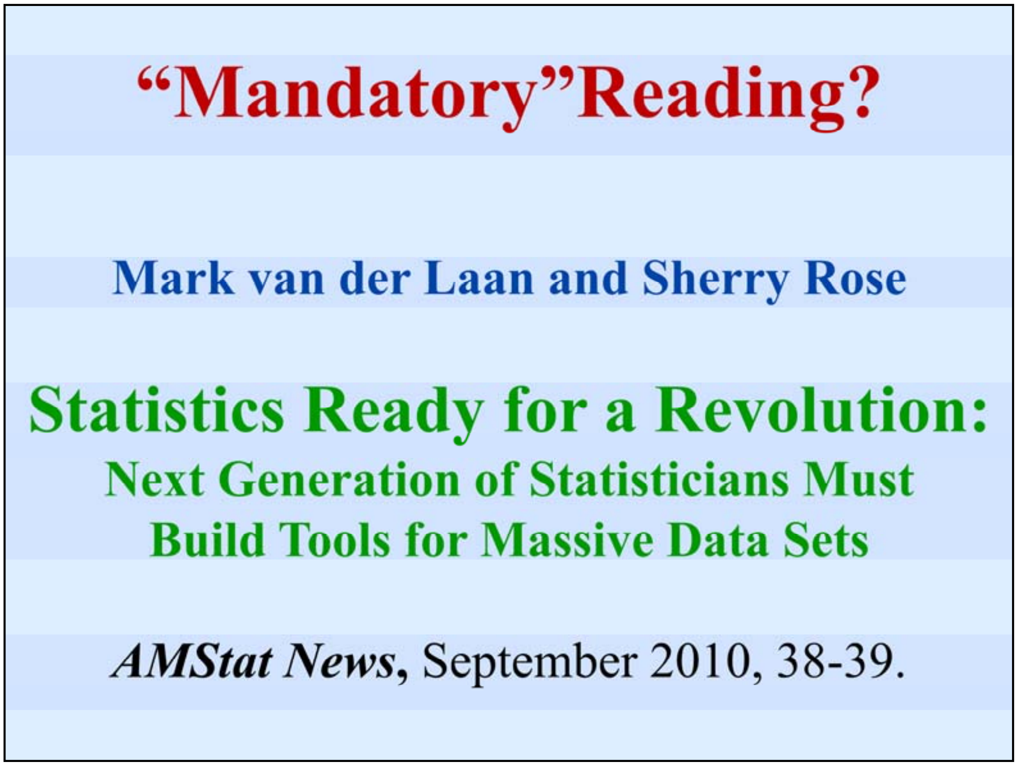
Most professionals passionately involved in health care practice and policy don't really know very much about experimental design or data analysis methodology. In fact, they probably don't realize that the health care study methods still being used today are completely inadequate to meet their health care information needs …needs that better match those of their patients.

Forward looking statisticians and epidemiologists should be actively (perhaps, passionately) working to discourage continuing use of traditional clinical trial methods that focus, exclusively, on main-effects of treatments. These are the "overall average" effects that can be (easily) measured most precisely. Unfortunately, high precision does not always translate into high accurately. And, even worse, this over-simplification can yield a false impression that uncertainty in study findings is much lower than it actually is. In other words, traditional clinical trial methods may be providing relatively "good" answers …but only to the "wrong" questions.

The recent federal Patient Centered Outcomes Research (PCOR) initiative is certainly encouraging researchers to take initial steps towards "individualized medicine." The canonical PCOR question is: "What works for patients like me?"

With the advent of "Big Data" from administrative claims, patient registries and electronic medical records, isn't it about time for health care researchers to finally, deliberately address real issues by quantifying "heterogeneous patient response"?

After all, doctors have been waiting for this for at least 120 years!



“Mandatory”Reading?

Mark van der Laan and Sherry Rose

Statistics Ready for a Revolution:

Next Generation of Statisticians Must Build Tools for Massive Data Sets

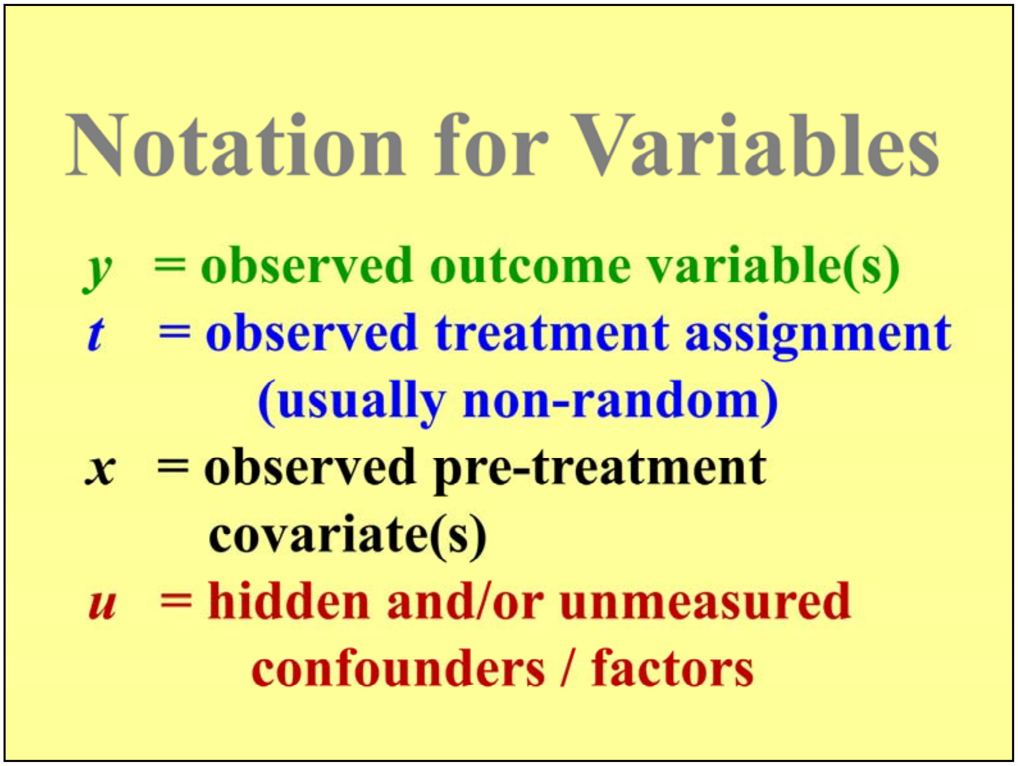
AMStat News, September 2010, 38-39.

“Big Data” on Health Care are coming, and traditional statistical methods used in clinical trials will be of no help. After all, traditional “covariate adjustment” methods and their ubiquitous p-values sprang up almost 100 years ago and were, perhaps, ideal for the relatively small samples then collected via well designed and controlled experiments and analyzed, essentially by hand.

Big data will be “dense” in all regions of widespread interest. There will be no need to make strong (but potentially quite wrong) assumptions and no need to interpolate between or extrapolate beyond just a few, sparse data points. If allowed to, big data will be literally capable of speaking, quite objectively, for themselves.

To prepare for this eventuality, I recommend that statisticians, epidemiologists and technically-inclined health outcomes researchers start by reading (and periodically re-reading) the above key 2-page article.

I am convicted that powerful “new” approaches will be based on incredibly simple and easily appreciated concepts, like patient matching (post-hoc blocking on pre-treatment characteristics). To avoid insertion of personal opinions or prejudices, calculations and graphical displays will be performed by “expert systems” that implement computer-intensive systematic sensitivity analyses, thereby revealing the “minimal uncertainty” in big data.



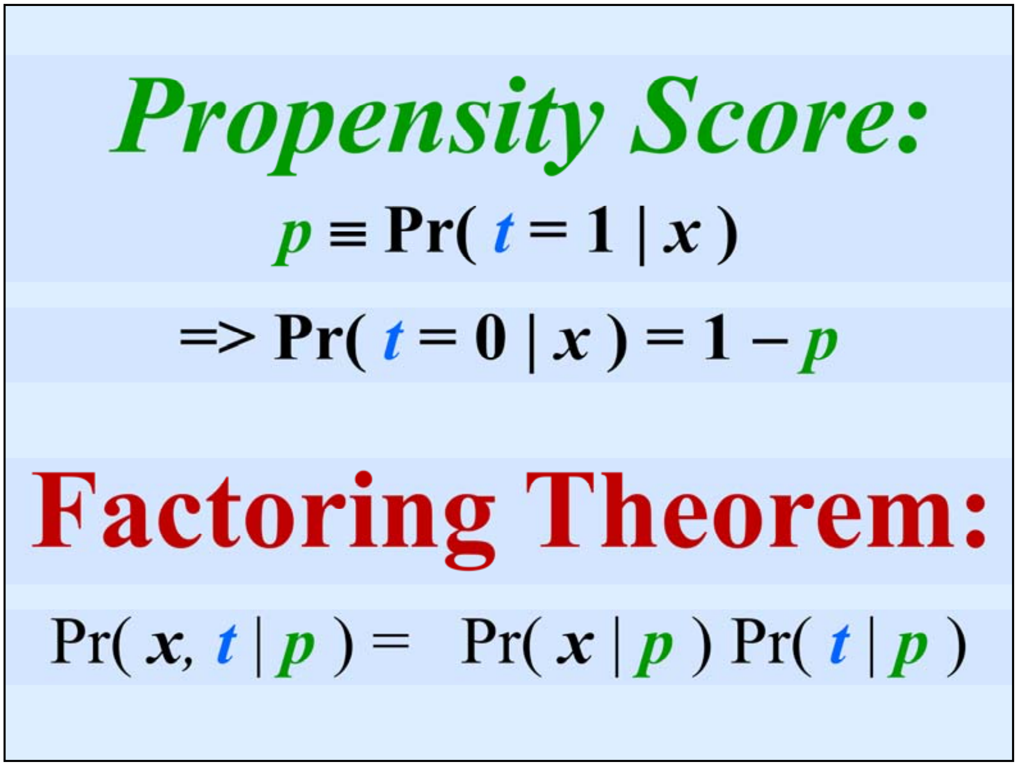
Notation for Variables

y = observed outcome variable(s)

