Reducing the Impact of Selection Bias with Propensity Scores

7th International Conference on Health Policy Statistics [ICHPS]
Philadelphia, January 18, 2008

Thomas E. Love, Ph.D. Thomas.Love@case.edu

Center for Health Care Research and Policy
Case Western Reserve University at MetroHealth Medical Center
Cleveland, Ohio www.chrp.org

Case Western Reserve University | Center for Health Care Research & Policy | MetroHealth

Reducing the Impact of Selection Bias with Propensity Scores

7th International Conference for Health Policy Research – Philadelphia – January 2008

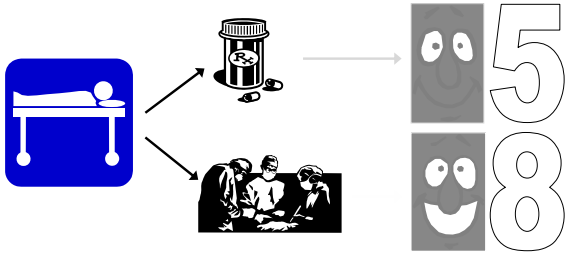
Thomas E. Love, Ph. D.
Center for Health Care Research & Policy
Case Western Reserve University

Workshop Outline

Propensity Methods: A Brief Review
- Estimating the Propensity Score Effectively
- Assessing and Displaying Covariate Balance
- Matching / Stratification / Regression Analyses
- Sensitivity Analysis: The Basics

Strategies and Recent Developments
- Using Propensity Scores Well
- Communicating Results and Thinking
- What's New and Exciting?

Causal Treatment (Exposure) Effects in terms of Potential Outcomes



Surgery-Drug Treatment Effect = 3

Thomas E. Love, Ph. D. Thomas.Love@case.edu www.chrp.org/love

The Database You Wish You Had Vs. Harsh Reality

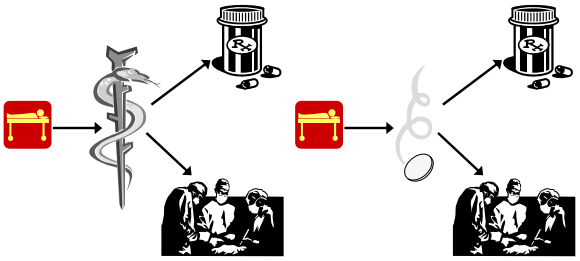
Patient	Outcome under Treatment	Outcome under No Treatment	Treatment's Impact
A	12	?	?
B	7	?	?
C	?	3	?
D	?	9	?

Rubin Causal Model Potential Outcomes Framework

- Strongly ignorable treatment assignment
  - Potential outcomes Y_T and Y_C are assumed conditionally independent of treatment assignment, given the covariates **X**
    - This is a "no hidden bias" assumption
    - Assume **X** contains all relevant information about treatment assignment
  - Data? Clinician-Researchers vs. Economists

Rubin (1997), Rosenbaum (2002)

Observational vs. Randomized Studies



In observational studies, the researcher does not randomly allocate the treatments.

Randomization ensures that subjects receiving different treatments are comparable.

Some Advantages of Smart Observational Studies

- Address chief criticism of RCTs - limited generalizability / external validity
- Data are widely (increasingly) available
  - Often reduced cost and time to get answer
- Enable examination of exposure in "real life"
- May enable examination of effect size and "entrenched practices"
- Large sizes permit investigation of exposures with smaller effect sizes

How Can We Avoid Being Misled?

1. What differentiates an observational study from a randomized controlled trial?
   - One key element: potential for selection bias.
2. What is selection bias, and why should I care about it?
   - Baseline characteristics of comparison groups are different in ways that affect the outcome.
3. What can be done to deal with selection bias in observational studies?
   - Propensity score methods for overt bias.
   - Sensitivity analyses to deal with hidden bias.

