Strategies for Using Propensity Scores Well

A Workshop given by

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

at the

6th International Conference for Health Policy Research

Friday, October 28, 2005
10:30 AM to 12:15 PM

Boston Park Plaza Hotel and Towers

Table of Contents

Strategies for Using Propensity Scores Well

6th International Conference for Health Policy Research – Boston – October 28, 2005

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

Propensity Scores in the News!

Firearm Violence Exposure and Serious Violent Behavior

Jeffrey B. Bingenheimer,[1*] Robert T. Brennan,[2] Felton J. Earls[2]

To estimate the cause-effect relationship between exposure to firearm violence and subsequent perpetration of serious violence, we applied the analytic method of propensity stratification to longitudinal data on adolescents residing in Chicago, Illinois. Results indicate that exposure to firearm violence approximately doubles the probability that an adolescent will perpetrate serious violence over the subsequent 2 years.

Bingenheimer et al. (2005), Holden (2005)

Acknowledgments

- Center for Health Care Research and Policy, including …
  - Randy Cebul, MD
  - Neal Dawson, MD
  - Charles Thomas
  - Stephen Lewis, MS
  - Scott Husak
  - Al Connors, MD
  - Irene Katzan, MD MS
  - Pat Murray, MD MS
  - Mark Votruba, PhD
- Students in CRSP 500
- Paul Rosenbaum, PhD
- Therese Stukel, PhD
- Sharon-Lise Normand, PhD
- Rickey Mehta
- Ralph D'Agostino, Jr., PhD
- Ben Hansen, PhD
- Michael Posner, MS
- Ali Ahmed, MD
- Society for Medical Decision Making
- ASA Cleveland Chapter
- ASA Health Policy Statistics Section

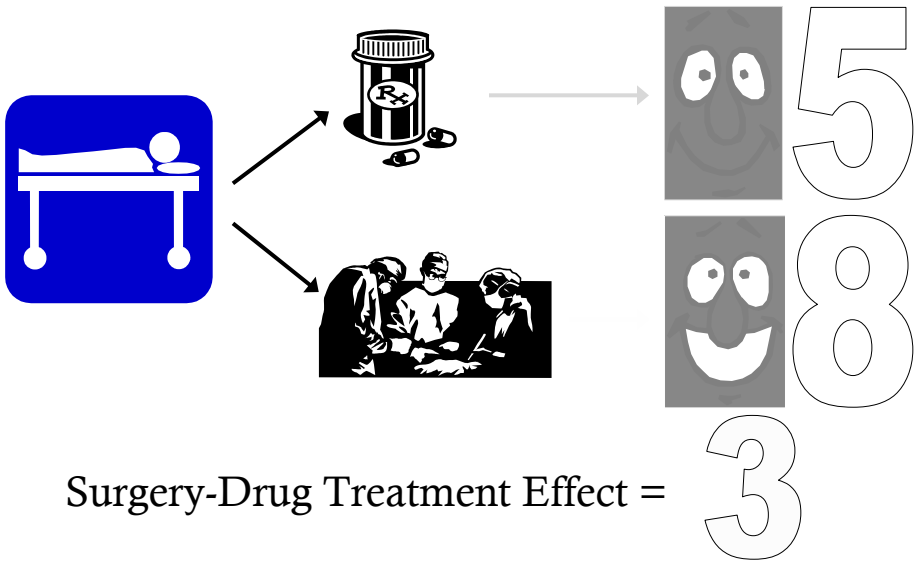
Workshop Outline

1. A Brief Review of Propensity Methods
2. Estimating the Propensity Score Effectively
3. Assessing and Displaying Covariate Balance
4. Choosing Analytic Techniques
5. Communicating Results to a Non-Statistical Audience

Part One

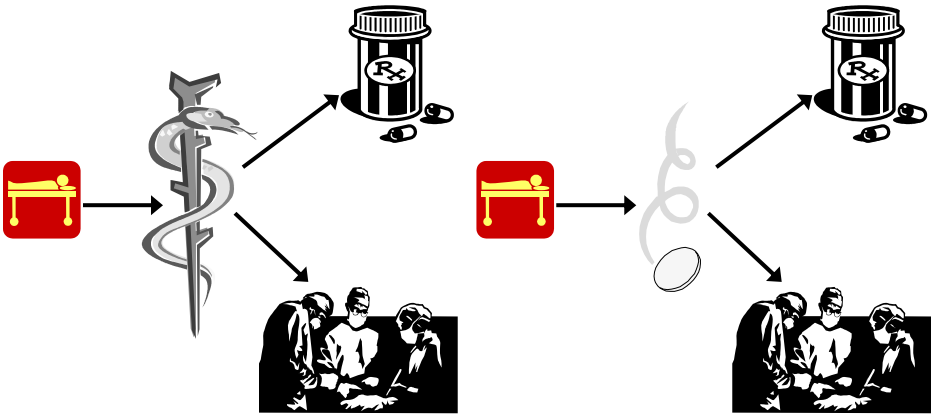
A (Brief) Review of Propensity Score Methods

Causal Treatment (Exposure) Effects in terms of Potential Outcomes



Surgery-Drug Treatment Effect =

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

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	?

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

Rubin Causal Model Framework

- Potential Outcomes posits two random variables to describe the outcome for each subject in the study
  - Y_T = Outcome under the treatment condition
  - Y_C = Outcome under the control condition
  - **X** = vector of covariates
  - Z = treatment assignment variable (1 if treated, 0 if control

Rubin (1997), Rosenbaum (2002)

Making Inference About Treatment Effects Possible

- Strongly Ignorable Treatment Assignment Assumption:
  - The outcome variables Y_T and Y_C are assumed to be conditionally independent of the treatment assignment variable Z, given the covariates **X**
  - This is a "no hidden bias" assumption
  - We assume that **X** contains all relevant information about treatment assignment

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.

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

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

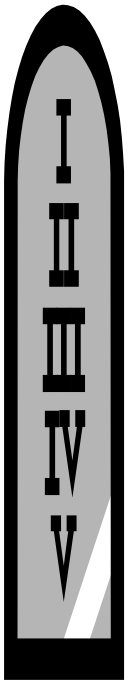
The New Database, Simply

Subject	Propensity to get Treatment	Outcome under Treatment	Outcome under No Treatment
A	.80	12	?
B	.50	7	?
C	.50	?	3
D	.80	?	9

- Match A to D and B to C and plug in resulting estimates for question marks…

So How Do We Build a Propensity Score Model?

- Need to estimate Pr(treatment | covariates)
- Usual tool: Logistic regression model for the treatment allocation decision
  - We therefore want to consider including any variables that have a relationship to the treatment decision (i.e. precede it in time, and are relevant)
  - Clearly, there's no information included on the actual treatment received, or on the outcome(s).



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

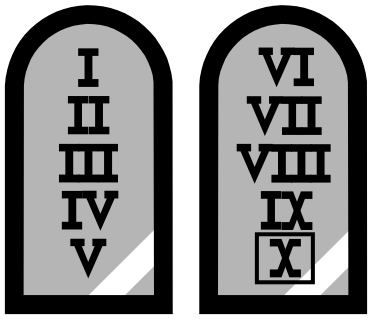
Thou Shalt Carefully Examine Thy Regression Coefficients (Beta Weights)

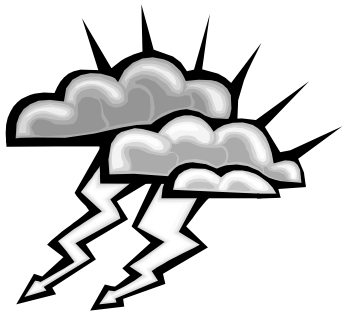
Thou Shalt Perform Bootstrap Analyses To Assess Shrinkage

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

Using the Propensity Score: Three Examples

- Matching using the Propensity Score
  - Aspirin Use and Mortality in Stress Echo patients
- Stratification / Subclassification using the Propensity Score (also includes weighting)
  - Right Heart Catheterization in the Seriously Ill
- Direct Adjustment with Propensity Scores
  - Prostate Cancer surgery vs. Radiotherapy

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)

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…
  - Unadjusted Hazard Ratio: 1.08 (0.85, 1.39)
- Logistic regression for Propensity for aspirin use: 31 covariates (good discrimination; C = 0.83)

Gum et al. (2001)

How Were The Aspirin Subjects Matched?

- Tried to match each aspirin user to a unique non-user with a PS identical to 5 digits.
- If not possible, proceeded to a 4-digit match, then 3-digit, 2-digit, and finally a 1-digit match (i.e., propensity scores within .099).
- Result: matches for 1351 (58%) of the 2310 aspirin patients to 1351 unique non-users.

SAS macro: 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. [Peak exercise capacity for men is p = .01]

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

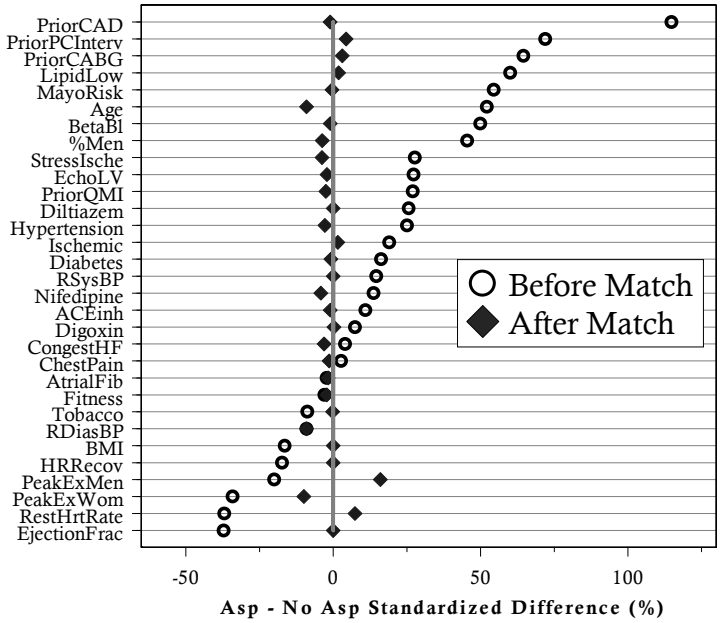
$$d = \frac{100\left(p_{Treatment} - p_{Control}\right)}{\sqrt{\frac{p_T\left(1-p_T\right) + p_C\left(1-p_C\right)}{2}}}$$

for binary variables

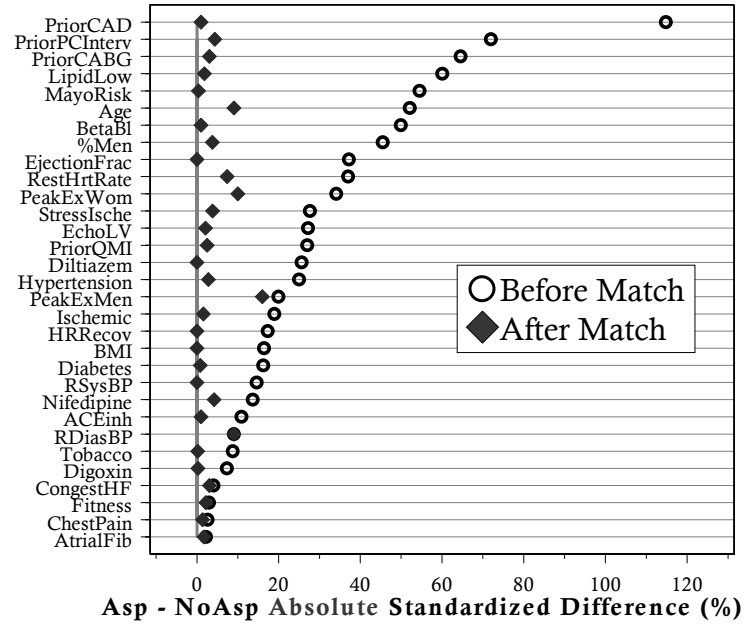
Absolute Standardized Differences > 10% indicate serious imbalance [Normand et al. (2001)]

	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%

Covariate Balance for Aspirin Study



Absolute Standardized Differences



Matching with Propensity Scores

- 1351 aspirin subjects matched well to non-aspirin subjects – big improvement in covariate balance. Matched group looks like an RCT…
- Matching still incomplete, but results on PS matched group mirrored the results for the covariate-adjusted group as a whole…
- Resulting matched pairs analyzed using standard statistical methods, e.g. Kaplan-Meier, Cox proportional hazards models.