t = observed treatment assignment (usually non-random)

x = observed pre-treatment covariate(s)

u = hidden and/or unmeasured confounders / factors



Propensity Score:

$$p \equiv \Pr(\, t = 1 \mid x \,)$$

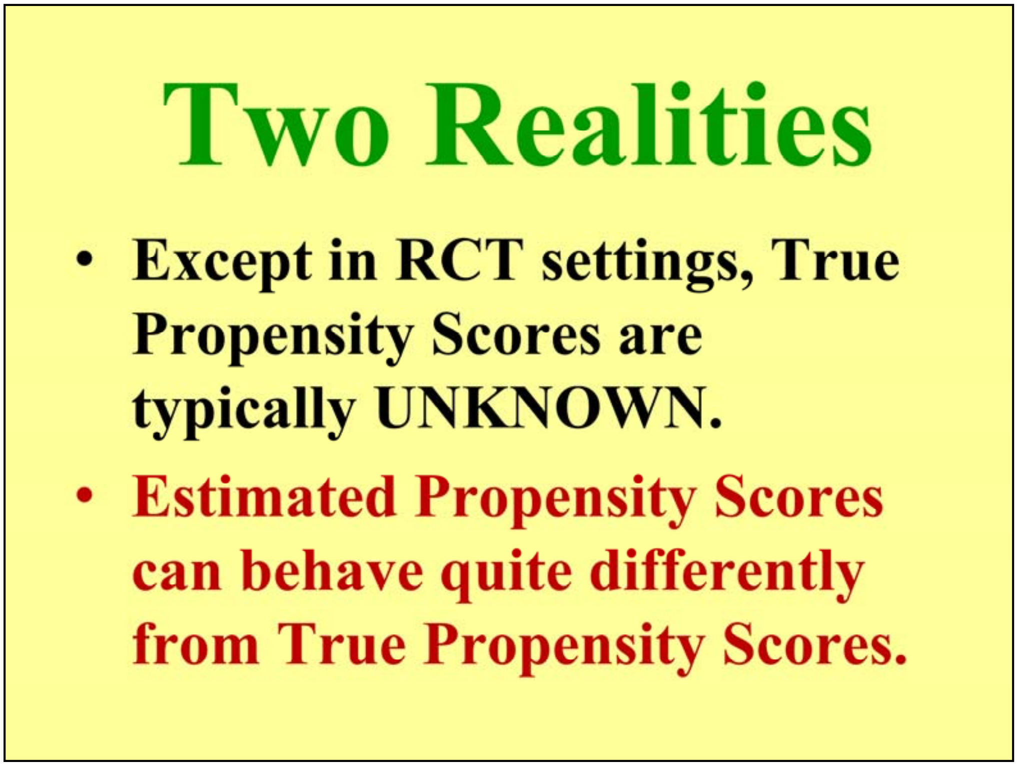
$$\Rightarrow \Pr(\, t = 0 \mid x \,) = 1 - p$$

Factoring Theorem:

$$\Pr(\, x, t \mid p \,) = \Pr(\, x \mid p \,) \Pr(\, t \mid p \,)$$

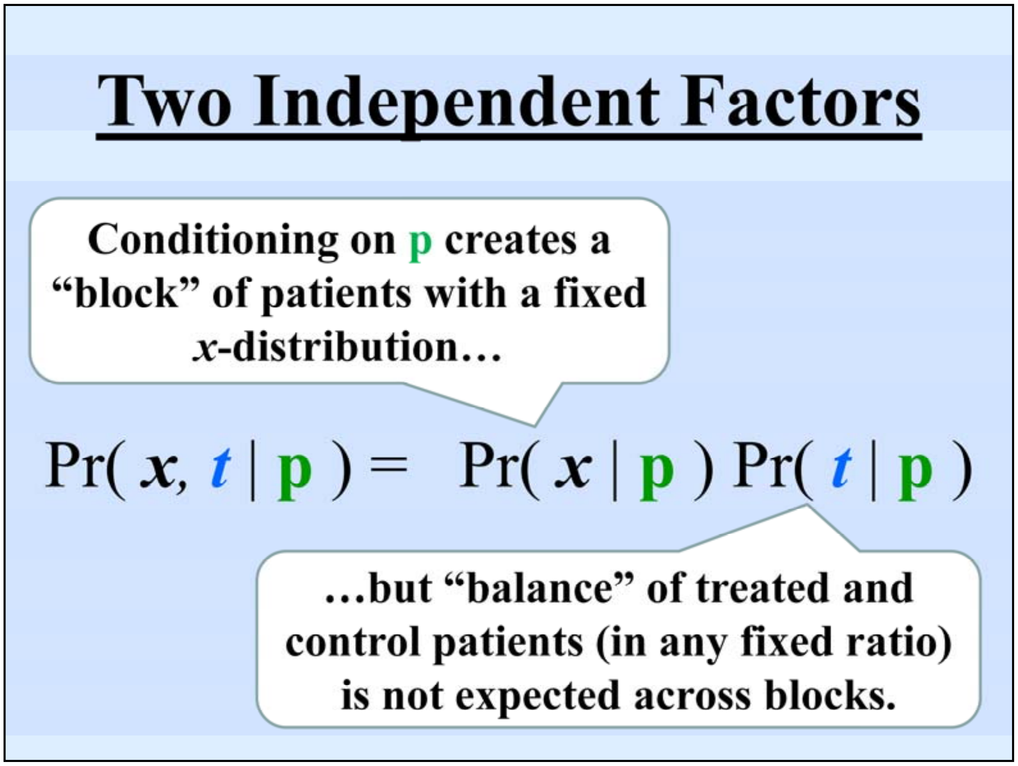
In words, the factoring theorem states that the joint conditional distribution of ***x*** and ***t*** given the true **p** must necessarily factor into the product of the conditional distribution of ***x*** given **p** and the conditional distribution of ***t*** given **p**. In statistics and probability theory, this factoring has profound implications, of course. The distribution of baseline patient ***x***-characteristics has thereby been shown to be *statistically independent* of the distribution of treatment choice, where both distributions are *conditional upon the given numerical value of* **p**.

The highly simplified notation used here is not intended to imply that only cases involving discrete variables are being addressed. Some (or all) of the component variables in the **x** vector may be continuous, and **p** is continuous. Measure theoretic details are being ignored here, just as they were in the original publication by Rosenbaum and Rubin in Biometrika in 1983.



Two Realities

- Except in RCT settings, True Propensity Scores are typically UNKNOWN.
- Estimated Propensity Scores can behave quite differently from True Propensity Scores.



Two Independent Factors

Conditioning on p creates a “block” of patients with a fixed *x*-distribution…

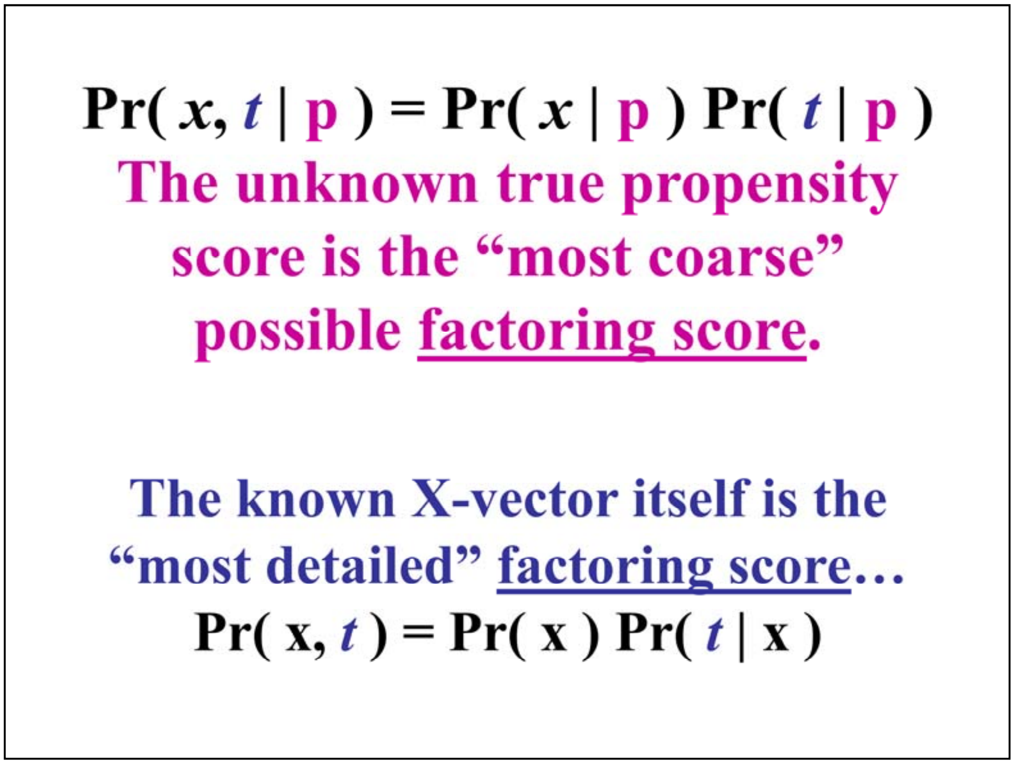
$$\Pr(\, \boldsymbol{x},\ \boldsymbol{t} \mid \mathbf{p}\,) = \Pr(\, \boldsymbol{x} \mid \mathbf{p}\,)\ \Pr(\, \boldsymbol{t} \mid \mathbf{p}\,)$$

…but “balance” of treated and control patients (in any fixed ratio) is not expected across blocks.