Assessing Causal Effect of Exposure on Outcome

- Objective: Draw causal inferences between [use of treatment vs. control] and outcome
- Standard Approach: Risk Adjustment
  - Problem: Selection Bias (people getting treatment are different from people getting control in ways that affect outcome)
- Idea: Compare treated to control subjects that looked similar (had similar propensity for treatment) prior to the treatment decision

Thomas E. Love, Ph. D. Thomas.Love@case.edu www.chrp.org/love

The Propensity Score
Pr(treatment given covariates)

- Definition: The conditional probability of receiving a given exposure (treatment) given a vector of measured covariates.
- Reduces baseline information to a single composite summary of the covariates.

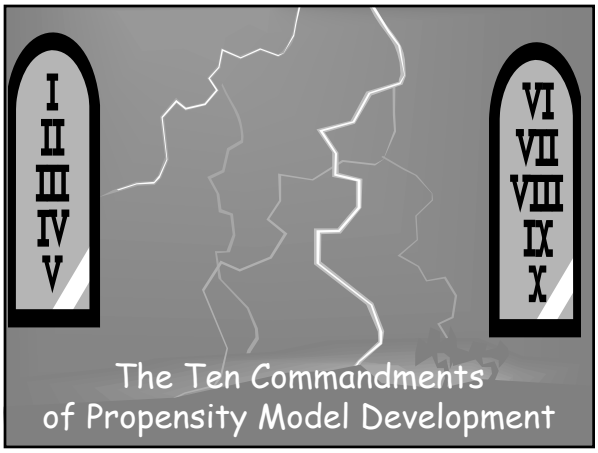
$$\ln\left(\frac{PS}{1-PS}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$$

$$PS = \frac{\exp\left(\beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p\right)}{1 + \exp\left(\beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p\right)}$$

The New Database, Simply

Subject	Propensity to get Treatment	Outcome under Treatment	Outcome under No Treatment
A	.81	12	?
B	.51	7	?
C	.48	?	3
D	.80	?	9

- Match subjects – then plug in estimates…
- So how do we model propensity for treatment?



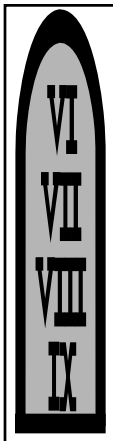
Thou Shalt Value Parsimony

Thou Shalt Examine Thy Predictors For Collinearity

Thou Shalt Test All Thy Predictors For Statistical Significance

Thou Shalt Have Ten Times As Many Subjects As Predictors

Thou Shalt Carefully Examine Thy Regression Coefficients (Beta Weights)

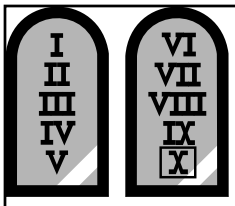


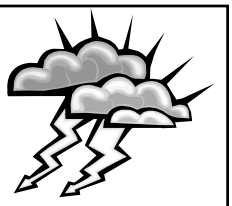
Thou Shalt Perform Bootstrap Analyses To Assess Shrinkage

Thou Shalt Perform Regression Diagnostics and Examine Residuals With Care

Thou Shalt Hold Out A Sample of Thy Data for Cross-Validation

Thou Shalt Perform External Validation on a New Sample of Data





Thou Shalt **IGNORE**
Commandments 1 through 9…
And Instead Simply Ensure
That The Model Adequately
Balances The Covariates

Apologies to Joe Schafer

Thomas E. Love, Ph. D. Thomas.Love@case.edu www.chrp.org/love

Estimating Propensity for Treatment given Covariates

- Usual tool: Logistic regression model for the treatment allocation decision
  - Consider including any variables that have a relationship to the treatment decision
    - Precede it in time
    - Relevant to treatment assignment
    - Relationship to outcome? (some controversy here)
  - No information included on the actual treatment received, or on the outcome(s)
  - I always err on the side of inclusion

DON'T Select Covariates Like This…

Covariate	Treatment	Control	T test Sig?	Include?
Age	49.8	50.1	**No**, p > .05	**No**
BMI	25.5	23.2	**Yes**	**Yes**
Heart Rate	82	76	**Yes**	**Yes**
Years Ed.	11.3	12.0	**No**, p > .05	**No**
Systolic BP	135	133	**No**, p > .05	**No**
SF-36 Phys	61	53	**Yes**	**Yes**

- Most common method: Compare treated and control groups on a long list of covariates, with a t test. Then adjust only for those with a significant difference.