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)

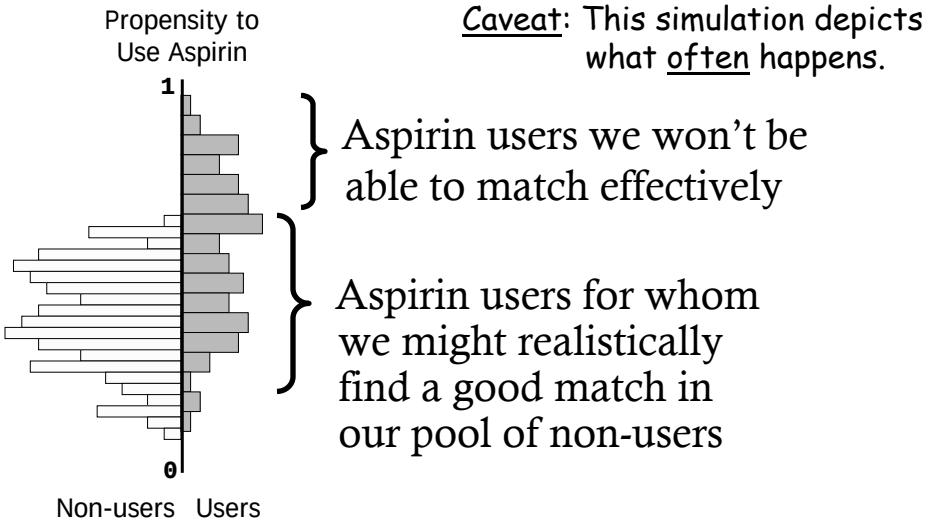
- During follow-up 153 (6%) of the 2702 propensity score-matched patients died.
- Aspirin use was associated with a lower risk of death in matched group (4% vs. 8%, p = .002).

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 for residual imbalances in the covariates.
- In practice, concern has been 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?



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?
  - RHC effect sizes may be small
- 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

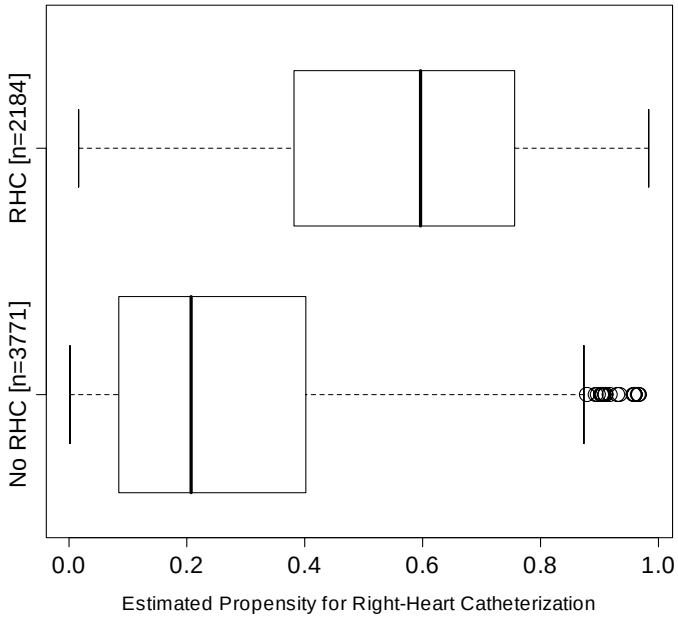
Connors et al. (1996)

Some Characteristics used to predict Propensity for RHC use

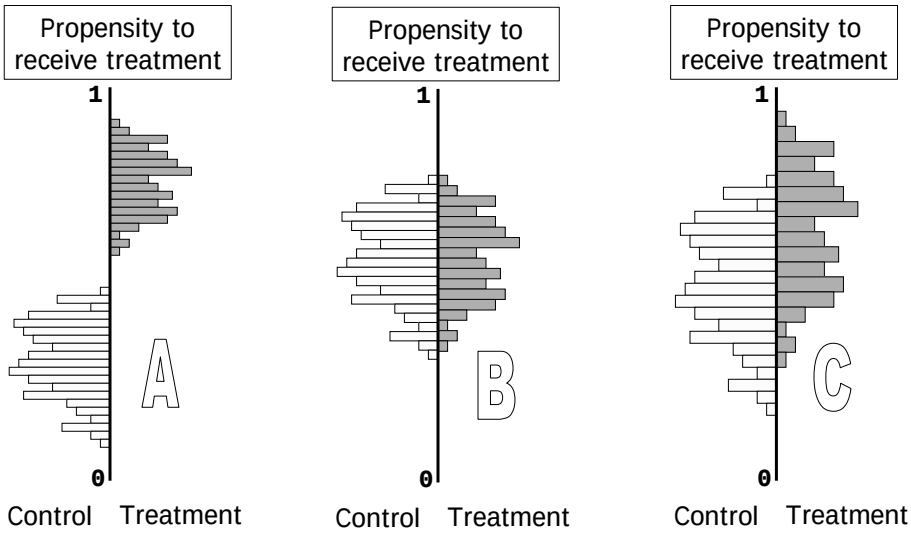
- Age, Sex, Race
- Education, Income
- Insurance type
- Primary and secondary disease category
- 12 categories of admission diagnosis
- ADL & DASI 2 weeks before admission
- DNR status on day 1
- Cancer (none, local, metastatized)
- 2-month survival model
- Weight, temperature, BP, heart rate, resp. rate
- 13 categories of comorbid illness
- Body chemistry (pH, WBC, $PaCO_2$, etc.)

Panel (7 specialists in clinical care) specified important variables related to the decision to use or not use a RHC.

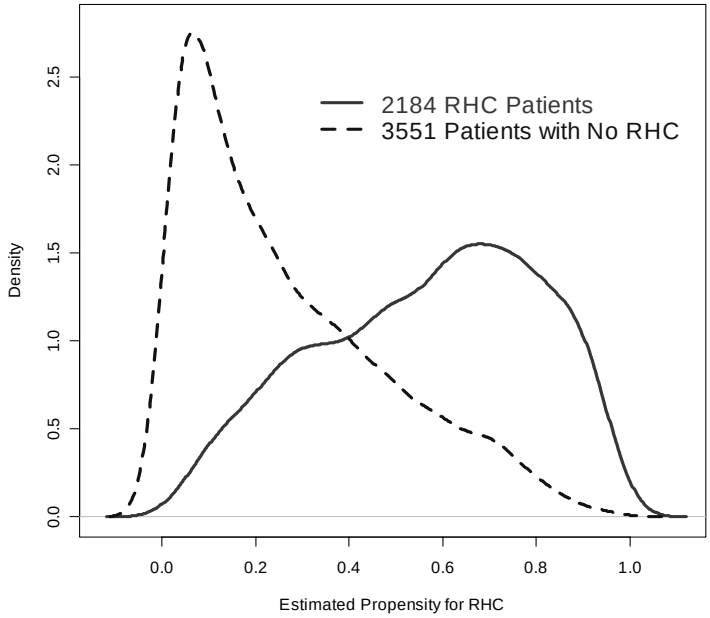
Boxplot of Propensity Score Overlap



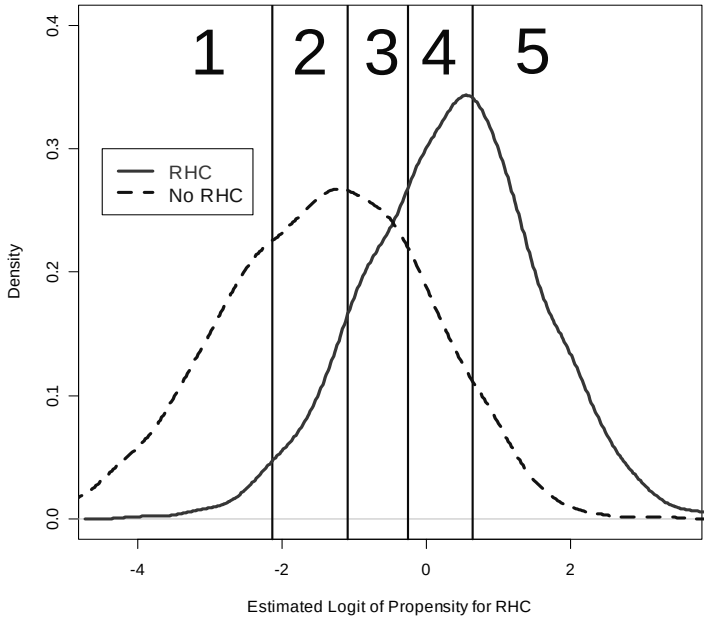
How Much Propensity Score Overlap Do We Want?



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.

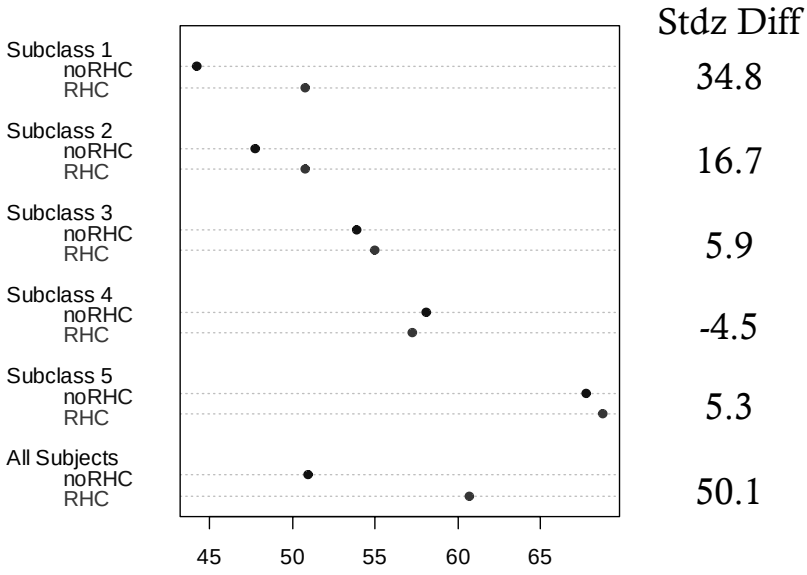
PS Subclass	Propensity Score ≈ Prob(RHC | covars.)	Actually got RHC	Actually got no RHC
5	Highest 1147 scores	888 (77%)	259 (23%)
4	2nd highest	635 (55%)	512 (45%)
3	Middle	398 (35%)	749 (65%)
2	2nd lowest	210 (18%)	937 (82%)
1	Lowest 1147 scores	53 (5%)	1094 (95%)

Subclassification & Covariate Balance

Age>80	Overall	S1	S2	S3	S4	S5
RHC (%)	7.6	18.9	11.0	10.8	7.2	5.1
noRHC (%)	14.1	22.2	14.1	11.1	5.9	4.6
Stzd. Diff.	**-20.8**	-8.3	-9.5	-0.9	5.6	2.0
***p*-value**	**< .00001**	0.73	0.26	0.92	0.40	0.87

Mean DBP	Overall	S1	S2	S3	S4	S5
RHC	68	94	86	78	70	57
noRHC	85	100	83	77	72	59
Stzd. Diff.	**-45.5**	-13.4	-5.4	3.1	-6.5	-6.2
***p*-value**	**< .00001**	0.33	0.47	0.62	0.28	0.37

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

How Should We Stratify?

- Hullsiek and Louis (2002) suggest balancing iteratively on the inverse variance of the subclass-specific treatment effects, rather than more simply on the number of observations within a subclass.
  - Having lots of subclasses reduces bias due to confounding.
  - Similar variances of treatment effect across subclasses maximizes effective bias control.