The true Propensity Score value, p, usually varies from block to block, so this value is usually NOT one-half.

In traditional Design-of-Experiments terminology, “blocking” and “balancing” are two very different concepts, with “blocking” rather clearly being the much more fundamental and important concept (either Number 1 or else Number 2 behind “randomization.”)

Achieving better “balance” typically means trying to make the X’X matrix more nearly block-diagonal, so that statistical test statistics will be closer to being uncorrelated …or nearly “clean.”



$\mathrm{Pr}(\,x, t \mid \mathrm{p}\,) = \mathrm{Pr}(\,x \mid \mathrm{p}\,)\ \mathrm{Pr}(\,t \mid \mathrm{p}\,)$

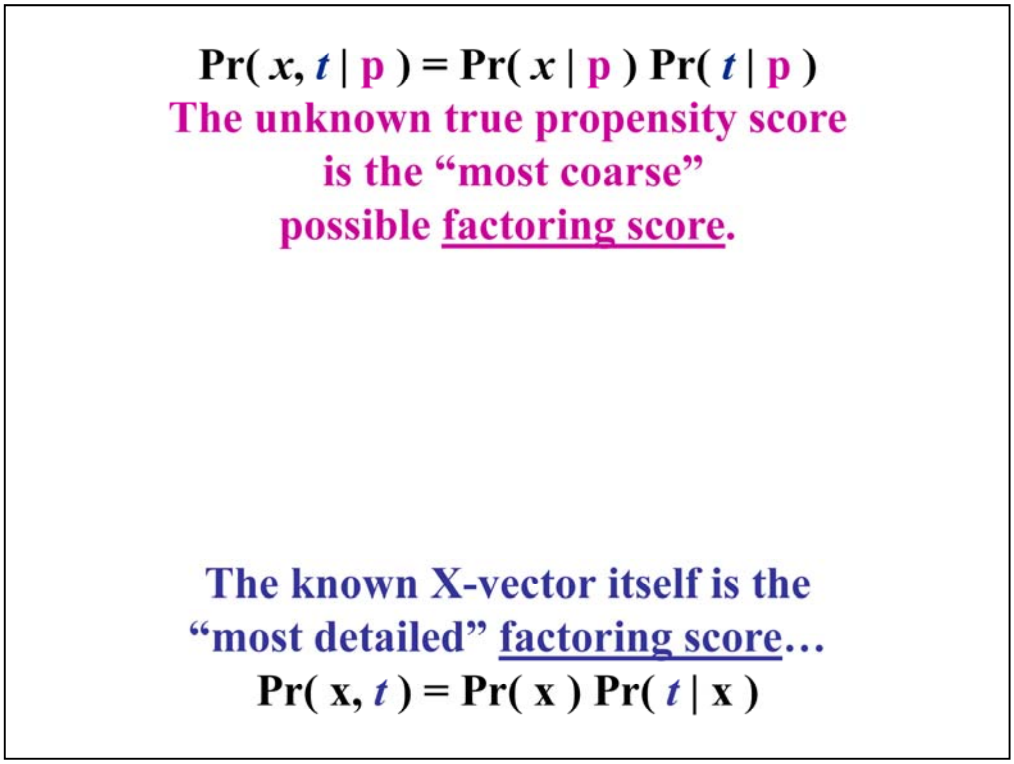
The unknown true propensity score is the “most coarse” possible factoring score.

The known X-vector itself is the “most detailed” factoring score…

$\mathrm{Pr}(\,\mathrm{x}, t\,) = \mathrm{Pr}(\,\mathrm{x}\,)\ \mathrm{Pr}(\,t \mid \mathrm{x}\,)$

The way the formulae factor has very strong statistical implications ..i.e. conditional independence

But PSs are the coarsest possible balancing scores, while x-vectors themselves are the finest.



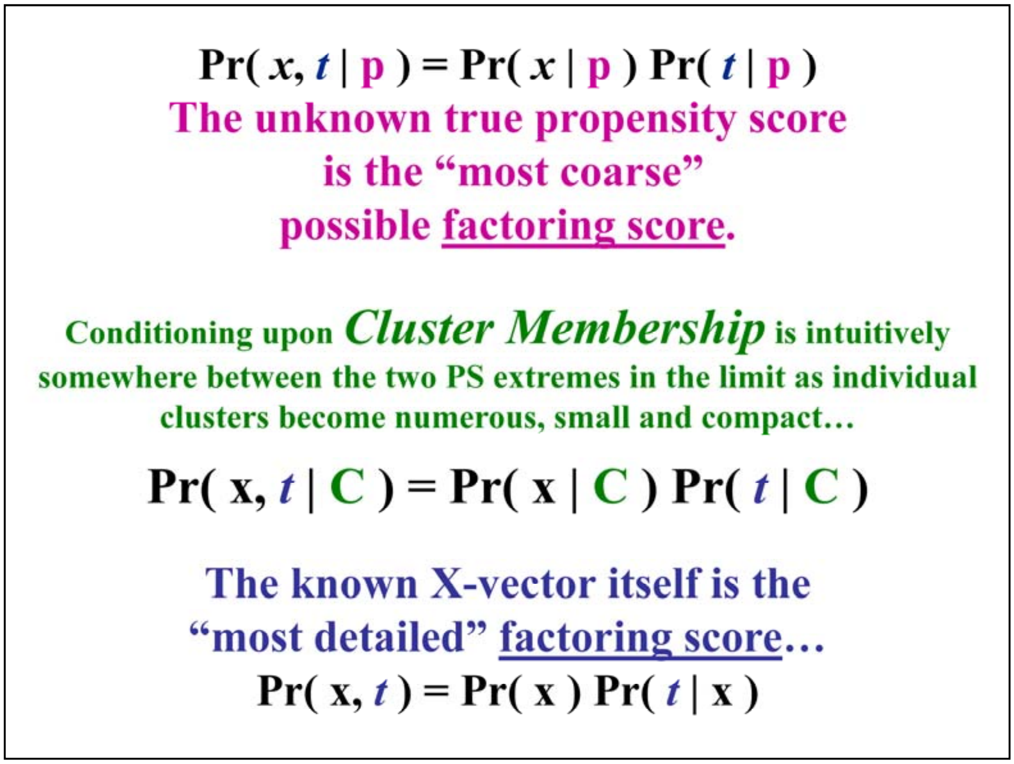
$\Pr(x, t \mid p) = \Pr(x \mid p)\,\Pr(t \mid p)$

The unknown true propensity score is the “most coarse” possible factoring score.

The known X-vector itself is the “most detailed” factoring score…

$\Pr(x, t) = \Pr(x)\,\Pr(t \mid x)$

Is there something between these two extremes?



$$\Pr(x, t \mid p) = \Pr(x \mid p)\,\Pr(t \mid p)$$

The unknown true propensity score is the "most coarse" possible factoring score.

Conditioning upon *Cluster Membership* is intuitively somewhere between the two PS extremes in the limit as individual clusters become numerous, small and compact…

$$\Pr(\mathrm{x}, t \mid \mathrm{C}) = \Pr(\mathrm{x} \mid \mathrm{C})\,\Pr(t \mid \mathrm{C})$$