OK, What Should We Do?

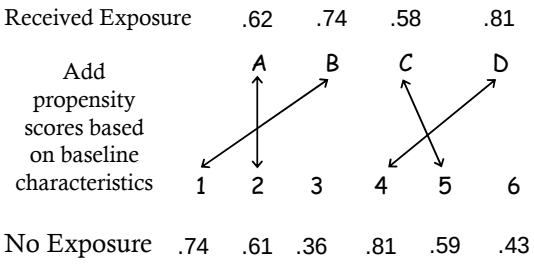
- Give the data multiple opportunities to call attention to potential problems.
- Select a tentative list of covariates for adjustments using problem knowledge and exploratory comparisons of treatment groups.
- Select tentative adjustment method and apply it to the covariates excluded from the list, identifying large imbalances after adjustment.
- Reconsider the tentative list in light of this.

Using the Propensity Score: Multivariate Matching

- Match subjects so that they balance on multiple covariates using one scalar score.
- Goal: Emulate a RCT in matching, then use standard analyses to compare matched sets.
- Design: Treated subjects matched to people who didn't receive treatment but who had similar propensity to receive treatment (match the treated to untreated "clones").

Seminal paper: Rosenbaum and Rubin (1985)

Causal Analysis via PS Match



Aspirin Use and Mortality

- 6174 consecutive adults at CCF undergoing stress echocardiography to evaluate coronary disease.
  - 2310 (37%) were taking aspirin (treatment).
  - 31 covariates – demographics, clinical history, medications, cardiovascular assessment, exercise capacity
- Outcome: All-cause mortality (median follow-up: 3.1 y)
  - Univariate Analysis: 4.5% of aspirin patients died, and 4.5% of non-aspirin patients died… Hazard Ratio: 1.08
- Logistic regression for Propensity for aspirin use: 31 covariates (good discrimination; C = 0.83)
- Greedy match ["5-4-3-2-1 decimal places"] approach

Gum et al. (2001) and http://www2.sas.com/proceedings/sugi26/p214-26.pdf

Baseline Characteristics By Aspirin Use (in %) (before matching)

Variable	Aspirin (n = 2310)	No Aspirin (n = 3864)	P value
Men	77.0	56.1	< .001
Clinical history: diabetes	16.8	11.2	< .001
hypertension	53.0	40.6	< .001
prior coronary artery disease	69.7	20.1	< .001
congestive heart failure	5.5	4.6	.12
Medication use: Beta-blocker	35.1	14.2	< .001
ACE inhibitor	13.0	11.4	< .001

- Baseline characteristics appear very dissimilar: 25 of 31 covariates have $p < .001$, 28 of 31 have $p < .05$.
- Aspirin user covariates indicate higher mortality risk.

Baseline Characteristics By Aspirin Use [%] (after matching)

Variable	Aspirin (n = 1351)	No Aspirin (n = 1351)	P value
Men	70.4	72.1	.33
Clinical history: diabetes	15.0	15.3	.83
hypertension	50.3	51.7	.46
prior coronary artery disease	48.3	48.8	.79
congestive heart failure	5.8	6.6	.43
Medication use: Beta-blocker	26.1	26.5	.79
ACE inhibitor	15.5	15.8	.79

- Baseline characteristics similar in matched users and non-users.
- 30 of 31 covariates show NS difference between matched users and non-users, but only 58% of aspirin users were matched.

