Society for Medical Decision Making
Boston Annual Meeting

Short Course #9:
Reducing Bias in Observational Studies: Propensity Methods

October 15, 2006 from 9 AM to 12:30 PM

Thomas E. Love, Ph. D., Course Director, thomas.love@case.edu
Randall D. Cebul, M. D. rdc@case.edu
Neal V. Dawson, M. D. nvd@case.edu

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

Additional Information Available at http://www.chrp.org/propensity

9:00 – 9:30	**Implications of Selection Bias in Observational Studies** Randall D. Cebul, M. D. rdc@case.edu
9:30 – 10:00	**Epidemiologic Principles: Bias** Neal V. Dawson, M. D. nvd@case.edu
10:00 – 10:45	**Doing Propensity Analysis: Basics and Matching** Thomas E. Love, Ph. D. thomaslove@case.edu
10:45 – 11:00	**Break**
11:00 – 12:00	**Stratification, Adjustment, Sensitivity and Strategy** Thomas E. Love, Ph. D. thomaslove@case.edu
12:00 – 12:30	**Discussion**

COURSE EXPECTED OUTCOMES:

All participants will gain theoretical and practical understanding of propensity score methods and their relevance to EMR-centered observational studies. Interested researchers will develop the skills to apply propensity scores to adjust for selection bias in their own modeling research.

Short Course #9:
Reducing Bias in Observational Studies: Propensity Methods

Presentations

COURSE BACKGROUND AND RATIONALE:

The increasing adoption of electronic medical records (EMRs) will accelerate the use of observational studies of the effects of treatments on subjects. Decision modelers frequently use data from observational studies to estimate treatment or exposure effect sizes, occasionally embedding relevant multivariate risk models to individualize results for specific patient subgroups. Our desire to avoid potentially biased decisions or policies requires us to understand and account for the inherent selection bias in observational studies. This session provides a state-of-the-art introduction to Propensity Methods for reducing selection bias and suggests approaches, using contemporary EMR systems, to better capture the selection process. Results from a variety of medical and health policy applications will be reviewed, and participants will be asked to suggest factors associated with selection in specific clinical problem areas. Course participants should have a basic understanding of epidemiologic principles and multivariate statistical techniques.

COURSE GOALS:

1. To introduce the conceptual and mathematical frameworks of selection bias and propensity methods.
2. To introduce the main analytical techniques associated with propensity score methods (multivariable regression adjustment, matching and stratification using the propensity score, as well as sensitivity analysis) and describe key formats for reporting results.
3. To identify opportunities for better capturing the selection process using contemporary EMR systems.

Part One: Implications of Selection Bias in Observational Studies

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

Part One
Implications of Selection Bias in Observational Studies

Randall D. Cebul, MD
rdc@case.edu
Case Western Reserve University
MetroHealth Medical Center
October 15, 2006

1

Introduction & Overview

- **What is selection bias?**
- **Why should we care about it?**
- **Why don't standard risk-adjustment methods always work?**
- **What methods exist to adjust for it?**
- **Overview of Course Parts 2-4**

2

What is Selection Bias?

- **Selection bias – the deviation in the *measured effect* of an *exposure* that is due to its non-random occurrence or assignment.**

… may occur *whenever we are unable to randomly assign who receives the exposure and who does not*; e.g., all Observational Studies

- **Examples: institutional databases (electronic medical records), patient registries (SEER), administrative data bases (MEDPAR), cohort studies (Framingham, RN Health Study)**

3

Instructor: **Randall D. Cebul, MD**
RDC@case.edu

Part One: Implications of Selection Bias in Observational Studies

Why should we care about Selection Bias?

- **Observational data sources are increasingly common and rich, yet they are *still observational***
- **We use Observational Studies increasingly as data sources for DAs and CEAs**
- **Selection creates the potential for:**
  - **Biased effect sizes (e.g., RR of Rx)**
  - **Biased thresholds & cost-effectiveness ratios**
  - **BAD POLICY**

4

What is the Potential Direction and Magnitude of Selection Bias?

As compared to unbiased data and random assignment, selection bias may result in:

- **Reversal of effect**
- **Blunting of effect**
- **Accentuation of effect**

5

Example: Reversal of Effect

HRT and Coronary Heart Disease

6

Instructor: **Randall D. Cebul, MD**
RDC@case.edu

Observational Studies Embedded in DA:

'Hormone Replacement Therapy Should Be Recommended for Nearly All Women'

Col NF, Eckman MH, *etal*. Patient-specific decisions about hormone replacement therapy in postmenopausal women. *JAMA.* 1997; 277: 1140-47.

7

Randomized Controlled Trial:

'Do not use estrogen/progestin to prevent chronic disease'[1,2]

"Estrogens with or without progestins should not be used for the prevention of cardiovascular disease"[3]

1. **Fletcher SW, Colditz G. Failure of estrogen plus progestin therapy for prevention. *JAMA.* 2002; 288: 366-67.**
2. **Writing group for the Women's Health Initiative. Risks and benefits of estrogen plus progestin in healthy postmenopausal women. *JAMA.* 2002; 288: 321-333.**
3. **Wyeth Product information insert. 2006**

8

Relative Risk for Coronary Heart Disease:
Observational Studies vs. RCT

	Relative Risk of HRT	
	Obs Studies*	WHI/RCT
CHD	0.60 (0.20-0.60)	1.29 (1.02 – 1.63)

In: Col NF *et al. JAMA*. 1997; 277:1140-47. From: Nurses' Health Study. *Am J Epidemiology.* 1995; 14:57; and Weiss NS *et al. NEJM.* 1980; 303: 1195.

9

Instructor: **Randall D. Cebul, MD**
RDC@case.edu

Part One: Implications of Selection Bias in Observational Studies

How Could Selection Reverse the Effect?