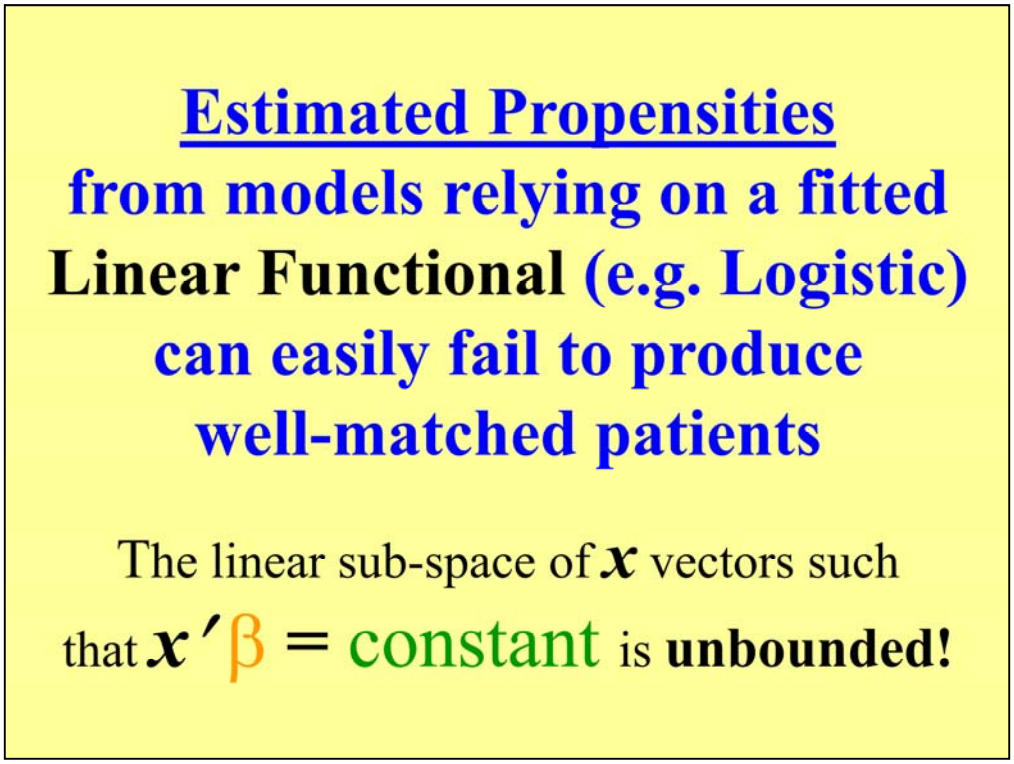
The known X-vector itself is the "most detailed" factoring score…

$$\Pr(\mathrm{x}, t) = \Pr(\mathrm{x})\,\Pr(t \mid \mathrm{x})$$

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

In the limit of interest, the distribution of x within cluster C is essentially UNIFORM.

No need to "check" for balance in this context …because this is being assured by "clustering" patients on X.

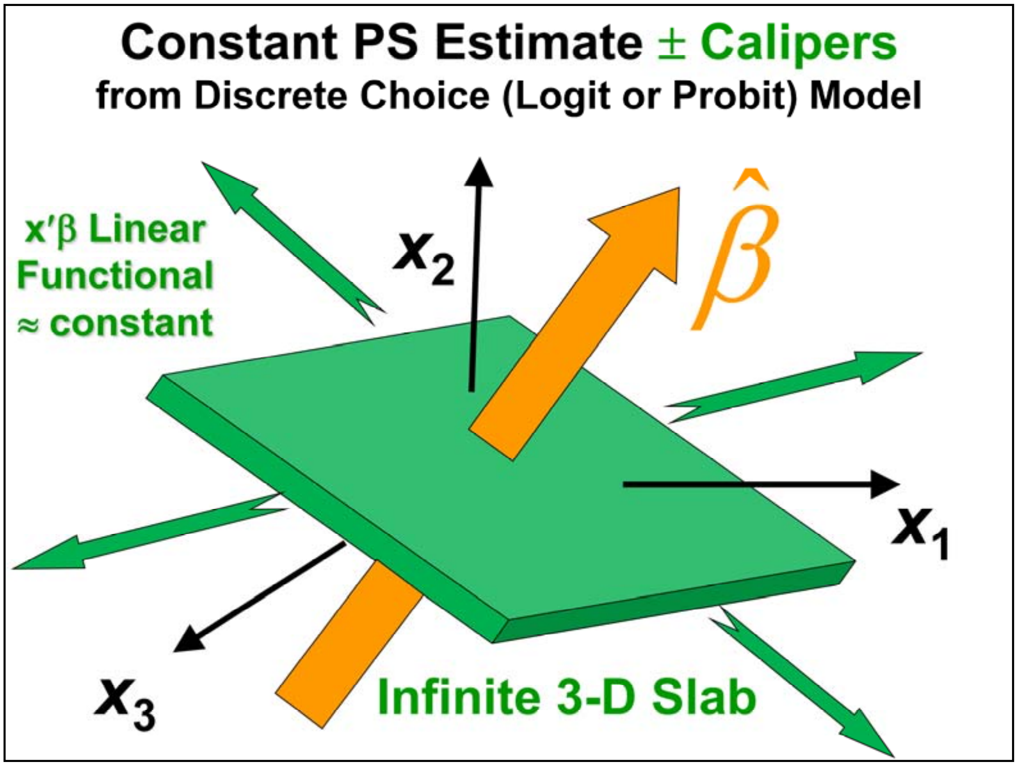


Estimated Propensities from models relying on a fitted Linear Functional (e.g. Logistic) can easily fail to produce well-matched patients

The linear sub-space of $\boldsymbol{x}$ vectors such that $\boldsymbol{x}'\beta = \text{constant}$ is **unbounded!**

Like true propensity scores, there is (unfortunately) a sense in which ESTIMATED propensity scores a COARSE. In fact, they can be so “coarse” that they fail to cause the joint distribution of X and t to conditionally FACTOR and successfully form conditional BLOCKS in X space.

We will examine a pair of displays that will enable us to literally SEE this happening in the 3-dimensional case.



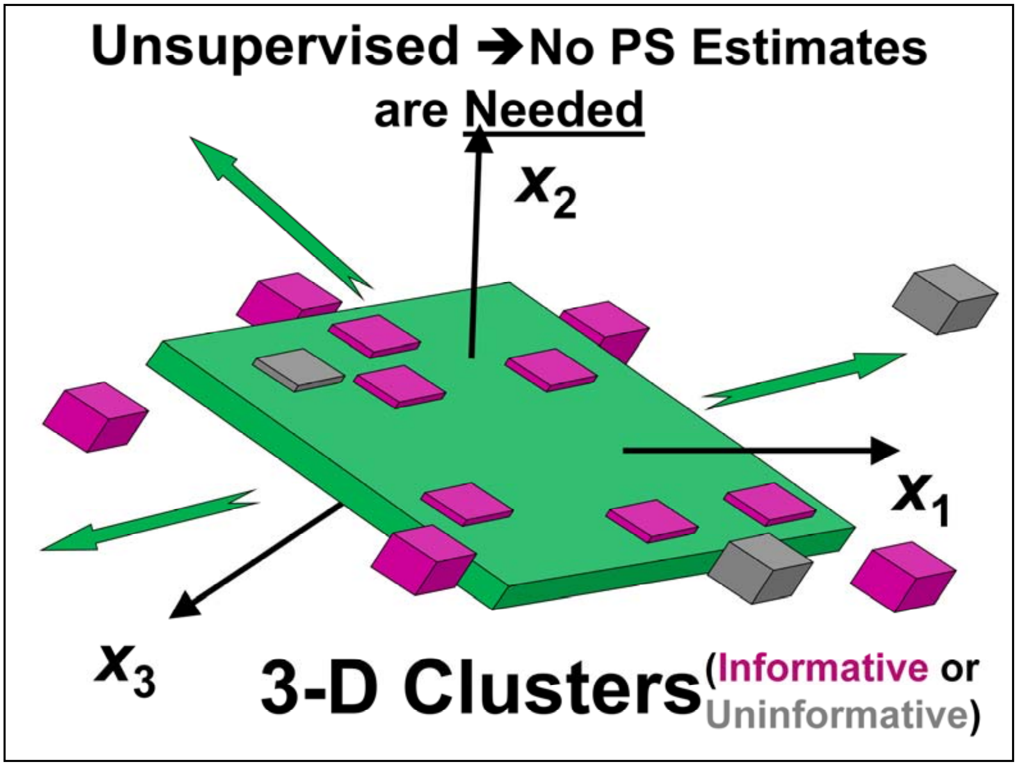
Constant PS Estimate ± Calipers

from Discrete Choice (Logit or Probit) Model

Slab extends to plus/minus infinity in all directions orthogonal to the beta-hat vector (2 dimensional space here.)

However, here the slab has finite thickness (PS plus/minus Calipers) and, thus, infinite volume.

Patients within this X-space slab could certainly have very different x1, x2 and x3 coordinates. Thus no blocking on these patient x-characteristics is automatic.



Unsupervised → No PS Estimates are Needed

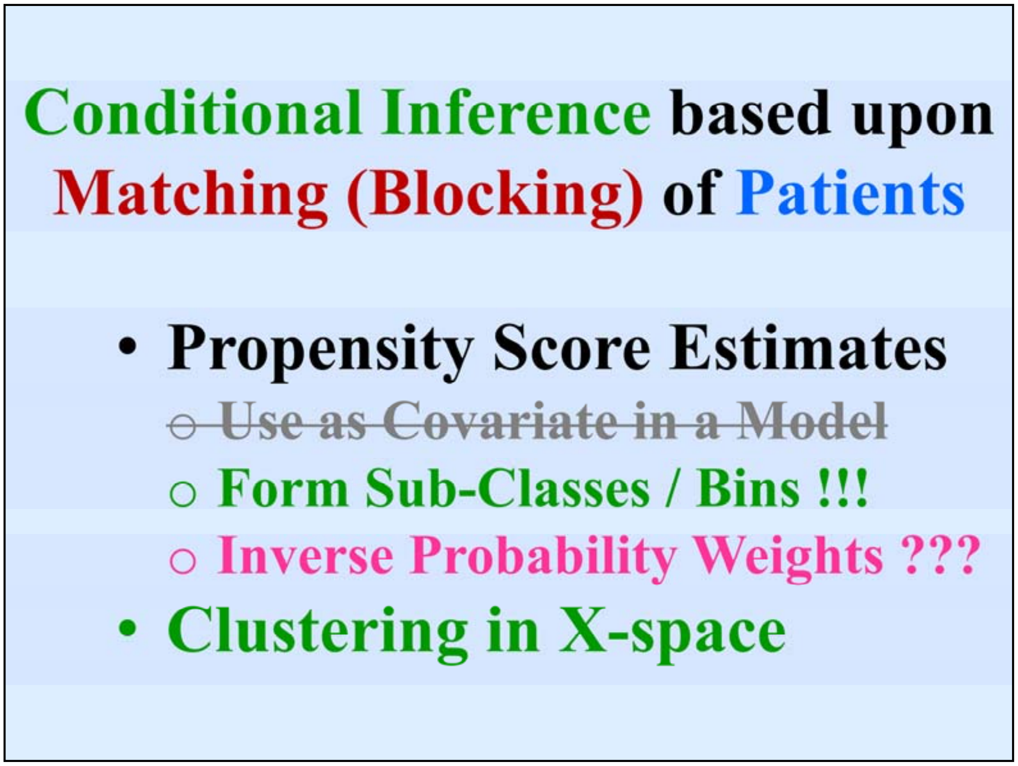
3-D Clusters (Informative or Uninformative)

For simplicity, only the clusters intersecting an x-space slab (linear subspace) are being displayed here.