Using Standardized Differences to Measure Covariate Balance

$$d = \frac{100\left(\bar{x}_{Treatment} - \bar{x}_{Control}\right)}{\sqrt{\frac{s^2_{Treatment} + s^2_{Control}}{2}}}$$

for continuous variables

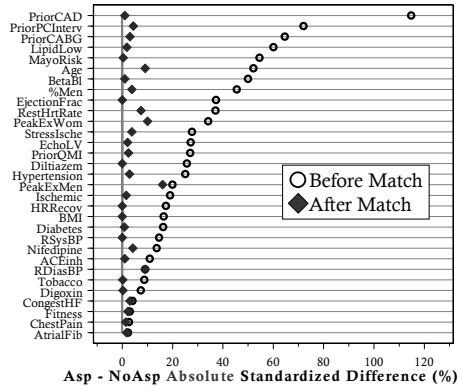
$$d = \frac{100\left(p_{Treatment} - p_{Control}\right)}{\sqrt{\frac{p_T\left(1-p_T\right) + p_C\left(1-p_C\right)}{2}}}$$

for binary variables

Absolute Standardized Differences > 10% indicate serious imbalance [Normand et al. (2001)]

	Aspirin	No Aspirin	P	Std. D.
Before Match	35.1% [811/2310]	14.2% [550/3864]	< .001	49.9%
After Match	26.1% [352/1351]	26.5% [358/1351]	0.79	-1.0%

Absolute Standardized Differences



Estimating The Hazard Ratios

Approach	n	Hazard Ratio	95% CI
Full sample, no adjustment	6174	**1.08**	(.85, 1.39)
Full sample with no PS, adjusted for all covariates	6174	**0.67**	(.51, .87)
PS-Matched sample	2702	**0.53**	(.38, .74)
PS-Matched, adjusted for PS and all covariates	2702	**0.56**	(.40, .78)

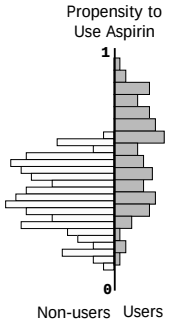
- During follow-up 153 (6%) of the 2702 propensity score-matched patients died.
- Aspirin use was associated with a lower risk of death in matched group (4% vs. 8%, p = .002).

Incomplete vs. Inexact Matching

- Trade-off between
  - Failing to match all treated subjects (incomplete)
  - Matching dissimilar subjects (inexact matching)
- Severe bias due to incomplete matching – it's usually better to match all treated subjects, then follow with analytical adjustments...
  - To clinicians, concern is inexactness.
  - Certainly worthwhile to define the comparison group and carefully explore why subjects match.

For more, see Rosenbaum (1985, 2002)

Who's Getting Matched Here?
Where Do The Propensity Scores Overlap?



Caveat: This simulation depicts what often happens.

Aspirin users we won't be able to match effectively

Aspirin users for whom we might realistically find a good match in our pool of non-users

Using Software to do the Matching

- SAS http://mayoresearch.mayo.edu/mayo/research/biostat/sasmacros.cfm
- SPSS http://sswnt5.sowo.unc.edu/VRC/Lectures/index.htm
- STATA: PSMATCH2 functions
  - Feed in: 1/0 for treated/control, PS, outcome, ID
  - Nearest-neighbor (greedy)
  - Caliper matching
  - Kernel-based matching
  - Common support (treated whose PS is larger than largest PS in the control group are left unmatched)
  - Mahalanobis metric matching

http://fmwww.bc.edu/RePEc/usug2001/psmatch.pdf gives STATA details (Barbara Sianesi)
http://ssw.unc.edu/jif/sacws/docs/Day1c.doc provides a worked example (Shenyang Guo)
http://elsa.berkeley.edu/~imbens/statamatching.pdf describes MATCH program in STATA

Matching Package in R
Jasjeet Sekhon http://sekhon.berkeley.edu/matching

- Build PS model, obtain scores – send to Match
- **Match** can do multivariate or PS matching…
  - Matching with or without replacement
  - Options for handling ties, weighting covariates
  - Exact and Caliper matching
- **MatchBalance** function evaluates the univariate and multivariate balance of match
  - Comparisons of means, empirical CDFs and QQ plots
  - Bivariate T tests (2-sample pre-match, paired-sample post-match), bootstrap Kolmogorov-Smirnov test (consistency)
  - Multivariate Tests: Chi-square and K-S null deviance tests

MatchIt programs in R

Gary King http://gking.harvard.edu/matchit

- Pre-processing via Matching
  - Exact matching
  - Subclassification
  - Greedy nearest neighbor matching
  - Optimal matching (and see Ben Hansen's paper)
  - Full matching (with Hansen's optmatch)
  - Genetic Matching
- Checking Balance (summaries and plots)
- Conducting Analyses after matching