Hullsiek and Louis (2002)

Propensity Score Weighting

- Idea: Re-weight treated and control observations to make them representative of the population of interest
- A treated subject's weight is the inverse of its propensity score. $$w_i = \frac{1}{PS_i}$$
- A control subject's weight is the inverse of 1 minus its propensity score. $$w_i = \frac{1}{(1 - PS_i)}$$

Rubin (2001), Rosenbaum (1987), Lunceford and Davidian (2004)

Right Heart Catheterization

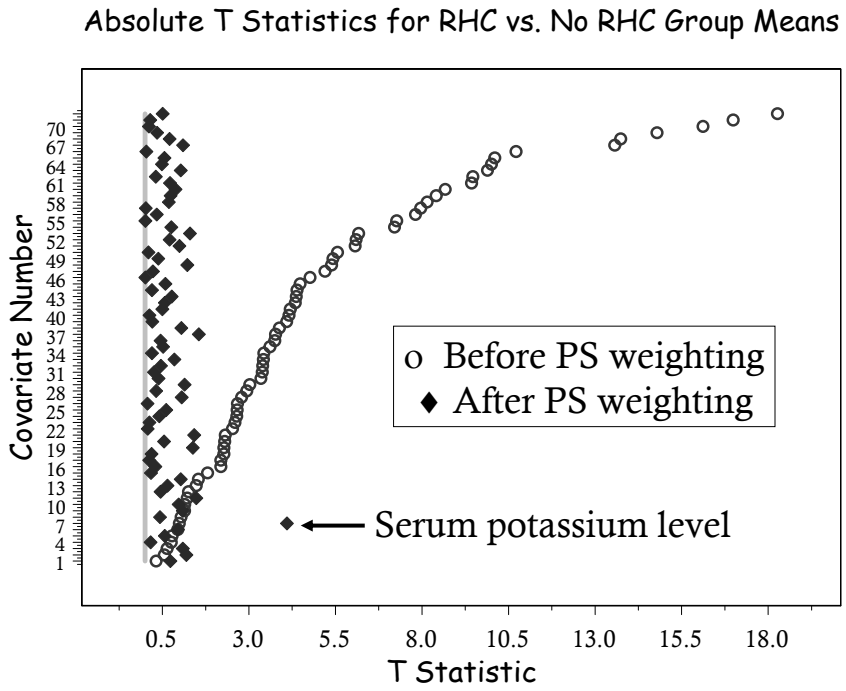
- 5735 hospitalized patients in the SUPPORT study: 2184 treated (RHC) and 3551 controls (no RHC).
- Reweight each treated patient by 1/PS, and each control patient by 1/(1-PS).
- PS model estimated by Hirano and Imbens using 57 of 72 available covariates (those with t > 2.0).

Hirano and Imbens (2001), Connors (1996), Hirano, Imbens and Ridder (2003)

RHC Covariate Balance Before/After Weighting on the Propensity Score

	Before Weighting			After Weighting		
Variable	No RHC	RHC	t	No RHC	RHC	t
Age	61.76	60.74	-2.28	61.25	61.15	-0.19
Sex	0.46	0.41	-3.42	0.44	0.43	-0.85
Edu	11.56	11.85	3.35	11.68	11.71	0.39
COPD	0.11	0.02	-13.6	0.07	0.06	-1.10
Album	3.16	2.97	-8.15	3.08	3.15	0.69
Potass	4.07	4.04	-0.99	4.15	3.97	-4.10

Hirano and Imbens (2001)



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.

Estimating the Treatment Effect

- Weighting: attractive theoretical properties, but rarely used (so far) in practice.
- Outcome model (logistic regression) is estimated using MLE with IPS weights

- Z is exposure assignment vector, X is vector of covariates for susceptibility adjustment, weights are as described above
- Estimated treatment effect:

Joffe et al. (2004) or Austin and Mamdani (in press) provide nice applications.

Using the Propensity Score in Direct Covariate Adjustment

- We sometimes use treatment indicator and the propensity score itself as predictors, along with other important covariates and risk adjusters.
- It's common to take a large set of covariates to form the PS, then use the PS and a subset of those covariates in the model for outcomes.
  - Analogous to matching within PS calipers, choosing the best match in terms of distance on key covariates from the target from among subjects inside the calipers.
- We could also regress within PS strata, etc.

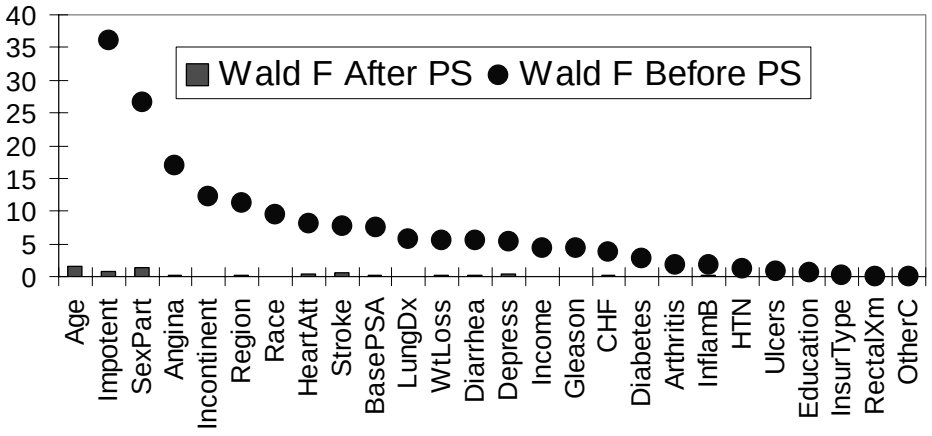
Prostate Cancer Surgery vs. Radiotherapy

- RP: Radical prostatectomy (n = 1156)
- RT: external beam radiotherapy (n = 435)
- Outcomes: relevant functioning, QOL
- PS = Pr(RP) model built for 26 covariates: used survey weights, and imputed missing data.
- Split into PS quintiles to check balance – no sig diffs. in covariate means, and reasonable overlap of PS.
- PS used as covariate in cross-sectional and longitudinal regressions comparing RP to RT.

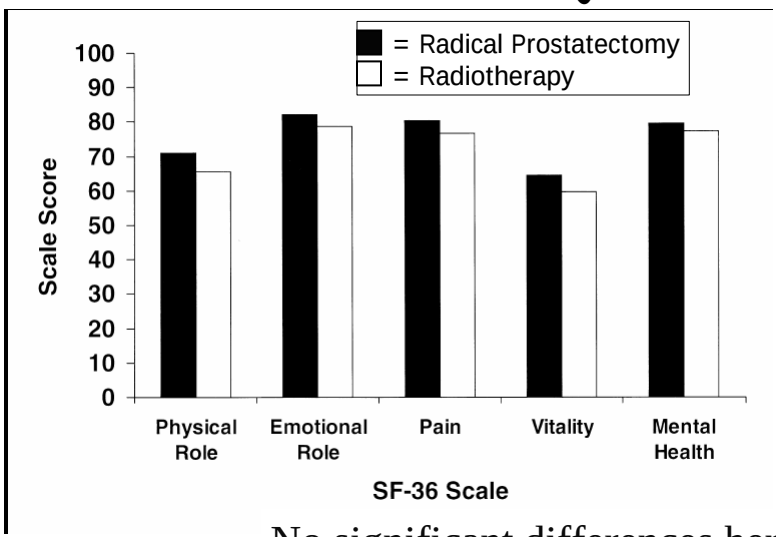
Potosky et al (2000)

131

Are the Covariates Balanced? Before/After PS Stratification



Comparing Outcomes Directly After PS and Covariate Adjustment



No significant differences here.

Could an Unmeasured Covariate Be Responsible for these Conclusions?

- 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.
- LTE: Logic, Theory & Empirical Evidence

Conclusions of Sensitivity Analysis in terms of an unobserved covariate

- When describing possible hidden bias, we refer to characteristics we did not observe, and therefore did not control for in the PS model.
- If our study was randomized, or somehow free of hidden bias, we would have strong evidence of a treatment effect.
- Structure of Argument: To explain away the observed effect, an unobserved covariate would need to increase the odds of exposure to treatment by more than a factor of ____.

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

Summary: Sensitivity Analysis

- Hidden bias is the great problem with observational studies, and for PS models.
- Sensitivity analysis can be applied to many statistical tests – we hope to find that an unobserved covariate would have to be very powerful to alter conclusions.
- This doesn't mean that such a covariate (or set of them) doesn't exist.

What Propensity Scores Can and Cannot Do

- If we match (subclassify) treated subjects to controls with similar propensity scores, we can behave as if they had been randomly assigned to treatments.
- Or, if we use regression to adjust for propensity to get treatment, we can compare treated to controls without worrying about the impact of baseline differences we've measured on selection to exposure.
- But 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.

Part Two

Estimating the Propensity Score Effectively

Building the Propensity Score Model

- If propensity scores are to be used in design to balance covariates, it is the distributional balance of those covariates that is achieved within blocks (strata, subclasses or matched pairs) that is the critical diagnostic tool.

Donald B. Rubin

Rubin (2004)

Should Adjustments Be Made for All Observed Covariates?

- If not, how should covariates be selected?
- No real reason to avoid adjustment for a true covariate – a variable describing subjects before treatment.
- In practice, though, this can increase both cost and complexity unnecessarily.
- There are issues of data quality and completeness to consider.

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.

Why Not Screen This Way?

- Doesn't consider relationship between covariate and outcome.
- No reason to believe that the absence of statistical significance implies the imbalance of the covariate is small enough to be ignored.
- Process considers covariates one at a time, while the adjustments will control the covariates simultaneously.

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.

Review of 47 Published Articles

- **Variable selection in the PS model**
  - 24 provided no information about variable selection
  - 6 used non-parsimonious logistic regression models
  - 7 identified variables through univariate sig. testing
  - 5 studies used stepwise or other selection algorithms
  - 4 studies used a priori (not data driven) selection
  - 1 study selected variables based on GOF tests
- **Interaction Inclusion Criteria**
  - 30 of 47 left use of interactions in PS model unclear
  - 12 clearly indicated that no interactions were used
  - 3 of the other 5 used p values to select interactions
  - 1 used improvement in discrimination of the PS model
  - 1 used improvement in balance across exposure groups

Weitzen et al (2004)

What People Have Done

- Adequacy of the Propensity Score Model
  - Goodness of Fit
    - 6 of the 47 studies considered the GoF of the PS model
    - 4 of the 6 provided the Hosmer-Lemeshow test p value
      - When the H-L was NS, one group didn't use the PS
    - 1 study used split-sample validation to evaluate fit
    - 1 study indicated model was "appropriately calibrated"
  - Discrimination
    - 18 of 47 studies reported C statistics (range: 0.52, 0.92)
  - Balance on Covariates
    - 22 of 47 included no information on covariate balance
    - In one article, when there was no balance in 3 of the 5 PS quintiles, they excluded patients from those quintiles

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.

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)

Dealing with Hidden Bias?

- Assess potential for unmeasured covariates to affect results with sensitivity analysis.
- The unmeasured covariate in question would have to be independent of all variables in the propensity model.
  - Either we missed a domain of the problem
  - Or our measure is so weak that we miss something important completely.

Does Propensity Matching Balance "Omitted" Covariates?

- We fit a published propensity model to data from the SUPPORT study on right heart catheterization, which used 82 covariates.
- Then we got data on 17 other covariates, not included in the propensity model.

Correlation with Propensity Score?	# of Covariates	Balance Improved After Match	Median Bias Reduction
Significant (α = .05)	10	9 (90%)	45%
Not Significant	7	2 (29%)	–36%

Love, Cebul, Thomas and Dawson (2003), Connors et al. (1996)

Conclusions: When is an omitted variable most dangerous?

- All observational studies are potentially affected by hidden bias. Sensitivity analyses are necessary in 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.

Missing Data and Generalized Propensity Scores

- Pattern of missing covariates can be prognostically important
- PS should, ideally, condition both on
  - Observed values of covariates, and
  - Observed missing-data indicators
- Generalized PS lead in expectation to balanced missing data patterns & covariate distributions in the treatment and control groups

D'Agostino and Rubin (2000), D'Agostino and Lang (2001)

Dealing with Missing Data: A Typical Approach

- MAR = assumed missing at random – mechanism by which data are missing is unrelated to information not in **X**.
  - Discrete: include binary "missing" x
  - Continuous: fit two predictors:
    [1] subject was measured/unmeasured,
    [2] if subject measured, then value.