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

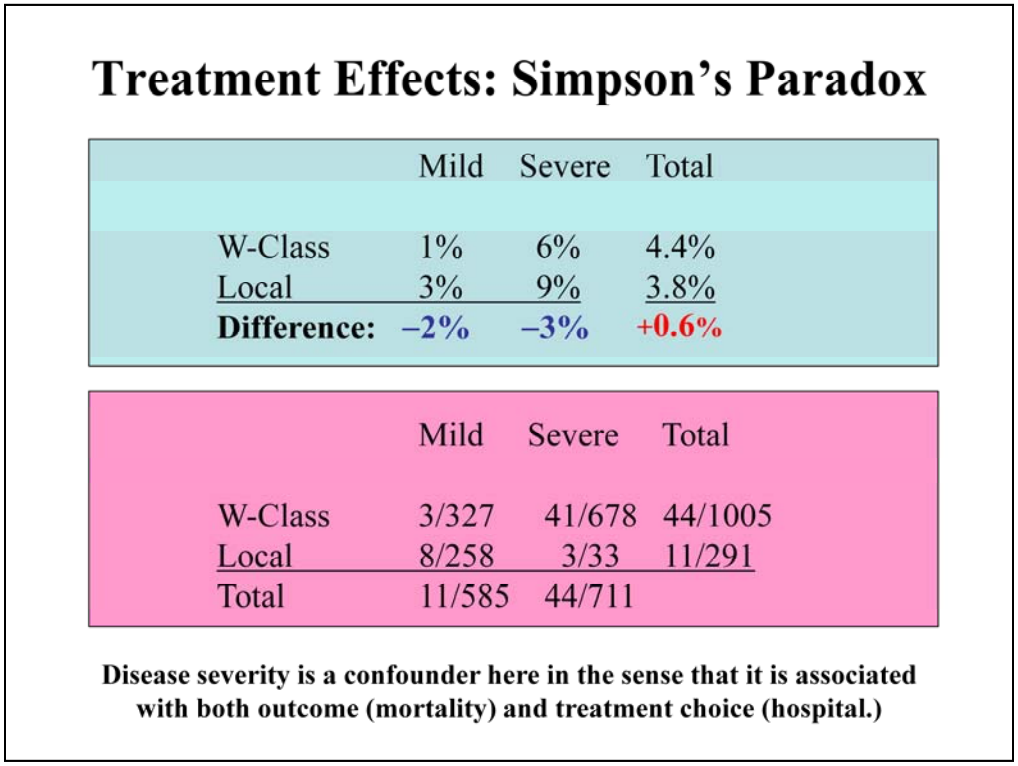
Observed Treatment Fractions within Clusters are Local, non-parametric PS estimates typically left un-used.

Note that “blocking” doesn’t need to be checked or validated here …clustering has assured that patients are VERY NEARLY MATCHED in X-space.



Conditional Inference based upon Matching (Blocking) of Patients

- Propensity Score Estimates
  - ~~Use as Covariate in a Model~~
  - Form Sub-Classes / Bins !!!
  - Inverse Probability Weights ???
- Clustering in X-space



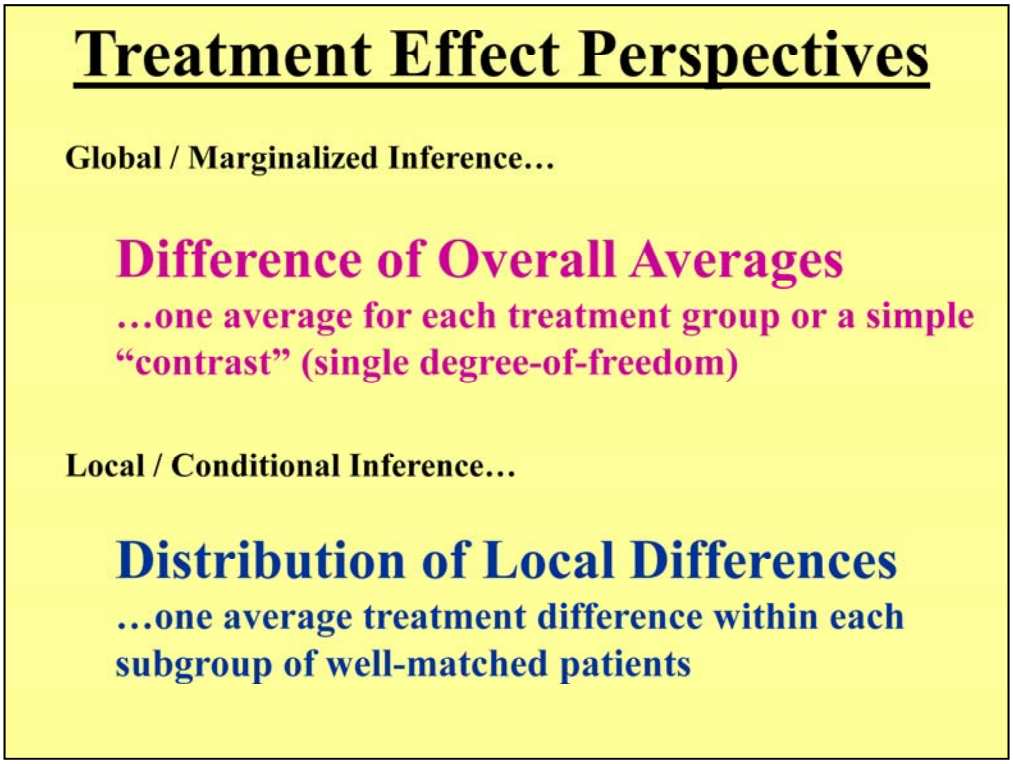
Treatment Effects: Simpson’s Paradox

	Mild	Severe	Total
W-Class	1%	6%	4.4%
Local	3%	9%	3.8%
Difference:	**–2%**	**–3%**	**+0.6%**

	Mild	Severe	Total
W-Class	3/327	41/678	44/1005
Local	8/258	3/33	11/291
Total	11/585	44/711	

Disease severity is a confounder here in the sense that it is associated with both outcome (mortality) and treatment choice (hospital.)

Where would you want to be treated? Your two choices are the World-Class Hospital or the Local Hospital with the above cardiac mortality rates.



Treatment Effect Perspectives

Global / Marginalized Inference…

Difference of Overall Averages

…one average for each treatment group or a simple "contrast" (single degree-of-freedom)

Local / Conditional Inference…

Distribution of Local Differences

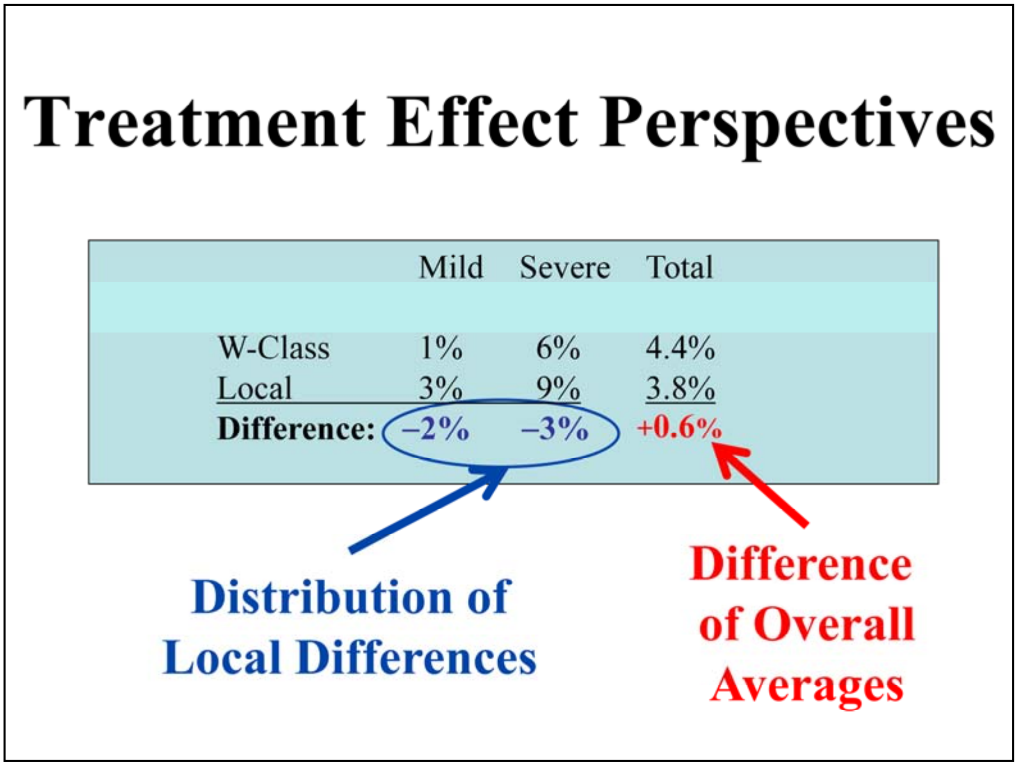
…one average treatment difference within each subgroup of well-matched patients

As in Simpson's Paradox, the “Difference of Overall Averages” approach corresponds to, say, comparing overall hospital mortality rates (a world-class facility vs a local hospital.) This tends to be an UNFAIR comparison when the two hospitals treat very different sorts of patients. Further, only “main effects” are being examined.