See Hansen (2004) for more on optimal matching

Advice on Getting A Better Match

- Two concerns: Covariate balance, and Matching as many subjects as possible.
  - If match is incomplete, consider both matching and non-matching (stratification/regression) analyses.
  - Match logit(PS) instead of raw PS, accept matches within a fraction (usually .6) of the pooled (across treatments) standard error of the PS. This is more defensible statistically, and should improve yield.
  - Matching on multivariate distance within PS calipers usually beats matching just on PS.

Using the Propensity Score to Stratify (Subclassify) Subjects

- Stratification by Propensity Score Quintile
  - Fit a PS model for each subject
  - Split the subjects into 5 strata (subclasses) of equal size by their propensity scores.
- Five strata of equal size (quintiles) constructed from the PS will usually suffice to remove over 90% of the selection bias due to each of the individual covariates in the PS model.

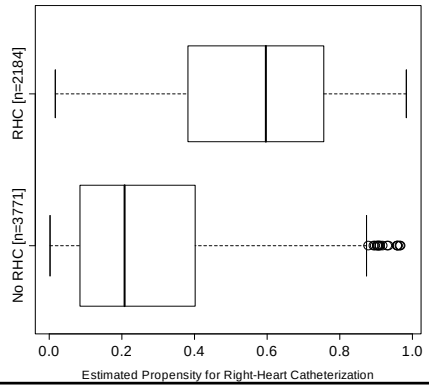
Seminal paper: Rosenbaum and Rubin (1984)

Right Heart Catheterization and Mortality

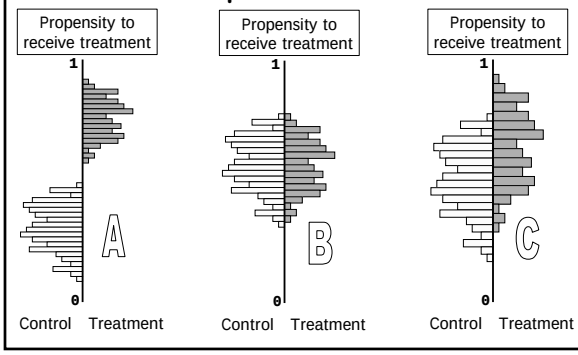
- Why an OS? RHC very popular – Equipoise?
  - RHC directly measures cardiac function – lots of reasons to think this would be helpful.
  - Physician makes the decision – ethical to participate?
- 5735 seriously ill hospitalized pts in SUPPORT
  - 2184 treated patients (RHC within 24 h of admission)
  - 3551 controls (no RHC in first 24 h after admission)
- Key Outcome: 30 day survival
- Panel (7 specialists in clinical care) specified important covariates related to decision to use RHC.

Connors et al. (1996)

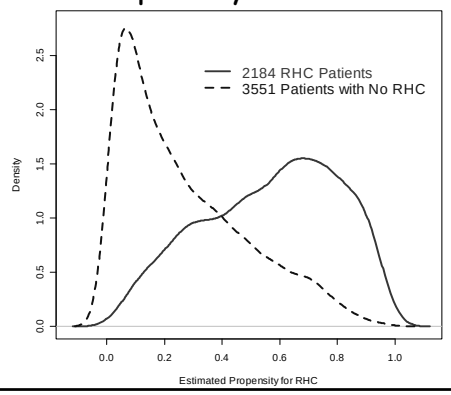
Assessing Propensity Score Overlap



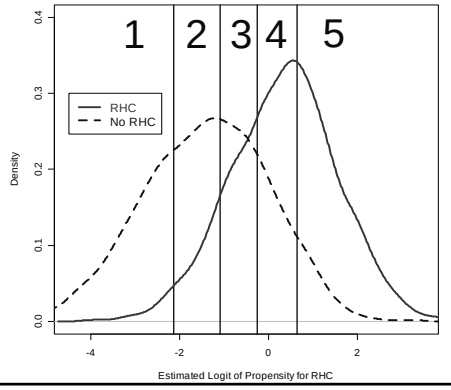
How Much Propensity Score Overlap Do We Want?