- **If factors favoring selection for the exposure (in this case, Use of HRT) protect against CHD, the estimated effect (protection) may be opposite the actual effect (harm).**
- **What factors favoring HRT use might also be associated with good CHD outcomes?**

10

Selection Bias: "Healthy User" effect and compliance biases

Table 1. **Potential Explanations for Discordant Findings from Randomized Trials and Observational Studies Regarding Postmenopausal Hormone Therapy and Coronary Heart Disease.**

Methodologic differences
- Confounding ("healthy user") bias
- Compliance bias
- Incomplete capture of early clinical events

} **Selection**

Biologic differences
- Hormone regimen (formulation and dose)
- Characteristics of study population (endogenous estrogen level, time since menopause, and stage of atherosclerosis)

Goldstein F, Clarkson TB, Manson JE. *N Engl J Med.* 2003; 348: 645-650.

11

How Could Selection Reverse the Effect?

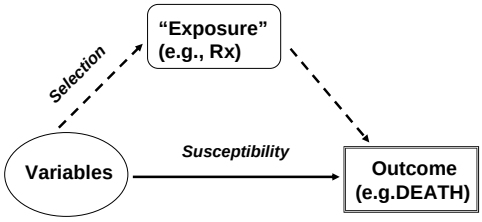
- **What factors favoring HRT use might also be associated with good CHD outcomes?**
- **e.g.: Was the higher SES, thin, white, jogging, non-smoker more likely to use HRT than the lower SES non-white sedentary obese smoker?**
- **Many of these factors are SOCIAL in nature, and were incompletely captured using standard risk-adjustment methods.**

12

Instructor: **Randall D. Cebul, MD**
RDC@case.edu

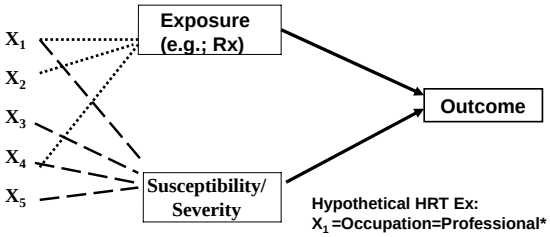
Part One: Implications of Selection Bias in Observational Studies

Isolating the Effect of Exposure: Standard risk-adjustment methods model severity/susceptibility for the outcome



13

Variables May Have Different Relationships to Exposure than to Outcome



Hypothetical HRT Ex:
X_1 =Occupation=Professional*
X_2= Lower BMI*
* Lambert *etal. J Epidem Comm Health* 2003; 57:294-300.

14

What Methods Exist to Adjust for Selection Bias in Observational Studies?

- Stratified analyses or careful selection of cohort sub-groups
- Strong expected influence of clinicians or systems of care: cluster analysis or hierarchical modeling
- Strong influence of economists: Instrumental variables
- Good data on factors likely to affect selection that also may affect outcome: Propensity Analysis

15

Part One: Implications of Selection Bias in Observational Studies

Designing an Observational Study: Comments of Lincoln Moses[1]

"If, after consideration, a data base approach still is chosen, then let the undertaking be carefully planned, with participation …from future users of the results.

Finally, it is time to respond to a challenge left us by David Byer: *arrange to record why* the patient is being given the therapy selected. This information should be a powerful adjustment variable…"

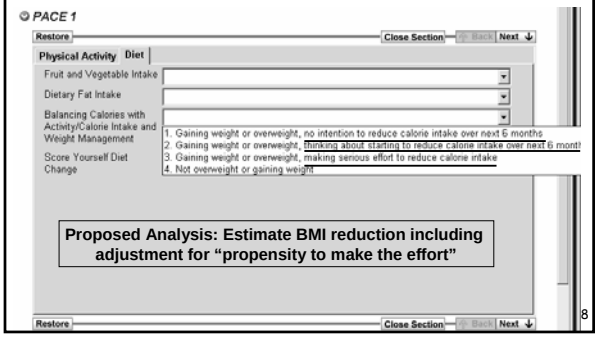
1. Measuring effects without randomized trials? Options, problems, challenges. *Med Care.* 1995; 33: AS8-AS14.

16

What Selection-related Data Would You Collect in an OS to Estimate the Effectiveness of a Weight Management Program?

17

Tailored EMR intake form for overweight patients referred to Weight Management –Incorporating Prochaska's Readiness-to-change Qs



8

Part Two: Epidemiological Principles: Bias in Estimating Effects

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

Part Two
Epidemiological Principles

Neal V. Dawson, MD
nvd@case.edu
Case Western Reserve University
MetroHealth Medical Center
October 15, 2006

1

Goals: To Understand the Following

- Need for risk adjustment
- Dimensions of risk
- Methods of adjusting for selection bias
- How adjusting for selection bias can be helpful in risk adjustment strategies
- How sensitivity analysis can help inform results from observational studies

2

Objectives: To Be Able To Discuss

- Seven key aspects of research architecture
- Five dimensions of risk
- The contribution of selection bias to the likelihood of susceptibility bias in observational studies
- How to use a propensity score in the evaluation of a treatment or exposure in an observational study
- The value of using sensitivity analysis in an observational study

3

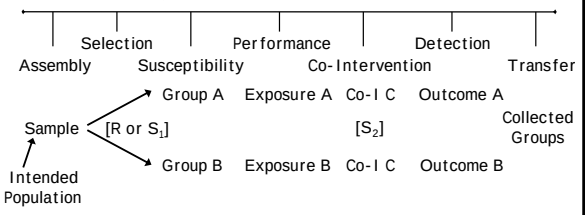
Part Two: Epidemiological Principles: Bias in Estimating Effects

Bias in Estimating Effects

- Distorted Assembly
- Selection bias
- Susceptibility bias
- Performance bias
- Co-Interventions (opportunity for selection)
- Outcome bias
- Transfer bias

4

7 Key Aspects of Research Architecture*: Judgments about Causation

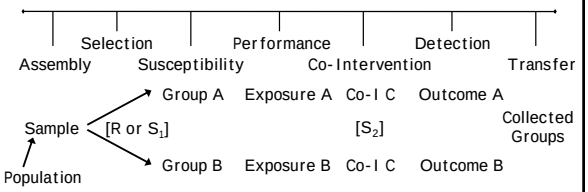


- Goal: Comparability of the groups who did and did not receive the exposure of interest (except for the actual receipt of the exposure)

*Adapted from Feinstein (Five key aspects)