The “Distribution of Local (Treatment) Differences” approach corresponds to examining, say, mortality rate differences only within relatively well-matched patient subgroups: Mild, Severe, etc. Furthermore, all possible treatment effects (main effects AND interactions) are pooled together and examined at the same time. The statistical model is nested, containing only (fixed) effects for average outcome in each cluster (regardless of treatment received) and (possibly random) effects for difference in outcome due to treatment within each cluster.

A key feature of methodology for implementing “unsupervised propensity scoring” concepts is that the patient subgroups (classes, strata) do not need to be known in advance. The clustering approach itself identifies the sense in which relatively homogeneous sub-groups of patients are to be formed.

Tukey’s “Sunset Salvo” talked at some length about EXPERT SYSTEMS and the huberis of thinking that every dataset has only one “best” way to be analyzed and, thus, can support only one possible conclusion.



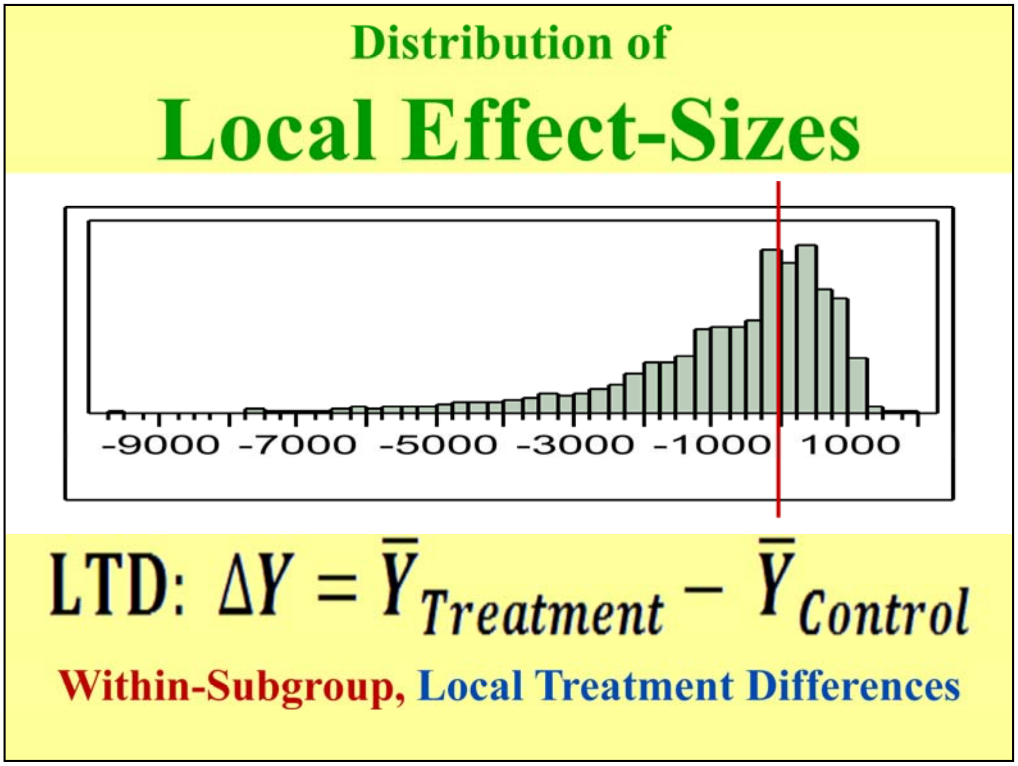
Treatment Effect Perspectives

	Mild	Severe	Total
W-Class	1%	6%	4.4%
Local	3%	9%	3.8%
Difference:	**–2%**	**–3%**	**+0.6%**

Distribution of Local Differences

Difference of Overall Averages

Where would you want to be treated? Your two choices are the World-Class Hospital or the Local Hospital with the above cardiac mortality rates.



Distribution of

Local Effect-Sizes

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

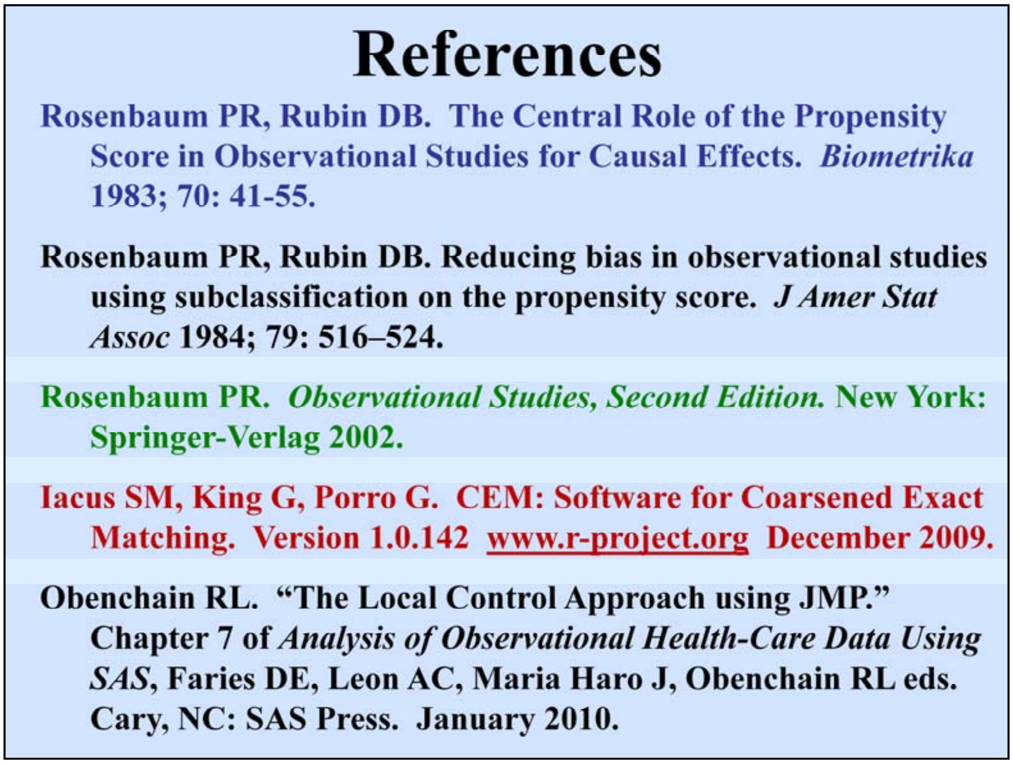
Within-Subgroup, Local Treatment Differences

Using the (newly proposed) AHRQ terminology for OCER, the basic strategy is to compute many LOCAL Average Treatment Effect (ATE) estimates within clusters of patients who are relatively well-matched in X-space.

When viewed as a distribution, the collection of LOCAL ATEs reveal Heterogeneous Treatment Effects (HTEs).

This analytical OCER approach is clearly of the “right” type: Form Treatment Differences First and, only, then display their variation over distinct types of patient subgroups.

If you cannot resist the temptation to compute and overall ATE, weight each LOCAL ATE by the total number of patients in its cluster. This estimate has been “adjusted” (conditionally, rather than marginally) for observed pre-treatment X-characteristics), and simulation studies have shown that it is more accurate that all traditional parametric-model-based ATE estimates in the MOST COMMON sorts of situations. I. E,. whenever the specified model is WRONG due to being either too simple or too complex.



References

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

Rosenbaum PR, Rubin DB. Reducing bias in observational studies using subclassification on the propensity score. *J Amer Stat Assoc* 1984; 79: 516–524.

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

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*, Faries DE, Leon AC, Maria Haro J, Obenchain RL eds. Cary, NC: SAS Press. January 2010.

Local Control Approach: More detail on LTD Distributions…

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 on Slide 21 that 41% of the 39,585 patients with estimated LTDs are **positive**, but the mean is **negative** $635.