Do the Propensity Scores Overlap?



Subclassification by PS Quintiles

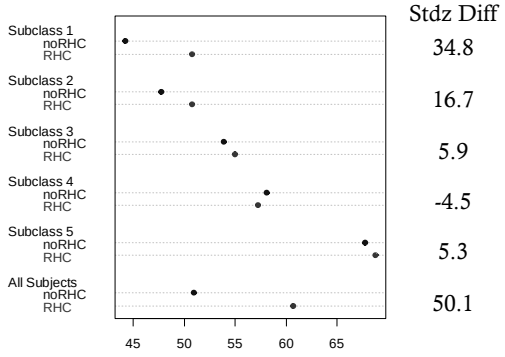


Propensity Score Quintile Subclassification

- 5735 patients were divided into five strata of 1147 patients each, by estimated linear propensity scores.

PS Subclass	Propensity Score ≈ Prob(RHC | covars.)	Actually got RHC	Actually got no RHC
5	Highest 1147 scores	888 (77%)	259 (23%)
4	2nd highest	635 (55%)	512 (45%)
3	Middle	398 (35%)	749 (65%)
2	2nd lowest	210 (18%)	937 (82%)
1	Lowest 1147 scores	53 (5%)	1094 (95%)

APACHE III Acute Physiology Score Across Subclasses



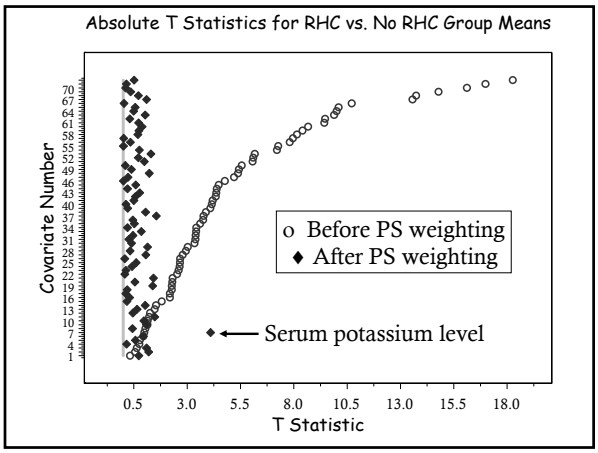
Subclass Specific Estimates

PS Subclass	Group	Patients	P(Survive 30 d)
1	RHC	53	.698
	No RHC	1094	.705
2	RHC	210	.619
	No RHC	937	.705
3	RHC	398	.643
	No RHC	749	.706
4	RHC	635	.649
	No RHC	512	.688
5	RHC	888	.595
	No RHC	259	.645

Propensity Score Weighting

- Idea: Re-weight observations to represent population of interest
- Treated subject's weight is $w_i = \frac{1}{PS_i}$
- Control subject's weight is $w_i = \frac{1}{(1 - PS_i)}$
- SUPPORT: Hirano & Imbens reweight based on PS model with 57 of 72 covars (t > 2.0)...

See Rubin (2001), Rosenbaum (1987), Lunceford & Davidian (2004), Hirano & Imbens (2001), Hirano, Imbens & Ridder (2003) for more. Joffe et al (2004) and Austin & Mamdani (2006) provide applications.



What To Include in the PS Model

- All covariates that subject matter experts (and subjects) judge important when selecting treatments.
- All covariates that relate to treatment and outcome, certainly including any covariate that improves prediction (of exposure group).
- Sop up as much "signal" as possible.
- Why? If our propensity model misses an important reason why subjects are selected to an exposure, we'll be in trouble, and we'll never know it.

What Makes an Unmeasured or Omitted Variable Dangerous?

- All observational studies are potentially affected by hidden bias. Sensitivity analyses are a necessary part of any such study.
- An omitted variable is most likely to affect conclusions about the exposure if it is:
  - closely related to outcome.
  - seriously imbalanced by exposure.
  - uncorrelated with propensity score.