5

7 Key Aspects of Research Architecture*: Judgments about Causation



- **Overt Bias** – adjustment techniques using measured covariates.
- **Hidden Bias** – Assess the potential for affecting results with sensitivity analysis
  - Create a false exposure effect (FP)
  - Masking a true exposure effect (FN)
  - Right direction – mis-estimation of effect

6

Instructor: **Neal V. Dawson, MD**
NVD@case.edu

Part Two: Epidemiological Principles: Bias in Estimating Effects

Feinstein's Intellectual Model for the evaluation of the scientific quality of cause-effect research

Distorted Assembly

- Defines the subjects who will subsequently be compared
- The sample assembled should reflect the population to which results will be generalized
  - If this is not true, then distorted assembly has occurred
- Subjects considered eligible for the study and
- Application of specific inclusion/exclusion criteria will determine the pool of baseline characteristics of the sample

7

Selection Bias

- Process determining who receives the exposure being compared
  - Bias can occur when subjects are selected to receive a given exposure (or co-intervention)
  - Exposure is based on their baseline characteristics (covariates) and
  - Covariates are related to different likelihoods of outcome

8

Selection Bias, continued

- Unmeasured covariates may or may not be associated with the measured baseline characteristics
- Contrasted with situations where patients receive a given exposure by a random process
  - Expectation that exposed and unexposed groups will have approximately equal distributions of covariates (measured and unmeasured).

9

Instructor: **Neal V. Dawson, MD**
NVD@case.edu

Part Two: Epidemiological Principles: Bias in Estimating Effects

Susceptibility Bias: Case Mix/Severity

- Susceptibility reflects the comparability of baseline characteristics of Group A and Group B
- Bias occurs when persons receiving the exposures being compared have importantly different expectations, at baseline, of the outcome of interest
- Differences in expectations of outcomes are a function of importantly different baseline characteristics

10

Performance Bias

- Comparison of how "well" patients receive Exposure A or Exposure B
- Bias occurs if the exposures being compared are not applied with the same proficiency
- Relates to differences in dosage schedules, compliance rates, etc.

11

Co-Interventions (opportunity for selection)

- Additional medical interventions (beyond the exposure of interest) that may influence the likelihood of achieving the outcome of interest

12

Instructor: **Neal V. Dawson, MD**
NVD@case.edu

Outcome Bias

- Comparison of outcomes in Group A with outcomes in Group B
- Bias occurs if the process for determining the status of the outcome of interest in each Group is applied unequally
- Arises from different surveillance, diagnostic testing, or diagnostic interpretation

13

Transfer Bias

- Comparison of members of "original" cohorts of subjects in Groups A and B
- Bias can occur when members of the original or complete cohorts of A and B are lost
  - Dropouts
  - Intra-study exclusions
  - Crossovers
  - During statistical manipulations (e.g., because of missing values of measured covariates in some types of regressions)

14

Dimensions of Risk

- Step 1:
  Delineate the goals of risk adjustment
- Risk
  - For whom?
  - Over what time period?
  - For what outcome?

15

Instructor: **Neal V. Dawson, MD**
NVD@case.edu

Part Two: Epidemiological Principles: Bias in Estimating Effects

Types of Variables to Consider: Susceptibility

- Age
- Gender
- Race/ethnicity
- Socioeconomic status
- Culture
- Genetic characteristics
- Acute clinical stability

16

Types of Variables to Consider: Susceptibility

- Principal diagnosis (case mix)
- Severity and chronicity of principal diagnosis
  - Clinical features:symptoms/signs
  - Pathological features: extent and location
- Extent, severity and type of comorbid disease

17

Types of Variables to Consider

- Function
  - Physical
  - Cognitive
  - Psychological
  - Psychosocial
- Health status
- Quality of Life
- Attitudes and preferences

18

Instructor: **Neal V. Dawson, MD**
NVD@case.edu

Part Two: Epidemiological Principles: Bias in Estimating Effects

Propensity Score

Pr(exposure given a set of baseline characteristics)

- Definition: the conditional probability of receiving a particular exposure given a vector of measured covariates
- Represents the likelihood that a person with a given set of baseline characteristics (covariates) will receive the exposure of interest
- Reduces an entire set of baseline characteristics to a single composite characteristic that adequately summarizes the collection of characteristics
- Multivariable logistic regression is often used to fit the model (although other statistical models can be used)

19

Use of a Propensity Score

- Allows a straightforward evaluation of the degree of overlap of baseline characteristics between exposed and unexposed groups
- Allows one to answer an important yet frequently unasked question…

  "Is this degree of overlap sufficient for sensible comparisons between effects related to exposure and non-exposure?"
- Propensity scores work better in larger samples (expected balance of covariates)

20

Use of a Propensity Score, Continued

- Can be used to produce strata (subclassification or stratification) with the same likelihood of exposure among those who actually did and did not receive the exposure of interest
  - Subclassification doesn't depend on any particular functional form (such as linearity), whereas models do
  - Generally, five (or more) strata are enough to remove ≥ 90% of bias

21

Instructor: **Neal V. Dawson, MD**
NVD@case.edu

Part Two: Epidemiological Principles: Bias in Estimating Effects

Propensity Scores

- Can be used for matching patients
- Can be used in a multivariable adjustment model
- Can develop and use more than one propensity score (e.g., one for primary treatment and one for a major co-intervention)
- **Outcome =**
  propensity + susceptibility + exposure + noise

22

Sensitivity Analysis for an Unmeasured Covariate

- Propensity scores can only adjust known (measured) covariates
- A priori – include measures of all known important predictors of selection for exposure (prospectively collect data on covariates)
- Estimates are made of the effects that unmeasured covariates (with varying levels of association with selection and outcome) would have on study results
- Allows one to assess the likely stability of observational study results

23

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)

25

Part Two: Epidemiological Principles: Bias in Estimating Effects

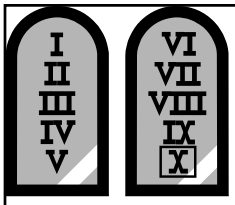
Thou Shalt Perform Bootstrap Analyses To Assess Shrinkage

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

26



Thou Shalt Ignore Commandments 1 through 9... And Instead Simply Ensure That The Model Adequately Balances The Covariates