Part Three

Assessing and Displaying Covariate Balance

Checking for Covariate Balance

Nursing Home Study

- Rx: Rehabilitation
- Control: No Rehab
- 21 covariates included in PS model (no interactions or polynomial terms)
- Matched patients with similar propensities for rehabilitation

Murray et al. (2003, 2005)

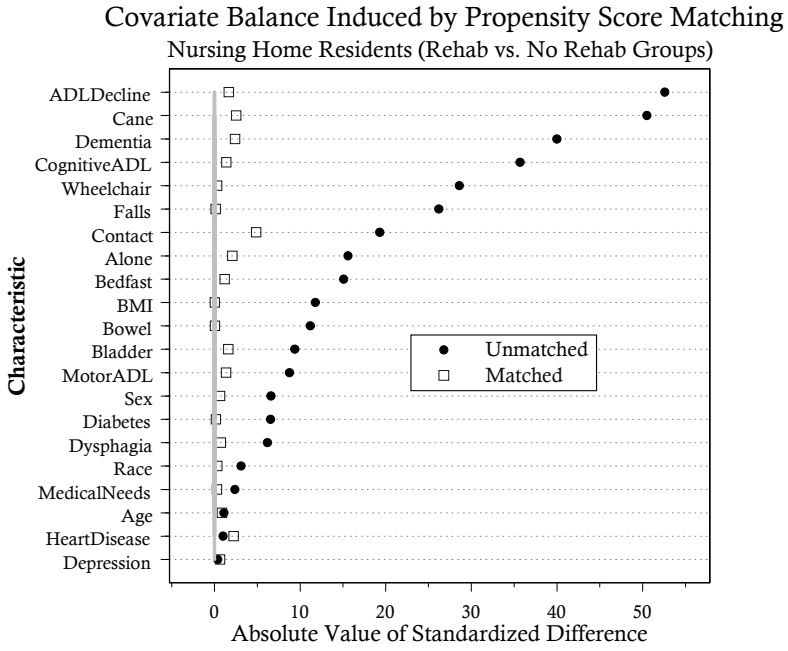
Variable	Rehab	No Rehab
Mean Age	78.1	78.0
% female	67	64
Mean BMI	24.2	23.6
Mean Cog ADL	79.6	71.0
% ath. ♥ disease	15	16
% dementia	20	38
% live alone	40	33
% with a cane	54	30
% recent ADL ⇓	66	41

Checking for Covariate Balance

Variable	Rehab	No Rehab	Matched Rehab	Matched No Rehab
Mean Age	78.1	78.0	78.1	78.0
% female	67	64	65	65
Mean BMI	24.2	23.6	23.6	23.6
Mean Cog ADL	79.6	71.0	75.0	74.6
% ath. ♥ disease	15	16	18	17
% dementia	20	38	28	29
% live alone	40	33	37	36
% with a cane	54	30	37	38
% recent ADL ⇓	66	41	52	52

Choices To Be Made

- How should I summarize balance within a covariate?
  - What if I have both continuous and categorical covariates?
  - Standardized differences
  - Significance test results
- How can I best present the summarized results across covariates?
- What are the key messages to get across?



Using Propensity Scores in Designing Observational Studies

- Specify treatments, outcomes, covariates.
- Collect treatment and covariate information, and model treatment assignment with the propensity score.
- Use propensity scores (through matching, stratification, reweighting) to reduce bias.
- Check for covariate balance across the treatment groups and iterate until happy…

Rubin (1997, 2001)

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.

When Can We Move On?

- 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.

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.

Comparing "Current Smokers" to "Never Smokers" using the National Medical Examination Survey

- Large nationally representative data base of nearly 30,000 adults, calendar year 1987
- Design Goal: Create samples of smokers and never smokers in NMES with the same multivariate distribution of covariates to assess causal effects
- Genders separately assessed. We focus on Male "Current Smokers" vs. Male "Never Smokers"
  - 3510 Male "Current Smokers" in the pool
  - 4297 Male "Never Smokers" as controls

Rubin (2001)

Propensity Model for "Current Smokers" vs. "Never Smokers"

- Logistic Regression with sampling weights
- 146 variables in the model including
  - Main effects
  - Quadratic effects
  - Interaction effects
- Separate models were built for "former" vs. "never" and for the female comparisons

Assessing the Degree of Overlap: Step 1: Looking for Mean Bias

- Bias B = standardized difference in the means of logit(propensity scores) between current smokers and never smokers for males
- We want the bias in the propensity score to be small, no greater than 0.50 in absolute value.
- Here, mean propensity score Bias B = 1.09
- In fact standardized difference > 0.5 for many of the individual covariates, as well.

Assessing the Degree of Overlap: Step 2: Looking at the Variance Ratio

- Ratio R = ratio of the variances of logit(propensity scores) between current smokers and never smokers for males.
- We want the variances to be homogeneous, so the ratio should be close to 1 (1/2 and 2 are far too extreme).
- Here, variance ratio for logit(PS) is R = 1.00
- Could look at ratio of covariate variances, also

Assessing the Degree of Overlap: Step 3: Comparing Residuals

- Regress each covariate on logit(PS) and look at ratio of variances of residuals for current smokers to variance of residuals for never smokers within the male population.
- Here, we get a separate result for each of the 146 covariates – we want results near 1.00
- 57% of covariates had $0.8 \leq$ resid ratio ≤ 1.2
- 5% of covariates had resid ratio > 2 or < 0.5

Mahalanobis Matching of Smokers & Never Smokers within PS Calipers

- For the 3510 male current smokers, 3510 "matching" male never smokers were chosen from the pool of 4297 male never smokers.
- Method: Mahalanobis metric matching within propensity score calipers (± 0.2 of a linear propensity score standard deviation).
- The four Mahalanobis distance variables used to do the matching were: age, education, body mass index, and sampling weight.

Impact of Matching on Overlap: Current vs. Never Smoking Males

- Before The Match

		Residual Variance Ratios					
B	R	≤.5	.5 - .8	.8 - 1.2	1.2 - 2	>2	Total
1.09	1.00	3	9	57	26	5	100

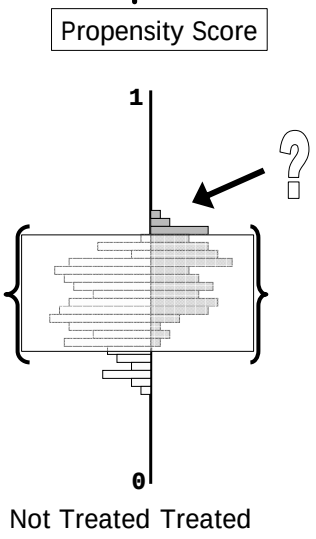
- After The Match

B	R	≤.5	.5 - .8	.8 - 1.2	1.2 - 2	>2	Total
0.08	1.16	1	3	90	6	0	100

What Do We Do If Some Treated Subjects Can't Be Matched Well?

- In this case, there were no "current smoker" Males that could not be matched within the PS calipers to "never smoker" Males.
- What if there had been a treated subject whose propensity score was not "matchable?"
- What if the "donor pool" of never smokers had been empty for one of the current smokers?

What if Treated and Untreated Groups Don't Overlap Completely?



- Inferences for the causal effects of treatment on the subjects with no overlap cannot be drawn without heroic modeling assumptions.
- Usually, we'd exclude these treated subjects, and explain separately.

Subclassification of Matched Samples of Smokers and Never Smokers

- Suppose we still weren't satisfied.
- Idea: Create two equal-size (weighted) subclasses, low and high on logit(PS).
- Treated and Control subjects with low PS are to be compared to each other.
- Treated and Control subjects with high PS are to be compared to each other.
- Weighted average of two comparisons = result.
- No reason to stop at just 2 subclasses, either…

Impact of Post-Matching Subclassification on Overlap

Current vs. never smoking **males**

			Residual Variance Ratios				
Subclasses	B	R	≤.5	.5 - .8	.8 - 1.2	1.2 - 2	>2
1	0.39	1.33	0	4	88	8	0
2	0.18	1.36	0	2	98	0	0
4	0.10	1.25	0	1	99	0	0
6	0.09	1.30	0	0	100	0	0
8	0.08	1.16	0	0	100	0	0
10	0.07	1.12	0	0	100	0	0

Why Can We Get Away With This?

- We can obtain dramatic reduction in initial bias through these propensity-based methods
- If substantial balance in covariates is obtained in this initial design stage, the exact form of the modeling adjustment is not critical.
- Similar treated and control covariate distributions implies only limited model-based sensitivity.
- Why doesn't this introduce a bias for our eventual conclusions and analytic results?
- Because we haven't got the outcomes.

What Should You Do About Residual Covariate Imbalance?

- Suppose a covariate appears seriously imbalanced after propensity matching.
- Could make a regression adjustment for that covariate after matching.
- Could use an additional or alternative measurement of the concept described by the covariate in the PS model.
- Consider re-matching starting with a different random order of treated patients, or by a different standard.
  - Consider Mahalanobis distance matching within propensity score calipers.

Part Four

Choosing Analytic Techniques

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.
  - Can do OPTIMAL full matching (see Bergstralh et al. macro at http://www.mayo.edu/hsr/sasmac.html)

Is Regression Adjustment for Observational Studies Obsolete?

- Matching and stratification are old and trusted methods of adjustment for observational studies, but the difficulty of implementing them led earlier practitioners to prefer regression.
- It is now possible to perform optimal full matching, and to achieve levels of bias reduction that were previously unattainable.

Hansen (2004)

Full Matching in a study of coaching for the SAT

- It has been tough to implement full matching in observational studies, even though it is the best in principle:
- Alignment of comparable treated and control subjects is as good as any alternate method, and potentially much better
- Hansen modifies full matching with modifications to minimize variance as well as bias
  - Optimal full matching removes as much as 99% of the bias along a PS on which treated and control means are separated by 1.1 SD's.
  - Reduces to insignificance biases along **27** covariates, while making use of more, not less, of the data than regression based analyses.

Hansen (2004)

SAT Coaching Study

- Survey of a random sample of 1995-1996 SAT test takers about their preparation
- 12% of respondents had completed extracurricular test preparation courses
- Matching looked unattractive to the original researchers due to significant reduction in sample size – but they only considered 1:1 matching.
- Do 1:k matching options look better?

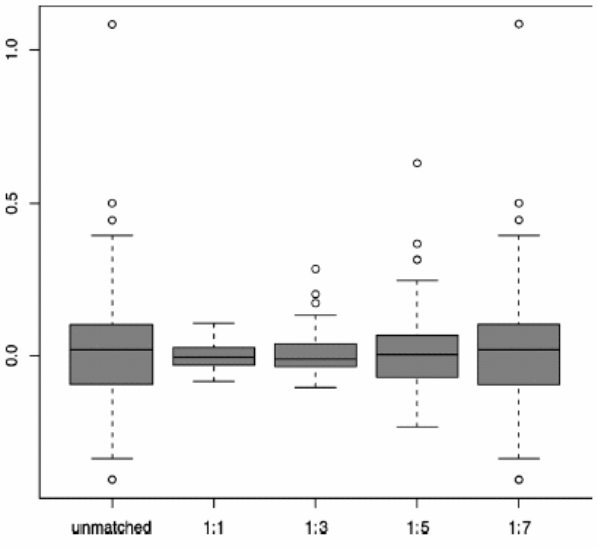
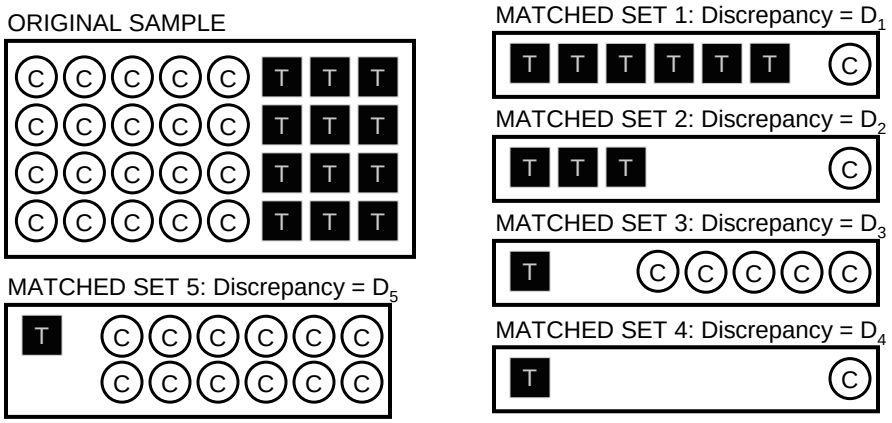


Figure 1. Covariate Imbalances in 1:k Matching. Each boxplot represents standardized biases in the 99 categories of the 27 categorical covariates along with standardized bias in the propensity score (which in each plot is the uppermost outlier). Strictly speaking, the matching represented at far right is not a 1:7 matching but a blend of six 1:6 and 494 1:7 matched sets.