Rosenbaum (1991, 2002), Rosenbaum and Rubin (1983)

Getting the Message Across

- Straightforward, Spreadsheet-Based Formal Sensitivity Analysis for Matched Samples
  - Separate tabs for Outcomes: Binary, Continuous, Survival (w/censoring)
  - All calculations based on base case formulas using sign-score tests as described in Rosenbaum (2002) – some nuances ignored (dealing with ties, etc.)
  - Appendix document provides the details for three examples – battle-tested in my Obs Studies course

Demonstration: A Survival Outcome

- Secondary analysis of DIG trial (Ali Ahmed, PI)
- Exposure: Either Normal or Low Serum Potassium Level in the DIG trial (Heart Failure Pts)
  - 1187 matched pairs of "normal potassium" and "low potassium" HF patient with similar baseline characteristics.
- Outcome: All cause mortality during the follow-up period (i.e. there is censoring)
- Main complicating issue – lots of censoring
- There are 440 "pairs with clear winners"
  - In 335 of these 440, winner is "normal potassium"

Spreadsheet Demo

	A	B	C
1	**Sensitivity Analysis for A Simple Comparison for Censored Survival**		
2	**Section 4.4.8. of Rosenbaum PR (2002) Observational Studies, 2nd Edition.**		
3	**INSERT VALUES (IN RED) IN CELLS HIGHLIGHTED IN YELLOW.**		
4			
5	**Data**		
6	**Total # of Pairs With A Clear Winner**	440	
7	**T = # Pairs Where Exposed Outlives Control**	335	
8			
9	**Sensitivity Analysis**		
10	**Gamma Values**	2-tail P value (lower bound)	2-tail P value (upper bound)
11	1.0	0.0000	0.0000
12	1.5	0.0000	0.0000
13	2.0	0.0000	0.0000
14	2.5	0.0000	0.0288
15	3.0	0.0000	0.5820
16	3.5	0.0000	1.0000
17	4.0	0.0000	1.0000
18	4.5	0.0000	1.0000
19	5.0	0.0000	1.0000
20	5.5	0.0000	1.0000
21	6.0	0.0000	1.0000
22			
23	**Insert Gamma Value Below**	2-tail P value (lower bound)	2-tail P value (upper bound)
24	2.563	0.0000	0.0498
25	**Stop when value for the upper bound of the P value (cell C24) is just below desired two-tailed significance level**		

Communicating The Results

- In the absence of hidden bias, a sign-score test for matched data with censoring provides strong evidence ($p < .001$) that low (vs. normal) potassium decreases survival time, even after adjustment via PS matching.
- To attribute the lower survival time to an unobserved binary covariate unrelated to our propensity model rather than the effect of low potassium, that covariate would need to both:
  1. increase the odds of low potassium more than 2.5-fold and
  2. be an excellent predictor of mortality.

Idealized Standards for Evaluation of The "Sensitivity" Problem

- LTE: Logic, Theory & Empirical Evidence
- It is unlikely that a hidden bias would substantially affect these conclusions:
  - Measured and incorporated every major known factor that they could identify.
  - Treatment effects on health outcomes were generally quite large, consistent with earlier studies, and clinically plausible.

Rich and Poor Covariate Sets

- With a rich set of covariates, adjustments for hidden covariates may be less critical.
- With less rich covariate sets, we may need to do more – say, try to find an instrument.
- Conclusion after the initial design stage may be that the treatment and control groups are too far apart to produce reliable effect estimates without heroic modeling assumptions.

A Few "Frequently Asked Questions"

1. Why bother? Why not just build a big regression model?
2. What about propensity model diagnostics? Should we be checking goodness of fit?
3. What about propensity scores when we have more than two possible exposures?
4. Do propensity scores have a role in designing observational studies, as well as analyzing?
5. What about instrumental variables?

Go to Wrapup

Why not model outcome using all variables in the propensity model?

- Two stages: fit PS, then use PS in model
- One stage: just fit big outcome model
- Pros of two-stage approach:
  - Forces you to think hard about selection.
  - You don't care about parsimony in the PS, so you get maximum predictive value there.
  - You can fit a very complicated PS model first with interactions, higher order terms, splines, etc.
  - You can fit a smaller outcome model, which may let you assess its validity more accurately.