27

Summary of Propensity

- 1. Model of selection: characteristics associated with assignment
- 2. Assure sufficient overlap of propensity for reasonable comparisons
- 3. Ensure that the model adequately balances the covariates

28

Part Three: Doing Propensity Analysis: Basics and Matching

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

Part Three
Doing Propensity Analysis: Basics and Matching

Thomas E. Love, Ph. D.
thomaslove@case.edu
Case Western Reserve University
MetroHealth Medical Center
October 15, 2006

1

The Propensity Score
Pr(treatment given covariates)

- Definition: The conditional probability of receiving a given exposure (treatment) given a vector of measured covariates.
- Usually estimated using logistic regression:

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

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

2

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

More examples: D'Agostino (1998), Austin and Mamdani (2006)

3

Instructor: **Thomas E. Love, PhD**
thomaslove@case.edu

Part Three: Doing Propensity Analysis: Basics and Matching

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)

4

Using Propensity Scores to Match Subjects

- Can matching on a scalar propensity score remove most of the bias due to a large pool of covariates?
- How should subjects be matched?
- Did the matching "work"? Can we demonstrate / display how well it worked?
- What's the tradeoff between matching inexactly and matching incompletely?

5

Aspirin Use and Mortality

- 6174 consecutive adults undergoing stress echocardiography for evaluation of known or suspected coronary disease.
- 2310 (37%) were taking aspirin (treatment).
- Main Outcome: all-cause mortality
- Median follow-up: 3.1 years
- 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)

Gum PA et al. (2001)

6

Instructor: **Thomas E. Love, PhD**
thomaslove@case.edu

Part Three: Doing Propensity Analysis: Basics and Matching

Propensity Score Model for Aspirin Use

- Logistic Regression predicting aspirin use
- 31 covariates included in the model:
  - Demographics, Clinical history, Medication use
  - Cardiovascular assessment and Exercise capacity
- Estimated propensity scores for aspirin use range from .03 to .98
  - ROC Area shows good discrimination (C = .83)
- But does the PS work?
- Are the covariates balanced?

7

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.

8

Matching with Propensity Scores

- For each patient, we have a propensity score.
- Randomly select an Aspirin user.
- Match to the non-user with closest propensity score (within some limit or "calipers")
- Eliminate both patients from pool, and repeat until you can't find an acceptable match.
  - Could match a non-user with Propensity Score inside "calipers" who matches exactly on characteristic X, or…
  - Match non-user with Propensity score inside "calipers" and smallest "distance" on some pre-specified covariates.

9

Instructor: **Thomas E. Love, PhD**
thomaslove@case.edu

Part Three: Doing Propensity Analysis: Basics and Matching

Matching on Gender within PS Calipers

- Shuffle "treatment" patients, and select one.
- Find all "non-treated" with PS inside calipers (here we'll set calipers at treated PS ± .03).
- Match patient within calipers of same gender.
- Repeat until no more matches are possible.

.80
.79
.78
.77
.76
.75
.74
.73
.72

Patient	Exposure	PS	Gender
A	**Treated**	**.76**	**Male**
B	Not Treated	.77	Female
C	Not Treated	.74	Male
D	Not Treated	.80	Male

10

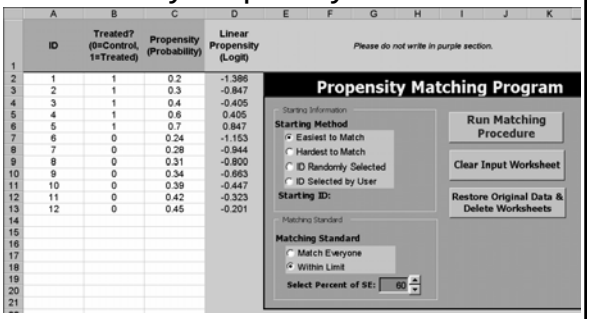
How Were The Aspirin Subjects Matched?

- 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

11

Excel Spreadsheet to Match Subjects by Propensity Score



Love TE and Husak SS – available at www.chrp.org/propensity

12

Instructor: **Thomas E. Love, PhD**
thomaslove@case.edu

Part Three: Doing Propensity Analysis: Basics and Matching

Propensity Matcher Results

ID	Treated?	Propensity	Linear Propensity	Match?	Partner ID
1	1	0.2	-1.386	No	-999
2	1	0.3	-0.847	Yes	8
3	1	0.4	-0.405	Yes	10
4	1	0.6	0.405	No	-999
5	1	0.7	0.847	No	-999

SE (Linear Propensity):	0.1829
x % Selected:	0.6
x % of SE:	0.1097

13

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]

14

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), Ahmed et al. (2006)]

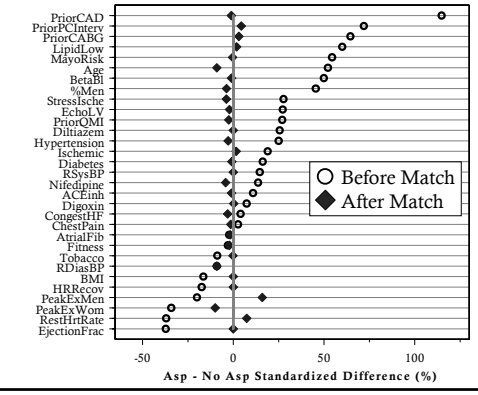
	Aspirin	No Aspirin	P	Std. D.
Before Match	35.1% [811/2310]	14.2% [550/3864]	< .001	49.9%
After Match	26.1% [352/1351]	26.5% [358/1351]	0.79	-1.0%

15

Instructor: **Thomas E. Love, PhD**
thomaslove@case.edu

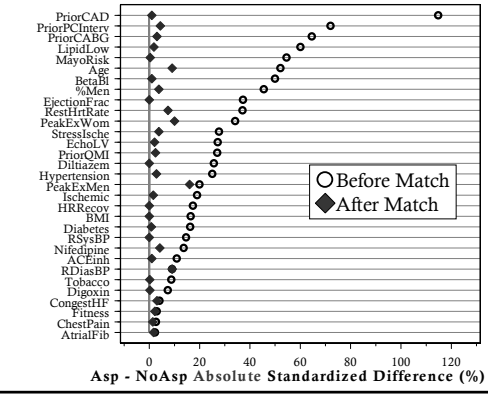
Part Three: Doing Propensity Analysis: Basics and Matching

Covariate Balance for Aspirin Study



16

Absolute Standardized Differences



17

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)

18

Instructor: **Thomas E. Love, PhD**
thomaslove@case.edu

Part Three: Doing Propensity Analysis: Basics and Matching

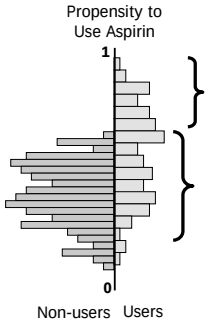
Which Aspirin Users Get Matched?