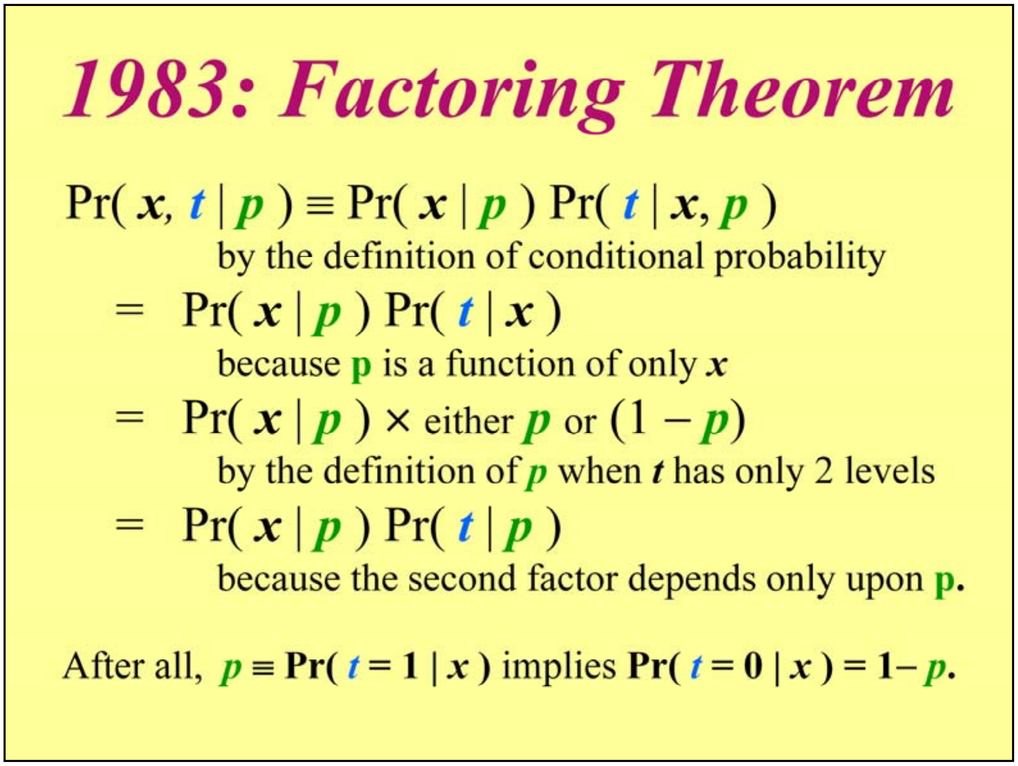
This sort of display provides an objective basis for individualized treatment choices. It depicts the distribution of local, observed effect-sizes estimates …using a simple histogram.

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 Full Distributions => Retain all of the information you can from all patients in "informative" blocks. Basic LC strategy automatically focuses

Extra



1983: Factoring Theorem

$$\Pr(\,\boldsymbol{x}, \boldsymbol{t} \mid \boldsymbol{p}\,) \equiv \Pr(\,\boldsymbol{x} \mid \boldsymbol{p}\,)\,\Pr(\,\boldsymbol{t} \mid \boldsymbol{x}, \boldsymbol{p}\,)$$

by the definition of conditional probability

$$= \Pr(\,\boldsymbol{x} \mid \boldsymbol{p}\,)\,\Pr(\,\boldsymbol{t} \mid \boldsymbol{x}\,)$$

because **p** is a function of only $\boldsymbol{x}$

$$= \Pr(\,\boldsymbol{x} \mid \boldsymbol{p}\,) \times \text{either } \boldsymbol{p} \text{ or } (1 - \boldsymbol{p})$$

by the definition of $\boldsymbol{p}$ when $\boldsymbol{t}$ has only 2 levels

$$= \Pr(\,\boldsymbol{x} \mid \boldsymbol{p}\,)\,\Pr(\,\boldsymbol{t} \mid \boldsymbol{p}\,)$$

because the second factor depends only upon **p.**

After all, $\boldsymbol{p} \equiv \mathbf{Pr}(\,t = 1 \mid x\,)$ implies $\mathbf{Pr}(\,t = 0 \mid x\,) = 1 - \boldsymbol{p}$.

This is a deceptively simple theorem in statistics / probability that requires only rather weak assumptions.

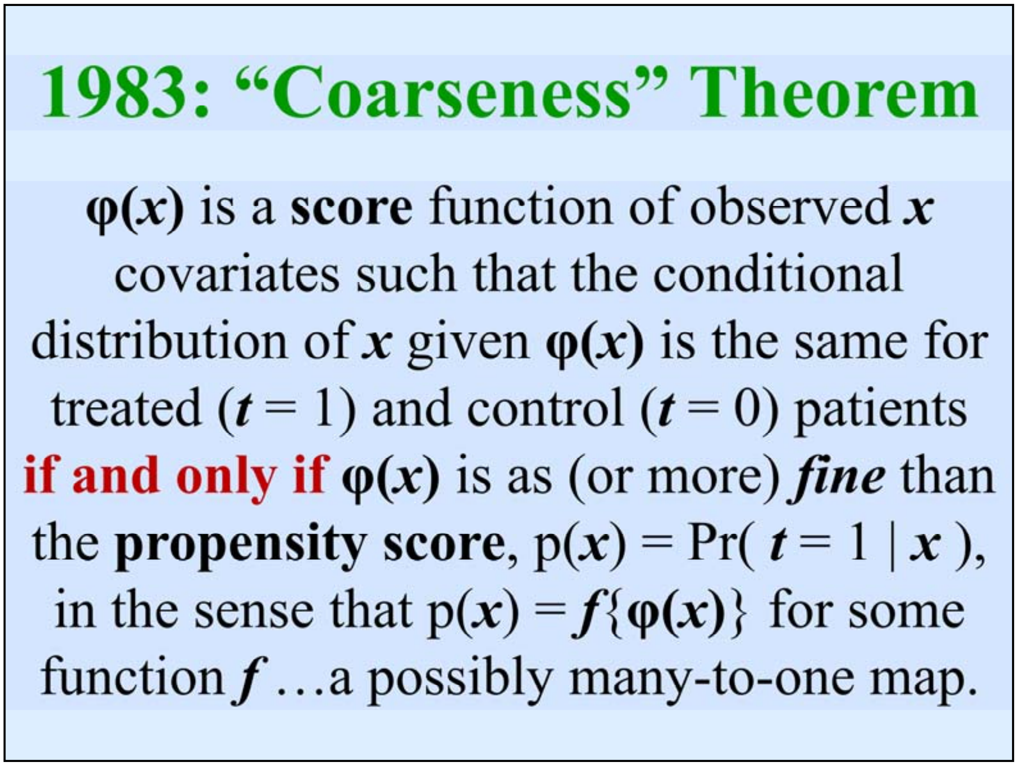
The first line above follows from the very definition of conditional probability.

The second line then follows from the fact that p is only a function of X: p = p(X).

The third line then follows because the final factor is the PS vector, with elements p and 1-p.

The fourth line then follows because the PS if a function of X only through the numerical value of p when there are 2 treatments.

I call this the “Fundamental Theorem” or the “Conditional Independence Theorem” of Propensity Scoring. I think it is misleading to refer to this as the “PS Balancing” Theorem because… NEXT SLIDE!!!



1983: "Coarseness" Theorem

$\varphi(x)$ is a **score** function of observed ***x*** covariates such that the conditional distribution of ***x*** given **$\varphi(x)$** is the same for treated ($t = 1$) and control ($t = 0$) patients **if and only if** **$\varphi(x)$** is as (or more) ***fine*** than the **propensity score**, $p(x) = \Pr(\, t = 1 \mid x\,)$, in the sense that $p(x) = f\{\varphi(x)\}$ for some function ***f*** …a possibly many-to-one map.

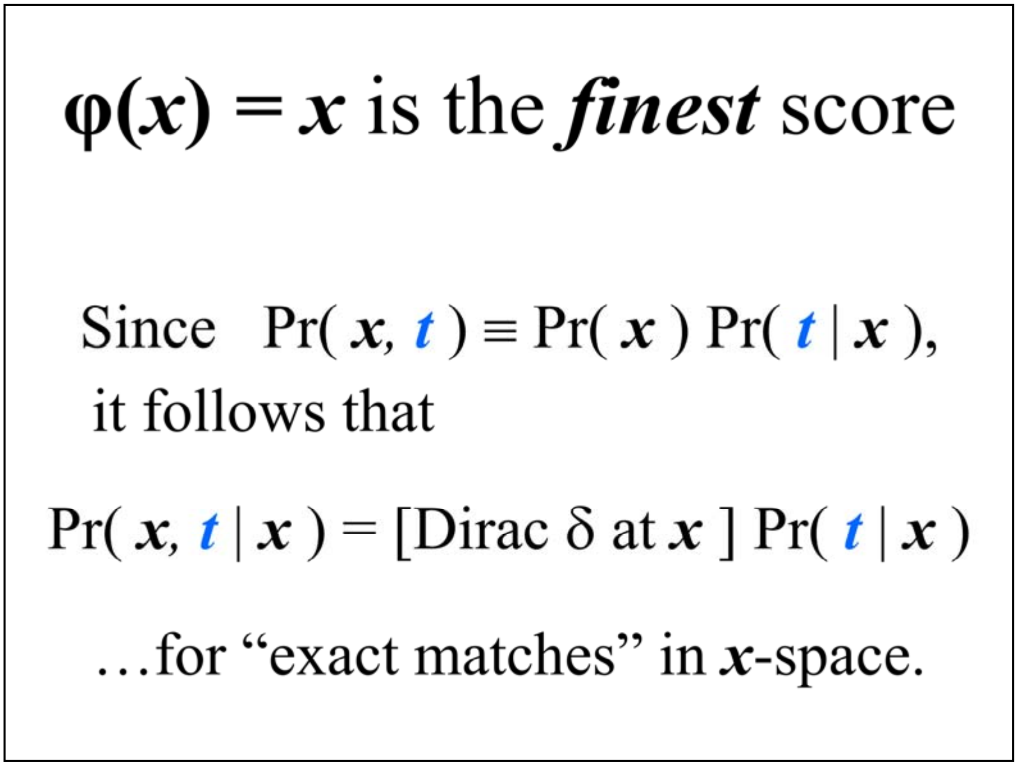
For example, observed patient x-vectors are scores of this type. Clearly, such scores do not need to be conditional probabilities like the unknown, true propensity for treatment choice t=1.