Covariate Imbalances in 1:k Matching

- In all of these cases, we're using less data
- Still some imbalance

Optimal Full Matching



- OFM minimizes propensity score distances (discrepancies) while using all treated and all control subjects (i.e. discarding no units).
- Here, infinite distances force matches on Race×Sex

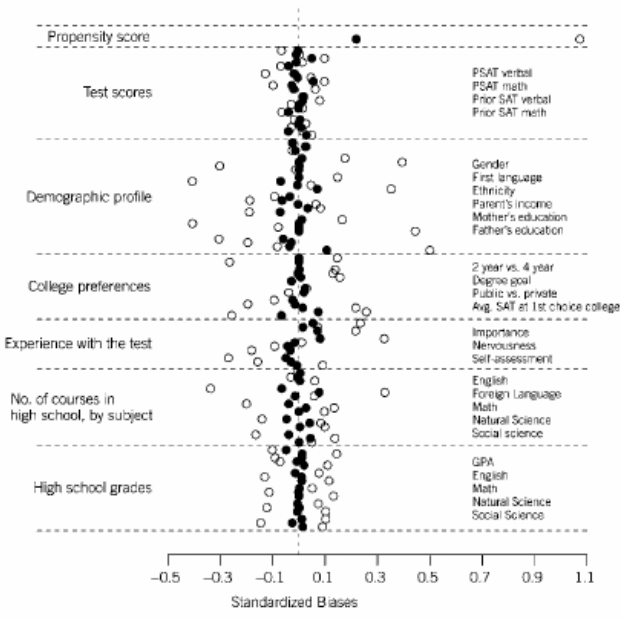


Figure 3. Standardized Biases Without Stratification or Matching, Open Circles, and Under the Optimal [.5, 2] Full Match, Shaded Circles.

Standardized Bias Plot

- Open circles are for standardized biases before matching
- Shaded circles describe results after full match

SAT Coaching Study Results

- Raw differences of treated and control group means were 41 points on Math and 9 on Verbal
- Full matching leads to aggregate contrasts of 26 points on Math and 1 point on the verbal.
  - Standard errors for these estimates are around 5 points.
- Surprised that Verbal effect is so small? Recall control condition is not "no prep at all" Estimated effect of treatment on the controls is 3 for Math and -8 on Verbal.
- Method doesn't require homogeneity of coaching effects – whether and to what degree coaching is beneficial appears to vary greatly across students

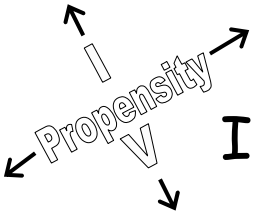
What About Instrumental Variables?

- Idea: Find a variable (the instrument)
  - strongly correlated with the treatment choice
  - but having no direct effect on the outcome (outside of the instrument's influence on treatment selection)
- If these two conditions are not met, then IV is not a useful approach.
- In health care, treatment selection is usually closely linked to outcome.

See Earle et al (2001), Landrum and Ayanian (2001), Posner et al (2001)

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.



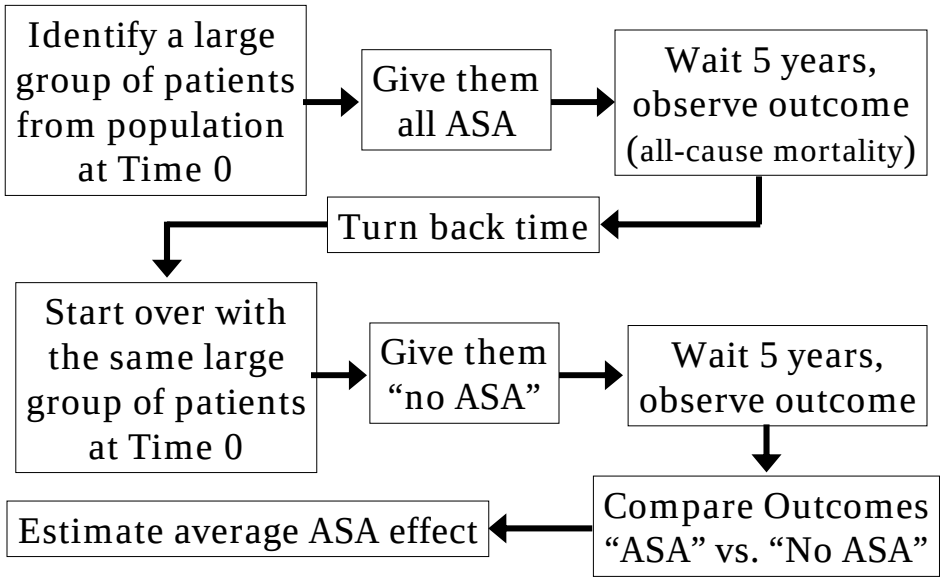
Propensity Scores vs. Instrumental Variables?

- 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).

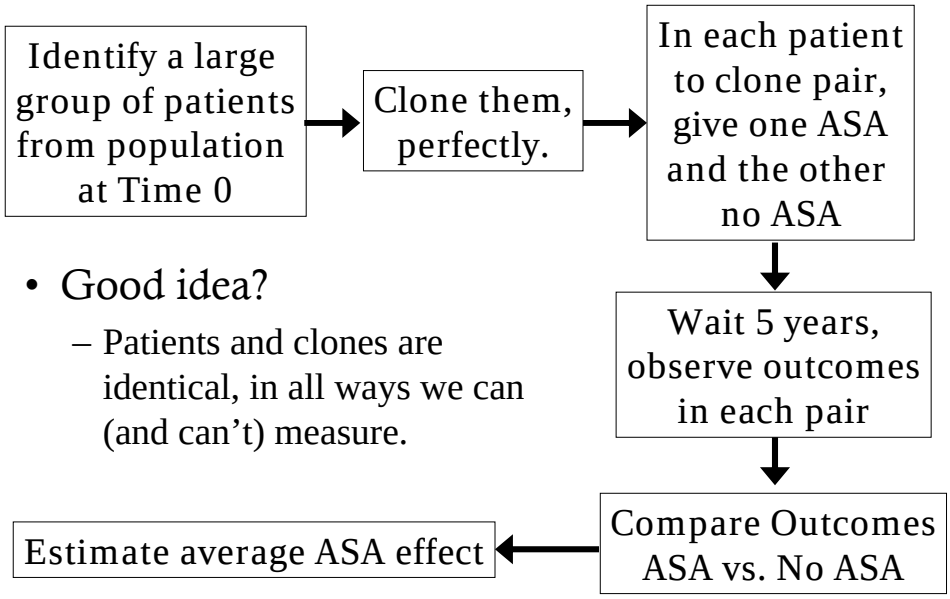
Part Five

Communicating Results to a Non-Statistical Audience

ASA and Mortality: Ideal Study

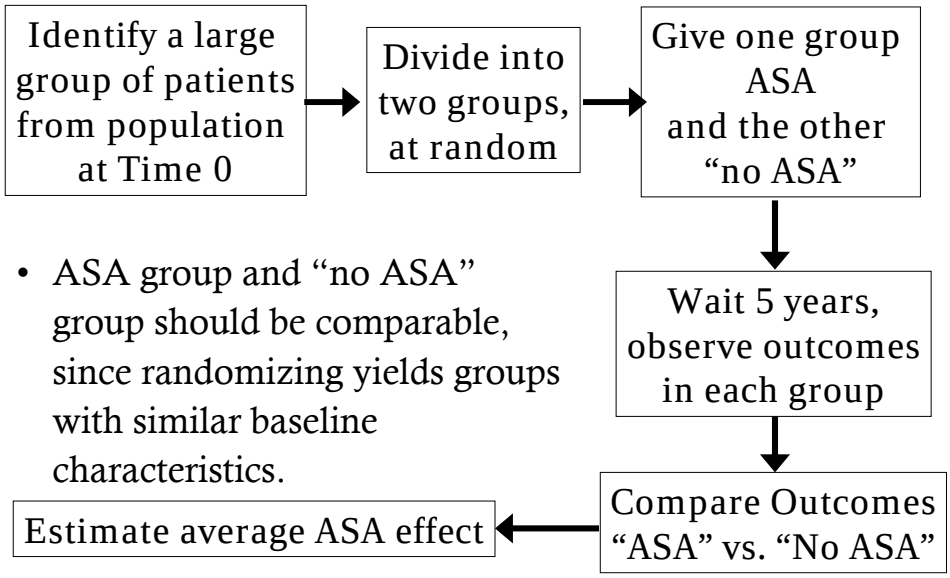


ASA & Mortality: Second Best Study



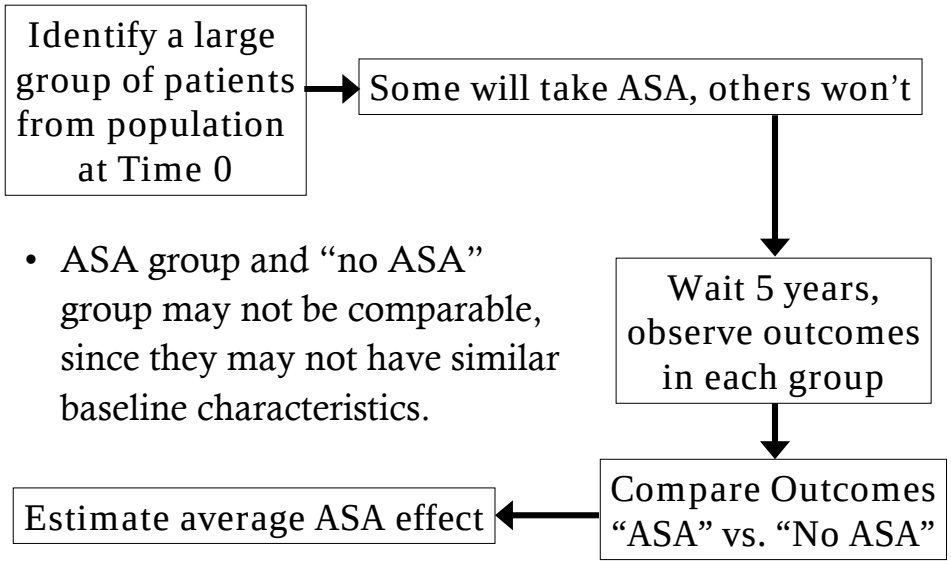
- Good idea?
  - Patients and clones are identical, in all ways we can (and can't) measure.

ASA & Mortality: RCT



- ASA group and "no ASA" group should be comparable, since randomizing yields groups with similar baseline characteristics.

ASA & Mortality: An Observational Study



- ASA group and "no ASA" group may not be comparable, since they may not have similar baseline characteristics.

Dealing With Selection Bias

- Analytical adjustments for Observed Bias
  - Matching, Stratification, Weighting, and Regression approaches are most common
- Goal: compare treated subjects to controls with same (or similar) baseline characteristics.
- Assuming no hidden bias (heroically), we use standard statistical comparisons of the groups after adjustment.

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.

A Few Advantages of Propensity 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?