- Generally, characteristics of unmatched aspirin users tend to indicate high propensity scores.
  - Overall, 37% of patients were taking aspirin.
  - The rate was much higher in some populations… 67% of Prior CAD patients were taking aspirin.
  - So prior CAD patients had higher propensity scores for aspirin use.
  - Of the unmatched aspirin users, 99.8% (957/959) had prior coronary artery disease.
  - So it's likely that the unmatched users tended towards larger propensity scores than the matched users.

19

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

20

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.

21

Instructor: **Thomas E. Love, PhD**
thomaslove@case.edu

Part Three: Doing Propensity Analysis: Basics and Matching

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).

22

Aspirin Conclusions / Caveats

- Patients included in this study may be a more representative sample of "real world" patients than an RCT would provide.
- PS matching is still not randomization: can only account for the factors measured, and only as well as the instruments can measure them.
- No information on aspirin dose, aspirin allergy, or duration of treatment, or on medication adjustments.

23

What Should You Do About Residual Covariate Imbalance?

- Suppose we had a covariate that appeared seriously imbalanced after matching by propensity score.
- Consider whether a regression adjustment for that covariate should be made after matching.
- Consider whether an additional or alternative measurement of the concept described by the covariate should be included 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 PS calipers.

24

Instructor: **Thomas E. Love, PhD**
thomaslove@case.edu

Part Four: Using Propensity Scores

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

Part Four
Using Propensity Scores: Stratification, Adjustment, Sensitivity and Strategies

Thomas E. Love, Ph. D.
thomaslove@case.edu
Case Western Reserve University
MetroHealth Medical Center

October 17, 2004

1

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 (includes weighting)**
  - **Right Heart Catheterization** in the **Seriously Ill**
- Direct Adjustment with Propensity Scores
  - Prostate Cancer surgery vs. Radiotherapy

2

Using Propensity Scores to Stratify (Subclassify) Subjects

- Can we break subjects out into a small number of strata (subgroups), so each strata is homogeneous?
  - Can we use the PS to define the strata?
  - How many strata do we need to use?
  - After stratification, are the covariates balanced?
  - Can we then combine the strata to estimate the treatment's effect on outcome?

Seminal paper: Rosenbaum and Rubin (1984)

3

Instructor: **Thomas E. Love, PhD**
thomaslove@case.edu

Part Four: Using Propensity Scores

How should we stratify on many covariates simultaneously?

- 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.

4

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)

5

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.)

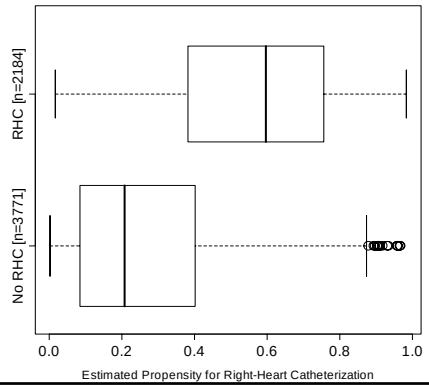
Panel (7 specialists in clinical care) specified important variables related to the decision to use or not use a RHC.

6

Instructor: **Thomas E. Love, PhD**
thomaslove@case.edu

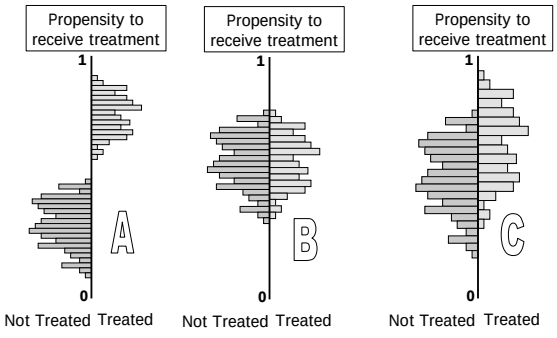
Part Four: Using Propensity Scores

Boxplot of Propensity Score Overlap



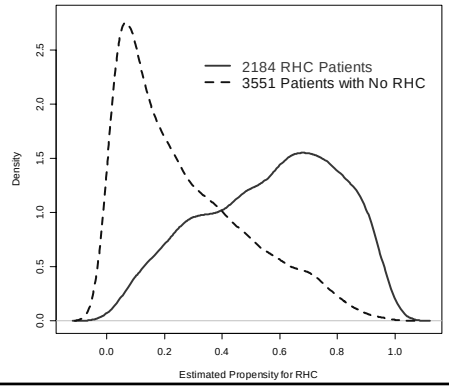
7

How Much Overlap In The Propensity Scores Do We Want?



8

Do the Propensity Scores Overlap?