IF: Note that Pr{t=1|phi(x)} = E{p(x)|phi(x)} by the definition of p(x)=Pr(t=1|x). But E{p(x)|phi(x)} = p(x) then follows whenever phi(x) is more fine than p(x).

ONLY IF: If phi(x) is more coarse [rather than more fine] than p(x), there always exist x1 and x2 exist such that p(x1) and p(x2) are different but phi(x1) = phi(x2). However, Pr(t=1|x, phi) is then clearly a function of x rather than of phi alone. This implies that t and x are not conditionally independent given phi, which is a contradiction.

Thus PSs are the coarsest possible factoring scores, while x-vectors themselves are the finest (Rosenbaum & Rubin *Biometrika* 1983; 70: 41-55.)

What, exactly, might be between those extremes???

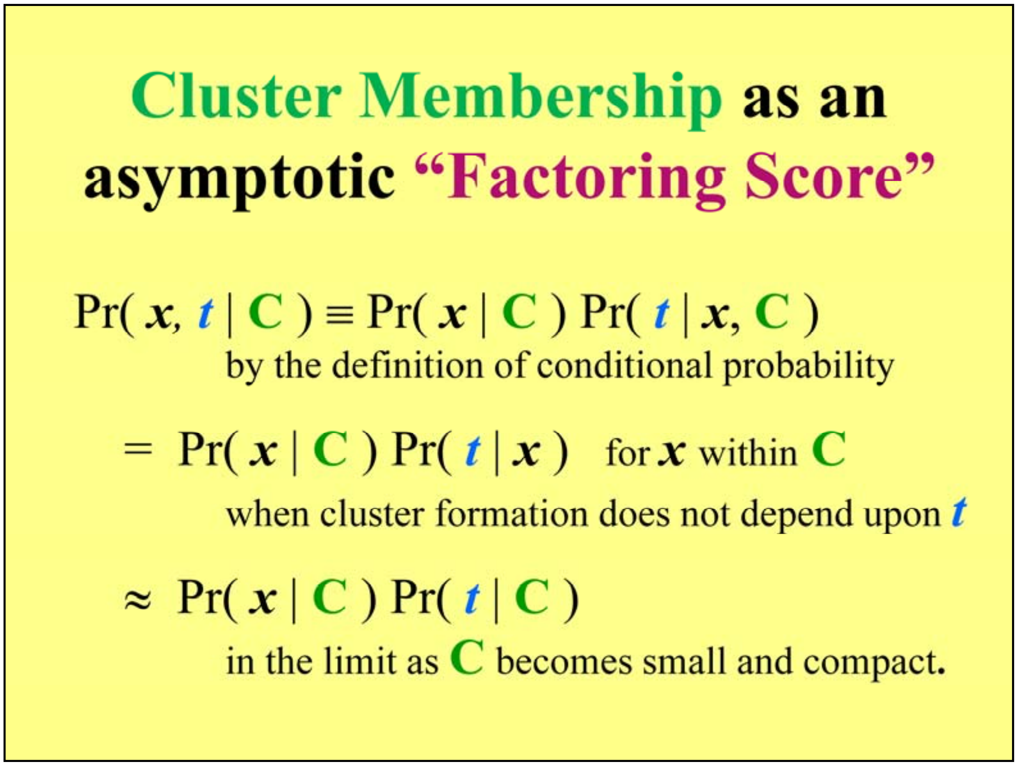


$\varphi(\boldsymbol{x}) = \boldsymbol{x}$ is the ***finest*** score

Since $\Pr(\boldsymbol{x}, \boldsymbol{t}) \equiv \Pr(\boldsymbol{x}) \Pr(\boldsymbol{t} \mid \boldsymbol{x})$,
it follows that

$$\Pr(\boldsymbol{x}, \boldsymbol{t} \mid \boldsymbol{x}) = [\text{Dirac } \delta \text{ at } \boldsymbol{x}] \Pr(\boldsymbol{t} \mid \boldsymbol{x})$$

…for "exact matches" in ***x***-space.



Cluster Membership as an asymptotic “Factoring Score”

$$\Pr(\, \boldsymbol{x}, \boldsymbol{t} \mid \mathrm{C}\,) \equiv \Pr(\, \boldsymbol{x} \mid \mathrm{C}\,)\Pr(\, \boldsymbol{t} \mid \boldsymbol{x}, \mathrm{C}\,)$$

by the definition of conditional probability

$$= \Pr(\, \boldsymbol{x} \mid \mathrm{C}\,)\Pr(\, \boldsymbol{t} \mid \boldsymbol{x}\,) \quad \text{for } \boldsymbol{x} \text{ within } \mathrm{C}$$

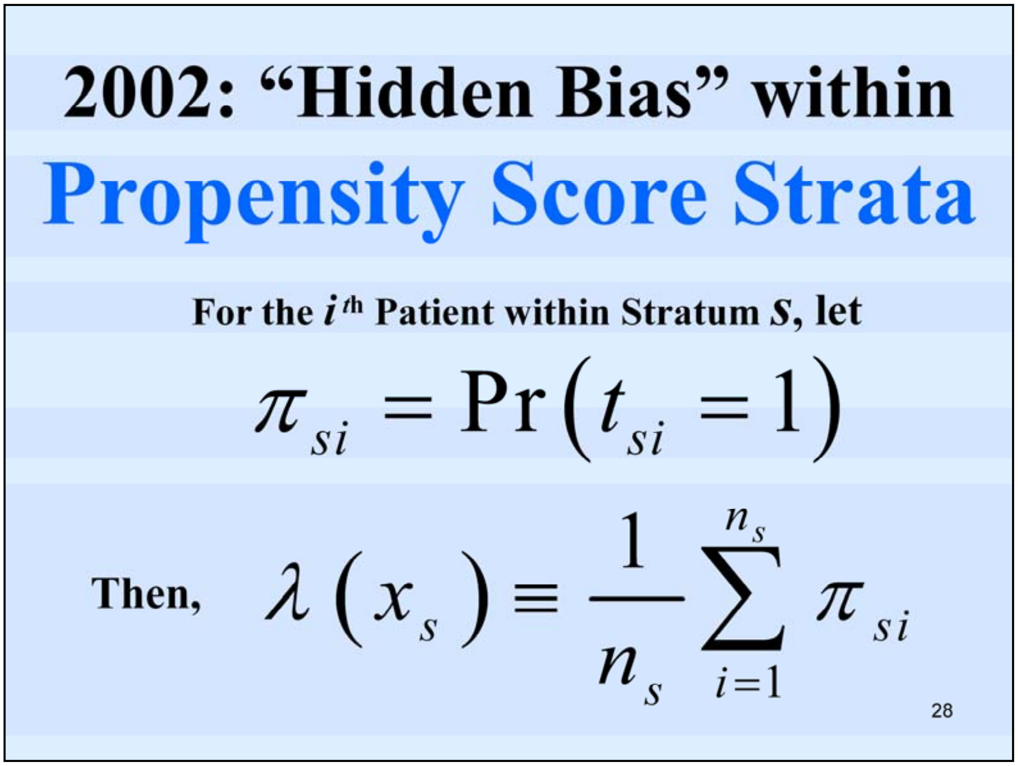
when cluster formation does not depend upon $\boldsymbol{t}$

$$\approx \Pr(\, \boldsymbol{x} \mid \mathrm{C}\,)\Pr(\, \boldsymbol{t} \mid \mathrm{C}\,)$$

in the limit as C becomes small and compact.

In this approach, clusters form the **strata** in the definition of “Local Propensity” in the 2-edition of Rosenbaum’s book (2002).

Why ESTIMATE propensity using a potentially quite un-realistic model when you can simply OBSERVE them (nonparametrically) in a way that essentially assures effective BLOCKING?



2002: "Hidden Bias" within Propensity Score Strata

For the i^{th} Patient within Stratum s, let

$$\pi_{si} = \Pr\left(t_{si} = 1\right)$$

Then, $$\lambda\left(x_s\right) \equiv \frac{1}{n_s} \sum_{i=1}^{n_s} \pi_{si}$$

28

Rosenbaum 2nd edition (2002), §10.2, page 297: Apparently, this is a formulation for propensities that can depend more upon the stratum that a patient is in than upon his/her specific baseline x-characteristics.

Unfortunately, no definition or explanation is given for the x_sub_s symbol !!! What does it MEAN? …Must it NOT depend upon differences between patients in the stratum? …Are all patients in each stratum to have been exactly matched on their x-vectors?

The propensity score, lambda, is the (marginal) probability that a randomly chosen patient from stratum s [that contains n_s patients] will receive treatment t = 1.

There is no (local) "hidden bias" when all of the Pi_sub_si terms are equal for all i within stratum s.