A Few Cautions and Limitations

- Hidden Bias: Beware unmeasured covariates which affect outcomes and/or treatment assignment.
- This is a reasonable method with fairly large samples.
- Options narrow as an investigation proceeds. What is easy early may become difficult or impossible later.
- Sadly, though OS work cries out for design, we're often working with secondary data, where a lot of design issues weren't considered in a serious way…

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)

What should always be done in an OS … and often isn't?

1. Collect data so as to be able to model selection
2. Demonstrate need for adjustment - selection bias
3. Carefully record intervention time - adjust only for things present before or at time of intervention.
4. Ensure baseline characteristic overlap [comparability]
5. Check baseline characteristic balance after adjustment
6. Specify relevant post-adjustment population with care
7. Estimate treatment effect in light of adjustment
8. Estimate sensitivity of results to potential hidden bias

Thank you very much.

THOMASLOVE@case.edu

http://www.chrp.org/propensity/ICHPR

http://www.chrp.org/love

Rosenbaum (2002 text, Springer)

Rubin (2001, *Health Services & Outcomes Research Methodology*)

Austin & Mamdani (in press, *Statistics in Medicine*)

D'Agostino (1998, *Statistics in Medicine*)

References of Interest, Including Those Used in the Talk

1. Austin, P. C. and M. M. Mamdani (in press). "A comparison of propensity score methods: A case-study estimating the effectiveness of post-AMI statin use." *Statistics in Medicine In Press*. [Examines multiple approaches to propensity score modeling applied to the same data, including matching, stratification, covariate adjustment and weighting. Also looks at several ways of assessing residual confounding after stratification or matching.]
2. Austin, P. C., M. M. Mamdani, et al. (2005). "The use of the propensity score for estimating treatment effects: Administrative versus clinical data." *Statistics in Medicine 24*: 1563-1578. [Stratifying on PS quintiles derived from administrative data did not also successfully balance clinical patient characteristics. Also, clinical data-based estimates of treatment effectiveness were attenuated compared to estimates from administrative data.]
3. Barker, F. G., II, S. Amin-Hanjani, et al. (2004). "Age-Dependent Differences in Short-Term Outcome after Surgical or Endovascular Treatment of Unruptured Intracranial Aneurysms in the United States, 1996-2000." *Neurosurgery 54*: 18-30. [PS Application: Stratified treatment effect analysis by propensity score quintile. PS for surgical treatment modeled for 1780 patients, using demographics, clinical characteristics, primary payer for care, type of admission, and hospital characteristics.]
4. Barnard, J., C. E. Frangakis, et al. (2003). "Principal Stratification Approach to Broken Randomized Experiments: A Case Study of School Choice Vouchers in New York City." *Journal of the American Statistical Association 98*(462): 299-323, including Comments and Rejoinder. [Application of the "principal stratification" framework to the problems of noncompliance and missing data in a randomized trial of school vouchers.]
5. Barsky, R., J. Bound, et al. (2002). "Accounting for the Black-White Wealth Gap: A Nonparametric Approach." *Journal of the American Statistical Association 97*(459): 663-673. [Interested in the effect of being black on "wealth-holding" in the black population. Approach involves reweighting the white wealth distribution based on a non-parametric estimate of the propensity score.]
6. Beddhu, S., M. H. Samore, et al. (2003). "Impact of timing of initiation of dialysis on mortality." *J Am Soc Nephrol 14*(9): 2305-12. [PS Application: 3 PS strata created and applied to a series of stratified Cox PH models. Some sensitivity analyses to account for the impact of censoring.]
7. Berg, G. D., A. Johnson, et al. (2003). "Clinical and Utilization Outcomes for a Pediatric and Adolescent Telephonic Asthma Care Support Program: A Propensity Score-Matched Cohort Study." *Dis Manage Health Outcomes 11*(11): 737-743. [PS Application: PS matching to find a suitable comparison cohort for 318 asthma care support program participants based on demographics, utilization, and procedures performed - simple 1:1 matching followed by Kruskal-Wallis comparisons.]
8. Berger, V. W. (2005). "The reverse propensity score to detect selection bias and correct for baseline imbalances." *Statistics in Medicine 24*: 2777-2787. [Deals with the problem of potentially systematic baseline imbalance in individual blinded randomized trials. The reverse propensity score (defined within) permits detection of and correction for selection bias.]
9. Bingenheimer, J. B., R. T. Brennan, et al. (2005). "Firearm violence exposure and serious violent behavior." *Science 308*(5726): 1323-6. [PS Application: Examines longitudinal data on Chicago adolescents using a 12-way PS stratification (PS below 0.05 or above 0.75 then excluded) to

look at relationship between exposure to firearm violence and subsequent perpetration of serious violence.]

10. Braitman, L. E. and P. R. Rosenbaum (2003). "Rare Outcomes, Common Treatments: Analytic Strategies Using Propensity Scores." *Ann Intern Med 137*(8): 693-696. [Propensity Scores: What They Are, Why They Work, and What They Can't Do. With subsequent Letter from Longstreth WT, Fahrenbruch CE and Koepsell TD "Propensity Scores for Rare Outcomes and Common Treatments" v. 139, pp. 152-153 which describes (briefly) the use of PS to deal with imbalances observed in RCTs which need to be summarized with risk differences.]
11. Cannon, C. P., G. M. Gibson, et al. (2000). "Relationship of Symptom-Onset-to-Balloon Time and Door-to-Balloon Time With Mortality in Patients Undergoing Angioplasty for Acute Myocardial Infarction." *JAMA 283*: 2941-2947. [PS Application: PS used to address possible confounding - no real detail, but led to somewhat reduced risk increase with increases in door-to-balloon time. With Editorial by Michael S. Lauer "Primary Angioplasty - Time Is of the Essence" with several positive comments about the PS, also identifying need for lots of data and the fact that PS does nothing about hidden biases.]
12. Cepeda, M. S., R. Boston, et al. (2003). "Comparison of Logistic Regression versus Propensity Score When the Number of Events Is Low and There Are Multiple Confounders." *American Journal of Epidemiology 158*(3): 280-287. [Simulations comparing standard logistic regression for the outcome incorporating a group of covariates to standard logistic regression incorporating only the exposure variable and the quintiles of PS, based on the same covariates. Not that interesting a comparison, in my view. Conclusion was that # of events per confounder was the issue (small -> PS, >8 ->logistic regression).]
13. Christakis, N. A. and T. J. Iwashyna (2003). "The health impact of health care on families: a matched cohort study of hospice use by decedents and mortality outcomes in surviving, widowed spouses." *Soc Sci Med 57*(3): 465-75. [PS Application: PS matching used to match couples where decedent was in hospice to couples where decedent was not.]
14. Chu, J. H., S. J. Gange, et al. (2005). "Hormonal contraceptive use and the effectiveness of highly active antiretroviral therapy." *Am J Epidemiol 161*(9): 881-90. [PS Application: 1:1 PS match (actually they first stratified, found no mean differences within most strata in terms of PS, then selected matches at random from within the stratum, they used nearest-neighbor in the first stratum) of 77 hormonal contraceptive users to 77 non-users on the basis of demographics, clinical and lab characteristics, and prior use of health-care services.]
15. Clark, C. and A. R. Rich (2003). "Outcomes of homeless adults with mental illness in a housing program and in case management only." *Psychiatr Serv 54*(1): 78-83. [PS Application: PS subclassification to form high-, medium- and low-impairment groups based on psychiatric symptoms, alcohol, and drug use.]
16. Cohen, G. (2005). "Propensity-Score Methods and the Lenin School." *J of Interdisciplinary History 36*(2): 209-232. [PS Application: PS matching to assess whether study at the International Lenin School in Moscow increased the chances of British students assuming the leadership roles that the Communist International (Comintern) expected of them. Confronts in some detail the issue of estimating PS using non-ignorably missing data.]
17. Cole, J. A., J. E. Loughlin, et al. (2002). "The effect of zanamivir treatment on influenza complications: a retrospective cohort study." *Clin Ther 24*(11): 1824-39. [PS Application: PS matching used to identify comparison groups with similar health service utilization, comorbidities and diagnosis (influenza) but no antiviral therapy.]

18. Connors, A. F., Jr., T. Speroff, et al. (1996). "The effectiveness of right heart catheterization in the initial care of critically ill patients. SUPPORT Investigators." *Jama 276*(11): 889-97. [PS Application: Detailed matching and multivariate adjustment - state of the art in 1996, I'd argue.]
19. Copas, A. J. and V. T. Farewell (1998). "Dealing with Non-Ignorable Non-Response by Using an "Enthusiasm-To-Respond" Variable." *Journal of the Royal Statistical Society, Series A 161*(3): 385-396. [Development of a "propensity-to-respond" score based on an embarrassment assessment combined with demographics. Applied to an estimation of virginity from the National Survey of Sexual Attitudes and Lifestyles.]
20. Craske, M. G., D. Golinelli, et al. (2005). "Does the addition of cognitive behavioral therapy improve panic disorder treatment outcome relative to medication alone in the primary-care setting?" *Psychol Med 35*(11): 1645-54. [PS Application: PS Weighting, using a generalized boosted model built from confounders including study site, insurance, demographics, and various psychological and clinical measures. Only retained significant bivariate differences in PS model.]
21. Crown, W. H., R. L. Obenchain, et al. (1998). "The application of sample selection models to outcomes research: The case of evaluating the effects of antidepressant therapy on resource utilization." *Statistics in Medicine 17*: 1943-1958. [Illustration of sample selection models, which attempt to control hidden biases in treatment selection which are also correlated with the outcome of interest.]
22. D'Agostino, R., Jr. and D. B. Rubin (2000). "Estimating and using propensity scores with partially missing data." *Journal of the American Statistical Association 95*: 749-759.
23. D'Agostino, R. B., Jr. (1998). "Propensity Score Methods for Bias Reduction in the Comparison of a Treatment to a Non-Randomized Control Group." *Statistics in Medicine 17*: 2265-2281. [Very readable introduction to matching, stratification, and regression using propensity scores, with useful examples. Very good "first paper" to read.]
24. D'Agostino, R. B., Jr., W. Lang, et al. (2001). "Examining the Impact of Missing Data on Propensity Score Estimation in Determining the Effectiveness of Self-Monitoring of Blood Glucose (SMBG)." *Health Services & Outcomes Research Methodology 2*: 291-315. [Assesses several PS-based and ANCOVA techniques for dealing with missing data in estimating the PS, and evaluating the outcome.]
25. Dehejia, R. H. (2004). "Practical propensity score matching: A reply to Smith and Todd." *J of Econometrics 125*: 355-364. [Discusses PS matching (1:1) in response to Smith and Todd's paper on matching for non-experimental data. Points out that a separate PS specification should be estimated for each comparison that you want to make, and that sensitivity analyses are very important, as well.]
26. Dehejia, R. H. and S. Wahba (2002). "Propensity Score-Matching Methods for Nonexperimental Causal Studies." *The Review of Economics and Statistics 84*(1): 151-161. [Focuses on settings where few controls are comparable to treated subjects. Uses data from the National Supported Work experiment and PS matching, comparing the estimates of the treatment obtained from matching on the PS to the benchmark results obtained from the associated experiment.]
27. Earle, C. C., J. S. Tsai, et al. (2001). "Effectiveness of chemotherapy for advanced lung cancer in the elderly: instrumental variable and propensity analysis." *J Clin Oncol 19*(4): 1064-70. [Interesting combination of instrumental variables and propensity score analyses.]
28. El-Bassel, N., L. Gilbert, et al. (2005). "Relationship between drug abuse and intimate partner violence: a longitudinal study among women receiving methadone." *Am J Public Health 95*(3): 465-70. [PS Application: PS Matching after multiple imputation using demographics, history of trauma, psychological distress, social support and HIV risks. Data were

collected in three waves: confounders measured pre-treatment, while outcomes were measured post-treatment.]