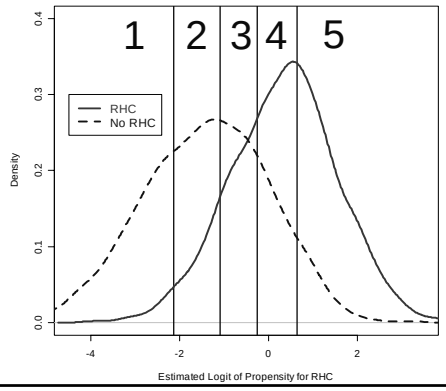


9

Instructor: **Thomas E. Love, PhD**
thomaslove@case.edu

Part Four: Using Propensity Scores

Subclassification by PS Quintiles



10

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%)

11

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

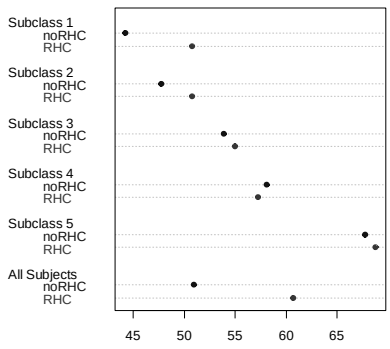
Mean DBP	**Overall**	**S1**	**S2**	**S3**	**S4**	**S5**
RHC	68	94	86	78	70	57
noRHC	85	100	83	77	72	59
Stzd. Diff.	**-45.5**	-13.4	-5.4	3.1	-6.5	-6.2
***p*-value**	**< .00001**	0.33	0.47	0.62	0.28	0.37

12

Instructor: **Thomas E. Love, PhD**
thomaslove@case.edu

Part Four: Using Propensity Scores

APACHE III Acute Physiology Score Across Subclasses



13

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

14

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)

15

Instructor: **Thomas E. Love, PhD**
thomaslove@case.edu

Part Four: Using Propensity Scores

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) 16

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) 17

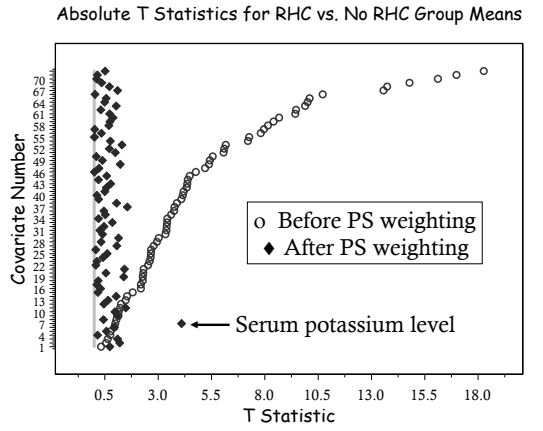
RHC Covariate Balance Before/After Weighting on the Propensity Score

	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) 18

Instructor: **Thomas E. Love, PhD**
thomaslove@case.edu

Part Four: Using Propensity Scores



19

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 (2006) provide nice applications.

20

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 (includes weighting)
  - Right Heart Catheterization in the Seriously Ill
- **Direct Adjustment with Propensity Scores**
  - **Prostate Cancer surgery vs. Radiotherapy**

21

Instructor: **Thomas E. Love, PhD**
thomaslove@case.edu

Part Four: Using Propensity Scores

Propensity Scores and Multivariate Adjustment

y = exposure + propensity + susceptibility + noise

- In modeling outcomes, we sometimes use treatment indicator and the propensity score itself as predictors, along with other important covariates and risk adjusters.
- When should we do this, rather than matching or stratifying?

22

How Is Multivariate Adjustment using the PS Done?

- 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.
- This is 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.

23

Example D: 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, imputed missing data.
- Split into PS quintiles to check balance – no sig diffs. in covariate means.
- Reasonable overlap of PS between groups.

Potosky et al. (2000)

24

Instructor: **Thomas E. Love, PhD**
thomaslove@case.edu

Part Four: Using Propensity Scores

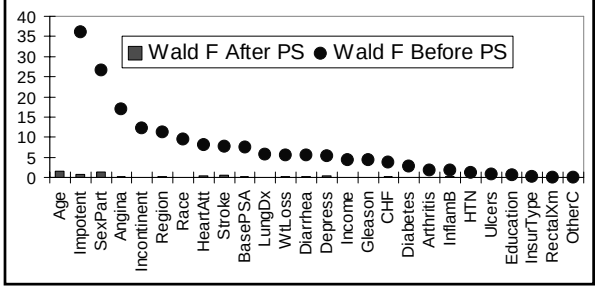
Checking for Covariate Balance after PS Adjustment

Variable	(Rx) RP %	(Ctrl) RT %	Unadjusted Wald F (p)	Wald F (p) **adj. for PS**
Incontinent	3	8	12.2 (<.001)	0.09 (.76)
Impotent	21	38	36.1 (<.001)	0.85 (.36)
CHF	5	8	3.8 (.05)	0.20 (.66)
Lung Dx	7	12	5.7 (.02)	0.02 (.89)
Hypertens.	41	45	1.2 (.28)	0 (.98)
Weight Loss	7	11	5.6 (.02)	0.20 (.66)
Angina	9	18	17.0 (<.001)	0.25 (.62)

Adapted from Table 1 in Potosky et al. (2000) p. 1585

25

Are the Covariates Balanced? Before/After PS Stratification



Stem and Leaf Plots of P Values Before/After PS Adjustment

Before		After	
.0	00000011112222445	.0	
.1	09	.1	
.2	08	.2	13
.3	9	.3	6
.4	4	.4	7
.5		.5	01
.6	5	.6	2566669
.7		.7	679
.8	69	.8	01244599
.9		.9	58

- Covariate balance much better after PS adjustment
- No significant differences remain

27

Instructor: **Thomas E. Love, PhD**
thomaslove@case.edu

Part Four: Using Propensity Scores

Multivariate Adjustment in Prostate Cancer Study

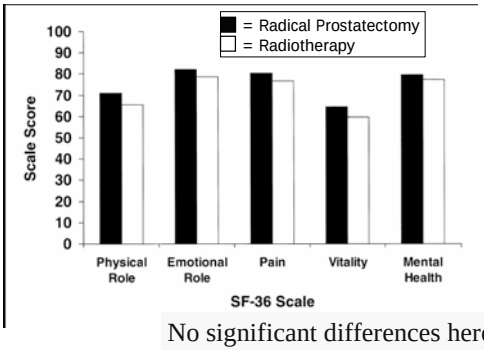
- Multivariate Logistic or Least Squares Regression Outcome Models…
- For instance, if p = Pr(Any Sexual Activity)…

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 \text{Treatment} + \beta_2 \text{PS} + \beta_{3\ldots} \text{Covs}$$

Covariates in model include age at diagnosis, baseline function, race/ethnicity, comorbidity, and educational attainment

28

Comparing Outcomes Directly After PS and Covariate Adjustment



No significant differences here.

29

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.

30

Instructor: **Thomas E. Love, PhD**
thomaslove@case.edu

Part Four: Using Propensity Scores

Sensitivity Analysis: Dealing with Hidden Bias

"How much hidden bias would have to be present to alter the study's conclusions?"