D'Agostino 1998 and Rubin 2001 provide some more details

Return to FAQs

What about Propensity Model Diagnostics?

- Rubin describes "confusion between two kinds of statistical diagnostics…"

(1) Diagnostics for the successful prediction of probabilities and parameter estimates underlying those probabilities

(2) Diagnostics for the successful design of observational studies based on estimated propensity scores.

- Basically, (2) has a role – (1) doesn't, here.

Rubin (2004)

Should we be checking propensity model goodness of fit?

- Are tests used to evaluate logistic model fit and discrimination helpful in detecting the omission of an important confounder?
  - Simulated data including an important binary confounder – compared inclusion to exclusion
- Hosmer-Lemeshow GOF and C statistic were of no value in detecting residual confounding in treatment effect estimates

Weitzen et al. (2005)

Return to FAQs

Propensity Scores for More Than Two Possible Exposures

- Several generalizations have been proposed, which attempt to maintain the balancing properties of the usual propensity score…
- Role of the PS in estimating dose-response functions **Imbens 2000 Biometrika**
- Matching with doses in an OS of a media campaign against drug abuse **Lu et al 2004 JASA**
- Propensity Scores with Continuous Treatments
  http://www.biostat.jhsph.edu/~dscharf/Causal/hirano-imbens-2004.pdf

Return to FAQs

Designing Observational Studies: Rubin's "Rules" for Regression in OS

- Three conditions which must all apply for regression adjustment to be trustworthy:
- Difference in the means of logit(PS) in the two groups being compared must be small.
- Ratio of variances of logit(PS) in the two groups must be close to 1.
- Ratio of variances of the "residuals" of the covariates after PS adjustment close to 1.

Rubin (2001)

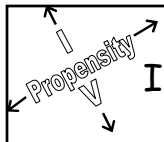
Why Work This Hard in the Initial Stage of Design?

- No harm, no foul.
  - Since no outcome data are available to the PS, nothing based on the PS biases estimation of treatment effects.
- Balancing covariates / PS makes subsequent model-based adjustments more reliable.
  - Model adjustments can be extremely unreliable when treatment groups are far apart on covariates.

Return to FAQs

When Are Instrumental Variables Methods Especially Attractive? An instrument is available, and …

- Assignment to a treatment is ignorable, but compliance with the assignment is not perfect so that the dose of treatment received is non-ignorable.
- Data are weak, in the sense that observed covariates provide insufficient insight into the background to allow estimated effects (adjusting for covariates) to be due to treatment.



Propensity Scores vs. Instrumental Variables?

- Some questions call for PS adjustment, others for IV models of Rx effect.
- Both have unverifiable assumptions:
  - PS adjusts for selection bias in terms of identified covariates – we must presume this is sufficient to also adjust for unobserved covariates. Sensitivity analysis can help.
  - IV presumes we can and do identify appropriate instrument(s).

Return to FAQs

Advantages of PS Methodology

- Results can be persuasive even to audiences with limited statistical training.
- Though estimating the PS requires some care, the comparability of treated and control patients can be verified simply.
- PS methods address selection bias well.
- PS methods may be combined with other sorts of adjustments.
- Methods appealing to grant review boards?

Strategic Issues in Observational Studies (Rosenbaum, 2002)

- Design observational studies
  - Exert as much experimental control as possible, carefully consider the selection process, and anticipate hidden biases
- Focus on simple comparisons
  - Increase impact of results on consumers
- Compare subjects who looked comparable prior to treatment
- Use sensitivity analyses to delimit discussions of hidden biases due to unobserved covariates

Rosenbaum (2002)

Some Cautions and Limitations

- Hidden Bias: Beware unmeasured covariates which affect outcomes and/or assignment.
  - Sensitivity Analysis helps quantify the problem
- This is a reasonable method with fairly large samples.
  - Matching vs. stratification vs. adjustment methods
- Options narrow as an investigation proceeds.
  - Sadly, though OS work cries out for design, we're often working with secondary data, where we have fewer options