29. Foody, J. M., C. R. Cole, et al. (2001). "A propensity analysis of cigarette smoking and mortality with consideration of the effects of alcohol." *Am J Cardiol 87*(6): 706-11. [PS Application: PS incorporated as a continuous variable in Cox PH models, after dropping largest and smallest quintiles.]
30. Foster, E. M. (2003). "Propensity Score Matching: An Illustrative Analysis of Dose Response." *Medical Care 41*(10): 1183-1192. [Use of PS methods for health services researchers, with particular attention to thinking about PS matching as a weighting for dealing with more than two exposure groups as recommended by Imbens et al.]
31. Greenland, S. (2000). "An introduction to instrumental variables for epidemiologists." *International Journal of Epidemiology 29*: 722-729. [Introductory discussion of instrumental variables, for epidemiologists, with an application dealing with non-parametric adjustment for non-compliance in randomized trials.]
32. Greenland, S. and H. Morgenstern (2001). "Confounding in Health Research." *Annu. Rev. Public Health 22*: 189-212. [Describes confounding the context of a counterfactual model of causation - source of bias in the estimation of causal effects.]
33. Gu, X. S. and P. R. Rosenbaum (1993). "Comparison of Multivariate Matching Methods: Structures, Distances, and Algorithms." *J of Computational and Graphical Statistics 2*(4): 405-420. [Optimal matching beats greedy (nearest available) matching, sometimes by quite a lot, in getting closely matched pairs, but not by much in terms of balancing the matched samples. Optimal full matching is substantially better than Optimal 1:k matching. Paper also compares several means of defining distances.]
34. Gum, P. A., M. Thamilarasan, et al. (2001). "Aspirin Use and All-Cause Mortality Among Patients Being Evaluated for Known or Suspected Coronary Artery Disease: A Propensity Analysis." *JAMA 286*: 1187-1194. [PS Application: PS matching with non-proportional hazards model for survival analysis - results show an interesting impact of selection bias adjustment With Editorial by MJ Radford and JM Foody "How Do Observational Studies Expand the Evidence Base for Therapy?" pp. 1228-1230 - editorial motivates observational studies vs RCT discussion, discusses clustering, hierarchical models]
35. Hahn, J. (1998). "On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects." *Econometrica 66*(2): 315-331.
36. Hansen, B. B. (2004). "Full Matching in an Observational Study of Coaching for the SAT." *Journal of the American Statistical Association 99*(467): 609-618.
37. Harraf, F., A. K. Sharma, et al. (2002). "A multicentre observational study of presentation and early assessment of acute stroke." *BMJ 325*: 17-21.
38. Heckman, J. J., N. Hohmann, et al. (2000). "Substitution and Dropout Bias in Social Experiments: A Study of an Influential Social Experiment." *The Quarterly Journal of Economics CXV*: 651-694.
39. Heckman, J. J., H. Ichimura, et al. (1997). "Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme." *The Review of Economic Studies 64*(4): 605-654.
40. Henry, A. D., A. M. Lucca, et al. (2004). "Inpatient hospitalizations and emergency service visits among participants in an Individual Placement and Support (IPS) model program." *Ment Health Serv Res 6*(4): 227-37.
41. Heuschmann, P. U., K. Berger, et al. (2003). "Frequency of thrombolytic therapy in patients with acute ischemic stroke and the risk of in-hospital mortality: the German Stroke Registers

Study Group." *Stroke 34*(5): 1106-13. [PS Application: Patients matched by propensity for tPA therapy in the context of comparing risk of inpatient death.]

42. Hill, A. B., D. Obrand, et al. (2000). "Hemispheric Stroke following Cardiac Surgery: A Case-Control Estimate of the Risk Resulting from Ipsilateral Asymptomatic Carotid Artery Stenosis." *Ann Vasc Surg 14*: 200-209. [PS Application: Case-control study with PS stratification - discussion includes a meta-analysis of related prospective cohort studies.]
43. Hirano, K. and G. W. Imbens (2001). "Estimation of Causal Effects using Propensity Score Weighting: An Application to Data on Right Heart Catheterization." *Health Services & Outcomes Research Methodology 2*: 259-278.
44. Hirano, K., G. W. Imbens, et al. (2003). "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score." *Econometrica 71*(4): 1161-1189.
45. Ho, D. E., K. Imai, et al. (2005). "MatchIt: Nonparametric Preprocessing for Parametric Causal Inference."
46. Hodges, K. and H. Grunwald (2005). "The use of propensity scores to evaluate outcomes for community clinics: identification of an exceptional home-based program." *J Behav Health Serv Res 32*(3): 294-305.
47. Holden, C. (2005). "Sociology. Controversial study suggests seeing gun violence promotes it." *Science 308*(5726): 1239-40. [Comment on Bingenheimer et al.]
48. Huang, I.-C., C. Frangakis, et al. (2005). "Application of a Propensity Score Approach for Risk Adjustment in Profiling Multiple Physician Groups on Asthma Care." *Health Services Research 40*(1): 253-278.
49. Hullsiek, K. H. and T. A. Louis (2002). "Propensity score modeling strategies for the causal analysis of observational data." *Biostatistics 2*(4): 179-193.
50. Huncharek, M., B. Kupelnick, et al. (2002). "Maternal smoking during pregnancy and the risk of childhood brain tumors: a meta-analysis of 6566 subjects from twelve epidemiological studies." *J Neurooncol 57*(1): 51-7.
51. Ichimura, H. and C. Taber (2001). "Propensity-Score Matching with Instrumental Variables." *The American Economic Review 91*(2): 119-124.
52. Imai, K. and D. A. van Dyk (2004). "Causal Inference with General Treatment Regimes: Generalizing the Propensity Score." *Journal of the American Statistical Association 99*(467): 854-866.
53. Imbens, G. W. (2000). "The role of the propensity score in estimating dose-response functions." *Biometrika 87*(3): 706-710.
54. Jaar, B. G., J. Coresh, et al. (2005). "Comparing the Risk for Death with Peritoneal Dialysis and Hemodialysis in a National Cohort of Patients with Chronic Kidney Disease." *Ann Intern Med 143*: 174-183. [Comment by Gerald Schuman, "Mortality and Treatment Modality of End-Stage Renal Disease"]
55. Joffe, M. M. and P. R. Rosenbaum (1999). "Invited commentary: propensity scores." *Am J Epidemiol 150*(4): 327-33. [Motivation for PS approaches geared to physician-epidemiologists, as well as a set of extensions - to case/control studies and to dose/response issues]
56. Joffe, M. M., T. R. Ten Have, et al. (2004). "Model Selection, Confounder Control, and Marginal Structural Models: Review and New Applications." *The American Statistician 58*(4): 272-279.
57. Jones, A. S., R. B. D'Agostino, Jr., et al. (2004). "Assessing the Effect of Batterer Program Completion on Reassault Using Propensity Scores." *J of Interpersonal Violence 19*(9): 1002-1020.

58. Katzan, I. L., R. D. Cebul, et al. (2003). "The effect of pneumonia on mortality among patients hospitalized for acute stroke." *Neurology 60*(4): 620-5. [PS Application: PS for pneumonia used for risk adjustment in logistic regression analyses.]
59. Kazmi, W. H., G. T. Obrador, et al. (2004). "Late nephrology referral and mortality among patients with end-stage renal disease: a propensity score analysis." *Nephrol Dial Transplant 19*(7): 1808-14.
60. Kennedy, J., H. Quan, et al. (2005). "Statins Are Associated With Better Outcomes After Endarterectomy in Symptomatic Patients." *Stroke 36*: 2072-2076.
61. Kilian, R. and M. C. Angermeyer (2005). "The effects of antipsychotic treatment on quality of life of schizophrenic patients under naturalistic treatment conditions: an application of random effect regression models and propensity scores in an observational prospective trial." *Qual Life Res 14*(5): 1275-89.
62. Kleiman, N. S., A. M. Lincoff, et al. (2000). "Early percutaneous coronary intervention, platelet inhibition with eptifibatide, and clinical outcomes in patients with acute coronary syndromes. PURSUIT Investigators." *Circulation 101*(7): 751-7.
63. Klungel, O. H., E. P. Martens, et al. (2004). "Methods to assess intended effects of drug treatment in observational studies are reviewed." *J Clin Epidemiol 57*(12): 1223-31.
64. Koch, C. G., F. Khandwala, et al. (2003). "Gender and outcomes after coronary artery bypass grafting: a propensity-matched comparison." *J Thorac Cardiovasc Surg 126*(6): 2032-43.
65. Kraemer, H. C., E. Stice, et al. (2001). "How Do Risk Factors Work Together? Mediators, Moderators, and Independent, Overlapping, and Proxy Risk Factors." *Am J Psychiatry 158*: 848-856.
66. Kubal, C., A. K. Srinivasan, et al. (2005). "Effect of risk-adjusted diabetes on mortality and morbidity after coronary artery bypass surgery." *Ann Thorac Surg 79*(5): 1570-6.
67. Landrum, M. B. and J. Z. Ayanian (2001). "Causal Effect of Ambulatory Specialty Care on Mortality Following Myocardial Infarction: A Comparison of Propensity Score and Instrumental Variable Analyses." *Health Services & Outcomes Research Methodology 2*: 221-245. [Direct comparison of propensity score and instrumental variable approaches.]
68. Landrum, M. B., B. J. McNeil, et al. (1999). "Understanding variability in physician ratings of the appropriateness of coronary angiography after acute myocardial infarction." *J Clin Epidemiol 52*(4): 309-19.
69. Lash, T. L., R. A. Silliman, et al. (2000). "The effect of less than definitive care on breast carcinoma recurrence and mortality." *Cancer 89*(8): 1739-47.
70. Lauby, J. L., P. J. Smith, et al. (2000). "A community-level HIV prevention intervention for inner-city women: results of the women and infants demonstration projects." *Am J Public Health 90*(2): 216-22. [PS Application: PS subclassification and regression runs/residuals for change score analyses.]
71. Leon, A. C. and D. Hedeker (2005). "A mixed-effects quintile-stratified propensity adjustment for effectiveness analyses of ordered categorical doses." *Statistics in Medicine 24*: 647-658.
72. Lieberman, E., J. M. Lang, et al. (1996). "Association of epidural analgesia with cesarean delivery in nulliparas." *Obstet Gynecol 88*(6): 993-1000.
73. Linden, A., J. L. Adams, et al. (2005). "Using Propensity Scores to Construct Comparable Control Groups for Disease Management Program Evaluation." *Dis Manage Health Outcomes 13*(2): 107-115.

74. Lindenauer, P. K., P. Pekow, et al. (2005). "Perioperative Beta-Blocker Therapy and Mortality after Major Noncardiac Surgery." *N Engl J Med 353*: 349-361. [Comment by Don Poldermans and Eric Boersma, "Beta-Blocker Therapy in Noncardiac Surgery"]
75. Logar, C. M., L. M. Pappas, et al. (2005). "Surgical revascularization versus amputation for peripheral vascular disease in dialysis patients: a cohort study." *BMC Nephrology 6*: 3.
76. Loughlin, J. E., J. A. Cole, et al. (2002). "Comparison of Resource Utilization by Patients Treated with Transdermal Fentanyl and Long-Acting Oral Opioids for Nonmalignant Pain." *Pain Medicine 3*(1): 47-55. [PS matching, hypothesis tests for differences in costs, visits, etc.]
77. Love, T. E., R. D. Cebul, et al. (2003). "Effect of matching for propensity on the balance of 'unmeasured' covariates (Abstract)." *J Clin Epidemiol 56*(9): 920.
78. Lu, B. (2005). "Propensity Score Matching with Time-Dependent Covariates." *Biometrics 61*: 721-28.
79. Lu, B., E. Zanutto, et al. (2002). "Matching with doses in an observational study of a media campaign against drug abuse." *Journal of the American Statistical Association 96*: 1245-1253.
80. Lunceford, J. K. and M. Davidian (2004). "Stratification and weighting via the propensity score in estimation of causal treatment effects: A comparative study." *Statistics in Medicine 23*: 2937-2960.
81. McCaffrey, D. F., G. Ridgeway, et al. (2004). "Propensity Score Estimation With Boosted Regression for Evaluating Causal Effects in Observational Studies." *Psychological Methods 9*(4): 403-425.
82. Mitra, N. and A. Indurkhya (2005). "A propensity score approach to estimating the cost-effectiveness of medical therapies from observational data." *Health Economics 14*: 805-815.
83. Mitra, N., F. R. Schnabel, et al. (2001). "Estimating the Effect of an Intensive Surveillance Program on Stage of Breast Carcinoma at Diagnosis: A Propensity Score Analysis." *Cancer 91*: 1709-1715.
84. Morant, S. V., D. Pettitt, et al. (2004). "Application of a propensity score to adjust for channelling bias with NSAIDs." *Pharmacoepidemiology and Drug Safety 13*: 345-353.
85. Mortensen, E. M., M. I. Restrepo, et al. (2005). "The effect of prior statin use on 30-day mortality for patients hospitalized with community-acquired pneumonia." *Respiratory Research 6*: 82. [Quote: "A propensity score technique was used to balance covariates associated with statin use between groups {ref to Stone Obrosky et al (1995) *Med Care*.} The use of the propensity score technique provides a way, in non-randomized studies, to control for pretreatment differences by defining sets of comparable patients. The propensity score was derived from a logistic regression model. A dichotomous indicator variable indexing whether a patient was on a statin was our response variable. The covariates used in the propensity score model were the pneumonia severity index score (which includes comorbid conditions such as congestive heart failure, liver disease, and history of stroke), history of alcoholism, history of diabetes mellitus, coronary artery disease, and current tobacco use. Variables were entered, and maintained, in the model if they had a p-value <0.20 in the univariate analysis (with statin use as the dependent variable) and had a p-value <0.20 in the final model."]
86. Moustaki, I. and M. Knott (2000). "Weighting for item non-response in attitude scales by using latent variable models with covariates." *Journal of the Royal Statistical Society, Series A 163*(3): 445-459.
87. Murray, P. K., T. E. Love, et al. (2005). "Rehabilitation services after the implementation of the nursing home prospective payment system: differences related to patient and nursing home characteristics." *Med Care 43*(11): 1109-15.