31

What To Do About 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 score model.
  - Either we have missed a domain of the problem
  - Or our measure of some domain is so weak that we've missed something important completely.

32

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 PS.
- If our study was randomized, or somehow free of hidden bias, we would have strong evidence of a treatment effect.
- To explain away the observed effect, an unobserved covariate would need to increase the odds of exposure to treatment and the odds of outcome by at least a factor of ____.

33

Instructor: **Thomas E. Love, PhD**
thomaslove@case.edu

Part Four: Using Propensity Scores

Surgery vs. Medicine for Coronary Artery Disease

- Coronary bypass surgery or medical/drug therapy for coronary artery disease?
  - 1515 subjects – 590 (39%) were surgical patients, the remaining 925 were medical patients.
  - PS included 74 observed covariates describing hemodynamic, angiographic, lab and exercise test results, as well as patient histories and demographics.
  - Stratified into quintiles based on PS, then combined estimates of Pr(sustained improvement at 6 mos.)

Rosenbaum PR & Rubin DB (1983, 1984) 34

Sensitivity Analysis for CAD Surgery vs. Medicine Study

Substantial Improvement at 6 mos.	Prob (SE)
Medical	.359 (.042)
Surgical	.669 (.059)

- Conclusion: Pr(improved|surgery) far exceeds Pr(improved|medicine).
- A hypothetical unobserved binary covariate would have to more than triple the odds of surgery and more than triple the odds of improvement, before altering the conclusion.

Rosenbaum PR and Rubin DB (1983, 1984) 35

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 ($\alpha = .05$)	10	9 (90%)	45%
Not Significant	7	2 (29%)	–36%

Love, Cebul, Thomas and Dawson (2003), Connors et al. (1996) 36

Instructor: **Thomas E. Love, PhD**
thomaslove@case.edu

Part Four: Using Propensity Scores

The Role of Sensitivity Analysis

- Cameron & Pauling's (1976) study concluded vitamin C increased survival time of colon cancer patients.
- Result not sensitive to an unmeasured binary covariate which led to a 10-fold increase in odds of exposure to vitamin C and was a perfect predictor of outcome.
- Sensitivity analysis looks great, yet the findings were contradicted in a RCT done at the Mayo Clinic.
- Conclusion: Sensitivity analysis cannot indicate what biases are present, it can only indicate the magnitude needed to alter the conclusion.

Rosenbaum PR (2002)

37

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.

38

Strategies for Designing Observational Studies Well Especially In Light of the EMR

39

Instructor: **Thomas E. Love, PhD**
thomaslove@case.edu

Part Four: Using Propensity Scores

Observational Studies Cry Out For Design

- These are reasonable methods with large samples, especially if we have a good selection model using multiple covariates.
- We're often working with secondary data, where design issues weren't considered…
- Options narrow as an investigation proceeds. What is easy early on may become difficult or impossible later.

40

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. EMR helps!

41

Screening for Covariates in the Propensity Score Model

- Most common method: **TERRIBLE IDEA!**
  - Compare treated to non-treated on many covariates.
  - Adjust only for significant differences.
- No reason that absence of significance implies imbalance is small enough to be ignored.
- Doesn't consider covariate-to-outcome relationship.
- This process considers covariates one at a time, while the PS adjustments will control the covariates simultaneously.

42

Instructor: **Thomas E. Love, PhD**
thomaslove@case.edu

Part Four: Using Propensity Scores

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.

43

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)

44

Review of 47 Published Articles

- **Variable selection in the PS model (6 did it right)**
  - 24 provided no information about variable selection
  - 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
  - 6 studies used non-parsimonious logistic regressions
- **Interaction Inclusion Criteria (1 did it right)**
  - 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)

45

Instructor: **Thomas E. Love, PhD**
thomaslove@case.edu

Part Four: Using Propensity Scores

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

46

Rubin's "Rules" for Checking Overlap for Multivariate Adjustment using PS

When stakes are high, consider propensity score overlap / balance in three ways to justify the use of direct PS adjustment:

1. Standardized Difference of Mean logit(PS) [Rx vs. Control] should be near 0, certainly < 50%.
2. Treated Var(PS) / Control Var (PS) should be close to 1. Certainly 1/2 or 2 is too extreme.
3. Regress each covariate (separately) on logit PS, look at variances of residuals as in part 2.

Rubin (2001)

47

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.

48

Instructor: **Thomas E. Love, PhD**
thomaslove@case.edu

Part Four: Using Propensity Scores

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?

49

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…

50

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

51

Partially Annotated Bibliography for SMDM Short Course #9:

REDUCING BIAS IN OBSERVATIONAL STUDIES: PROPENSITY METHODS

REFERENCES INCLUDED IN THE COURSE MATERIALS

1. Ahmed, A, Husain A, Love TE et al. (2006). "Heart failure, chronic diuretic use, and increase in mortality and hospitalization: An observational study using propensity scores." *European Heart Journal 27:* 1431-1439. [Uses propensity score matching combined with survival analyses to assess the impact of diuretics in HF patients. Incorporates formal sensitivity analysis and careful assessment of covariate balance, including the standardized differences plot.]
2. Austin, P. C. and M. M. Mamdani (2006). "A comparison of propensity score methods: A case-study estimating the effectiveness of post-AMI statin use." *Statistics in Medicine 25:* 2084-2106. [Examines multiple approaches to propensity score modeling applied to the same data, including matching, stratification, covariate adjustment and weighting. **A great "second paper" to read**. Also looks at several ways of assessing residual confounding after stratification or matching.]
3. 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. [Detailed PS matching and multivariate adjustment to account for selection bias, including sensitivity analysis - state of the art in 1996.]
4. 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 PS, with useful examples. **Excellent "first paper" to read.**]
5. 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. There is also an 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]
6. 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]
7. Rosenbaum, P. R. (1991). "Discussing hidden bias in observational studies." *Ann Intern Med 115*: 901-5. [Some of the fundamentals of sensitivity analysis]
8. 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. [Seminal paper on propensity matching]
9. 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 key propensity score issues.]
10. 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. [Detailed (yet readable) description of PS matching, subclassification, weighting, under high stakes. **Excellent "third paper" to read**.]

OTHER REFERENCES OF INTEREST

11. 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.]
12. 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.]
13. 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.]
14. 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.]
15. 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.]
16. 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.]
17. 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 PS (defined within) permits detection of and correction for selection bias.]
18. 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.]
19. 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 the use of PS to deal with imbalances observed in RCTs which need to be summarized with risk differences.]
20. 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.]

21. 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).]
22. 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.]
23. 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.]
24. 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.]
25. 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.]
26. 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.]
27. 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.]
28. 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.]
29. 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.]
30. 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.

31. 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.]
32. 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.]
33. 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. [What to do when few controls are comparable to treated subjects. Uses data from the National Supported Work experiment and PS matching, comparing treatment effect estimates obtained from matching on the PS to benchmark results obtained from an experiment.]
34. 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.]
35. 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: 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, outcomes post-treatment.]
36. 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.]
37. 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, PS matching as a weighting for dealing with more than two exposure groups.]
38. 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.]
39. 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.]
40. 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.]
41. Hahn, J. (1998). "On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects." *Econometrica 66*(2): 315-331.
42. 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.
43. Harraf, F., A. K. Sharma, et al. (2002). "A multicentre observational study of presentation and early assessment of acute stroke." *BMJ 325*: 17-21.
44. 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.

45. 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.
46. 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.
47. 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.]
48. 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.]
49. 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.
50. Hirano, K., G. W. Imbens, et al. (2003). "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score." *Econometrica 71*(4): 1161-1189.
51. 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.
52. Holden, C. (2005). "Sociology. Controversial study suggests seeing gun violence promotes it." *Science 308*(5726): 1239-40. [Comment on Bingenheimer et al.]
53. 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.
54. Hullsiek, K. H. and T. A. Louis (2002). "Propensity score modeling strategies for the causal analysis of observational data." *Biostatistics 2*(4): 179-193.
55. 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.
56. Ichimura, H. and C. Taber (2001). "Propensity-Score Matching with Instrumental Variables." *The American Economic Review 91*(2): 119-124.
57. 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.
58. Imbens, G. W. (2000). "The role of the propensity score in estimating dose-response functions." *Biometrika 87*(3): 706-710.
59. 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"]
60. 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.
61. 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.

62. 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.]
63. 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.
64. Kennedy, J., H. Quan, et al. (2005). "Statins Are Associated With Better Outcomes After Endarterectomy in Symptomatic Patients." *Stroke 36*: 2072-2076.
65. 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.
66. 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.
67. 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.
68. 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.
69. 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.
70. 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.
71. 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.]
72. 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.
73. 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.
74. 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.]
75. 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.
76. Lieberman, E., J. M. Lang, et al. (1996). "Association of epidural analgesia with cesarean delivery in nulliparas." *Obstet Gynecol 88*(6): 993-1000.
77. 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.
78. 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"]
79. 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.

80. 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, tests for differences in costs, visits, etc.]
81. 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.
82. Lu, B. (2005). "Propensity Score Matching with Time-Dependent Covariates." *Biometrics 61*: 721-28.
83. 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.
84. 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.
85. 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.
86. 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.
87. 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.
88. 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.
89. 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. ["A propensity score technique was used to balance covariates associated with statin use between groups {ref to Stone et al (1995) *Med Care.*} The covariates used 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."]
90. 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.
91. 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.
92. 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.]
93. Nash, I. S., R. R. Corrato, et al. (1999). "Generalist Versus Specialist Care for Acute Myocardial Infarction." *Am J Cardiol 83*: 650-654.
94. 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.]
95. 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.

96. 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.
97. 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).]
98. 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.]
99. 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.]
100. 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.]
101. Reiter, J. (2000). "Using Statistics to Determine Causal Relationships." *The American Mathematical Monthly 107*(1): 24-32.
102. Rosenbaum, P. R. (1987). "Model-Based Direct Adjustment." *Journal of the American Statistical Association 82*(398): 387-394.
103. Rosenbaum, P. R. (1991). "Sensitivity Analysis for Matched Case-Control Studies." *Biometrics 47*(1): 87-100.
104. Rosenbaum, P. R. (1999). "Choice as an Alternative to Control in Observational Studies." *Statistical Science 14*(3): 259-278.
105. Rosenbaum, P. R. (1999). "Reduced Sensitivity to Hidden Bias at Upper Quantiles in Observational Studies with Dilated Treatment Effects." *Biometrics 55*(2): 560-564.
106. Rosenbaum, P. R. (2001). "Effects attributable to treatment: Inference in experiments and observational studies with a discrete pivot." *Biometrika 88*(1): 219-231.
107. **Rosenbaum, P. R. (2002). *Observational Studies*. New York, Springer. [The major recent work on the subject of observational studies.]**
108. Rosenbaum, P. R. (2005). "Heterogeneity and Causality: Unit Heterogeneity and Design Sensitivity in Observational Studies." *The American Statistician 59*(2): 147-152.
109. 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.
110. 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. [Seminal paper - propensity score subclassification in detail.]
111. 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.]
112. Rubin, D. B. (2004). "On principles for modeling propensity scores in medical research." *Pharmacoepidemiol Drug Saf 13*(12): 855-7.
113. 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.

114. Rubin, D. B. and N. Thomas (1996). "Matching Using Estimated Propensity Scores: Relating Theory to Practice." *Biometrics 52*(1): 249-264.
115. 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.
116. 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.
117. 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.
118. 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"]
119. 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.
120. 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.]
121. Smith, H. L. (1997). "Matching with Multiple Controls to Estimate Treatment Effects in Observational Studies." *Sociological Methodology 27*: 325-353.
122. Smith, J. A. and P. E. Todd (2005). "Does matching overcome LaLonde's critique of nonexperimental estimators?" *J of Econometrics 125*: 305-353.
123. 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.
124. 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.
125. 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.
126. 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.
127. 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.
128. 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.
129. 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.
130. Winkelmayer, W. C. and T. Kurth (2004). "Propensity scores: help or hype?" *Nephrol Dial Transplant 19*: 1671-1673.
131. 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.