88. Murray, P. K., M. Singer, et al. (2003). "Outcomes of rehabilitation services for nursing home residents." *Arch Phys Med Rehabil 84*: 1129-1136. [PS Application: PS for rehab using 112 variables, with matching and multivariate adjustment (including dose-response) with detailed evaluation of effectiveness of the PS model and sensitivity analyses.]
89. Nash, I. S., R. R. Corrato, et al. (1999). "Generalist Versus Specialist Care for Acute Myocardial Infarction." *Am J Cardiol 83*: 650-654.
90. Normand, S.-L. T., M. B. Landrum, et al. (2001). "Validating recommendations for coronary angiography following acute myocardial infarction in the elderly: a matched analysis using propensity scores." *J Clin Epidemiol 54*(4): 387-98. [PS Application: PS Matching with calipers - interesting discussion of covariate balance (standardized differences) - outcome assessment through odds ratios, Mantel-Haenszel and survival rates.]
91. O'Connor, C. M., E. J. Velazquez, et al. (2002). "Comparison of Coronary Artery Bypass Grafting Versus Medical Therapy on Long-Term Outcome in Patients With Ischemic Cardiomyopathy (A 25-Year Experience from the Duke Cardiovascular Disease Databank)." *Am J Cardiol 90*: 101-107.
92. O'Keefe, S. (2004). "Job creation in California's enterprise zones: A comparison using a propensity score matching model." *J of Urban Economics 55*: 131-150.
93. Oo, A. Y., A. D. Grayson, et al. (2004). "Effect of training on outcomes following coronary artery bypass graft surgery." *Eur J Cardiothorac Surg 25*(4): 591-6.
94. Petersen, L. A., S. L. Normand, et al. (2000). "Outcome of myocardial infarction in Veterans Health Administration patients as compared with medicare patients." *N Engl J Med 343*(26): 1934-41. [PS Application: Presents both PS matching and the standard multivariate adjustments (through logistic regression).]
95. Peterson, J. G., E. J. Topol, et al. (2001). "Prognostic importance of concomitant heparin with eptifibatide in acute coronary syndromes. PURSUIT Investigators. Platelet Glycoprotein IIb/IIIa in Unstable Angina: Receptor Suppression Using Integrilin Therapy." *Am J Cardiol 87*(5): 532-6. [Dropped top two PS quintiles in this secondary analysis of an RCT - then included PS in Cox PH models.]
96. Posner, M. A., A. S. Ash, et al. (2001). "Comparing Standard Regression, Propensity Score Matching, and Instrumental Variables Methods for Determining the Influence of Mammography on Stage of Diagnosis." *Health Services & Outcomes Research Methodology 2*: 279-290. [Comparison of PS matching, instrumental variables, and standard risk adjustment techniques for dealing with selection bias.]
97. Potosky, A. L., J. Legler, et al. (2000). "Health outcomes after prostatectomy or radiotherapy for prostate cancer: results from the Prostate Cancer Outcomes Study." *J Natl Cancer Inst 92*(19): 1582-92. [PS Application: Direct PS adjustment in cross-sectional and longitudinal regression analyses - conclusions section motivates a study of hidden biases.]
98. Reiter, J. (2000). "Using Statistics to Determine Causal Relationships." *The American Mathematical Monthly 107*(1): 24-32.
99. Robins, J. M. (1998). "Correction for non-compliance in equivalence trials." *Statistics in Medicine 17*: 269-302.
100. Rosenbaum, P. R. (1987). "Model-Based Direct Adjustment." *Journal of the American Statistical Association 82*(398): 387-394.
101. Rosenbaum, P. R. (1991). "Sensitivity Analysis for Matched Case-Control Studies." *Biometrics 47*(1): 87-100.
102. Rosenbaum, P. R. (1999). "Choice as an Alternative to Control in Observational Studies." *Statistical Science 14*(3): 259-278.

103. Rosenbaum, P. R. (1999). "Reduced Sensitivity to Hidden Bias at Upper Quantiles in Observational Studies with Dilated Treatment Effects." *Biometrics 55*(2): 560-564.
104. Rosenbaum, P. R. (2001). "Effects attributable to treatment: Inference in experiments and observational studies with a discrete pivot." *Biometrika 88*(1): 219-231.
105. Rosenbaum, P. R. (2002). Observational Studies. New York, Springer. [The major recent work on the subject of observational studies. Packed with useful stuff.]
106. Rosenbaum, P. R. (2005). "Heterogeneity and Causality: Unit Heterogeneity and Design Sensitivity in Observational Studies." *The American Statistician 59*(2): 147-152.
107. Rosenbaum, P. R. and D. B. Rubin (1983). "Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome." *Journal of the Royal Statistical Society, Series B 45*: 212-218.
108. Rosenbaum, P. R. and D. B. Rubin (1984). "Reducing Bias in Observational Studies Using Subclassification on the Propensity Score." *Journal of the American Statistical Association 79*(387): 516-524. [A seminal paper - propensity score subclassification in detail.]
109. Rosenbaum, P. R. and D. B. Rubin (1985). "The Bias Due to Incomplete Matching." *Biometrics 41*(1): 103-116. [A seminal paper - some thoughts on sensitivity analysis, and a detailed argument about why statisticians want to match everyone.]
110. Rosenbaum, P. R. and D. B. Rubin (1985). "Constructing a Control Group Using Multivariate Matched Sampling Methods That Incorporate the Propensity Score." *The American Statistician 39*(1): 33-38.
111. Rubin, D. B. (1997). "Estimating Causal Effects from Large Data Sets Using Propensity Scores." *Ann Intern Med 127*(Part 2): 757-763. [A very readable introductory discussion of propensity score issues.]
112. Rubin, D. B. (2001). "Using Propensity Scores to Help Design Observational Studies: Application to the Tobacco Litigation." *Health Services & Outcomes Research Methodology 2*: 169-188.
113. Rubin, D. B. (2004). "On principles for modeling propensity scores in medical research." *Pharmacoepidemiology and Drug Safety*.
114. Rubin, D. B. (2004). "On principles for modeling propensity scores in medical research." *Pharmacoepidemiol Drug Saf 13*(12): 855-7.
115. Rubin, D. B. and N. Thomas (1992). "Characterizing the Effect of Matching Using Linear Propensity Score Methods with Normal Distributions." *Biometrika 79*(4): 797-809.
116. Rubin, D. B. and N. Thomas (1996). "Matching Using Estimated Propensity Scores: Relating Theory to Practice." *Biometrics 52*(1): 249-264.
117. Rubin, D. B. and N. Thomas (2000). "Combining propensity score matching with additional adjustments for prognostic covariates." *Journal of the American Statistical Association 95*: 573-585.
118. Seeger, J. D., P. L. Williams, et al. (2005). "An application of propensity score matching using claims data." *Pharmacoepidemiology and Drug Safety 14*: 465-476.
119. Shah, B. R., A. Laupacis, et al. (2005). "Propensity score methods gave similar results to traditional regression modeling in observational studies: a systematic review." *J Clin Epidemiol 58*(6): 550-9.
120. Shlipak, M. G., W. S. Browner, et al. (2001). "Comparison of the effects of angiotensin converting-enzyme inhibitors and beta blockers on survival in elderly patients with reduced left ventricular function after myocardial infarction." *Am J Med 110*(6): 425-33. [PS Application: PS models for patients on beta-blockers, ACE inhibitors, or both, vs. neither (control).

Classified by quintiles, yielding multiple regressions. With Editorial by MS Lauer "Medical Therapy for Coronary Artery Disease Works, Even (Especially) in the Real World"]

121. Silber, J. H., P. R. Rosenbaum, et al. (1995). "Comparing the Contributions of Groups of Predictors: Which Outcomes Vary with Hospital Rather Than Patient Characteristics." *Journal of the American Statistical Association 90*(429): 7-18.
122. Silber, J. H., P. R. Rosenbaum, et al. (2001). "Multivariate matching and bias reduction in the surgical outcomes study." *Med Care 39*(10): 1048-64. [Development of multivariate PS-matched case-control pairs using Medicare claims data.]
123. Smith, H. L. (1997). "Matching with Multiple Controls to Estimate Treatment Effects in Observational Studies." *Sociological Methodology 27*: 325-353.
124. Smith, J. A. and P. E. Todd (2005). "Does matching overcome LaLonde's critique of nonexperimental estimators?" *J of Econometrics 125*: 305-353.
125. Sturmer, T., S. Schneeweiss, et al. (2005). "Adjusting effect estimates for unmeasured confounding with validation data using propensity score calibration." *Am J Epidemiol 162*(3): 279-89.
126. Sturmer, T., S. Schneeweiss, et al. (2005). "Analytic strategies to adjust confounding using exposure propensity scores and disease risk scores: nonsteroidal antiinflammatory drugs and short-term mortality in the elderly." *Am J Epidemiol 161*(9): 891-8.
127. Tan, Z. (2005). "A Distributional Approach for Causal Inference Using Propensity Scores."
128. Ten Have, T. R., M. R. Elliott, et al. (2004). "Encouragement Trials in Treating Primary Care Depression." *Journal of the American Statistical Association 99*(465): 16-25.
129. Tu, W. and X.-H. Zhou (2003). "A Bootstrap Confidence Interval Procedure for the Treatment Effect Using Propensity Score Subclassification." *Health Services & Outcomes Research Methodology 3*: 135-147.
130. Wang, J., P. T. Donnan, et al. (2001). "The multiple propensity score for analysis of dose-response relationships in drug safety studies." *Pharmacoepidemiology and Drug Safety 10*: 105-111.
131. Weitzen, S., K. L. Lapane, et al. (2004). "Principles for modeling propensity scores in medical research: a systematic literature review." *Pharmacoepidemiol Drug Saf 13*(12): 841-53.
132. Weitzen, S., K. L. Lapane, et al. (2005). "Weaknesses of goodness-of-fit tests for evaluating propensity score models: the case of the omitted confounder." *Pharmacoepidemiol Drug Saf 14*(4): 227-38.
133. Winkelmayer, W. C. and T. Kurth (2004). "Propensity scores: help or hype?" *Nephrol Dial Transplant 19*: 1671-1673.
134. Zacharias, A., T. A. Schwann, et al. (2005). "Is hospital procedure volume a reliable marker of quality for coronary artery bypass surgery? A comparison of risk and propensity adjusted operative and midterm outcomes." *Ann Thorac Surg 79*(6): 1961-9.