

Copyright © 2004 Thomas E. Love, Ph.D.
Contact author [thomaslove@case.edu] for permission to copy or quote from this material.

Using Propensity Score Methods Effectively

**Fall Workshop
Cleveland Chapter of the
American Statistical Association**

**Monday, October 11, 2004
from
12:45 – 4:45 PM**

Thomas E. Love, Ph. D., Instructor

Director, Biostatistics and Evaluation Unit
Center for Health Care Research and Policy
Case Western Reserve University
MetroHealth Medical Center

Assistant Professor of Medicine
Case Western Reserve University School of Medicine

Assistant Professor of Operations
Weatherhead School of Management
Case Western Reserve University

Web: <http://www.chrp.org/love>

Email: thomaslove@case.edu

Using Propensity Score Methods Effectively

COURSE OUTLINE

Part One: Fundamentals

What are propensity scores, and why do I need to know about them?

RANDOMIZED EXPERIMENTS AND OBSERVATIONAL STUDIES
THE NEED TO DEAL WITH SELECTION BIAS
WHAT IS THE PROPENSITY SCORE?
POTENTIAL OUTCOMES AND THE RUBIN CAUSAL MODEL
DOES THIS HAVE ANYTHING TO DO WITH SALES AND MARKETING?
HOW DO WE BUILD A PROPENSITY SCORE MODEL?
“TEN COMMANDMENTS” OF PROPENSITY MODEL DEVELOPMENT

Part Two: Mechanics

Using propensity scores in causal analyses for observational studies

MULTIVARIATE MATCHING USING THE PROPENSITY SCORE
PROPENSITY SCORE SUBCLASSIFICATION (STRATIFICATION) & WEIGHTING
REGRESSION ADJUSTMENT USING THE PROPENSITY SCORE
HIDDEN BIAS AND THE IMPORTANCE OF SENSITIVITY ANALYSES

Part Three: Strategies

What are best practices? How do I use propensity ideas effectively?

ADVANTAGES, CAUTIONS AND LIMITATIONS OF PROPENSITY METHODS
ESTIMATING PROPENSITY MODELS WELL: SELECTING COVARIATES
SMART WAYS TO ASSESS COVARIATE BALANCE
WHEN SHOULD WE MATCH / STRATIFY / WEIGHT / ADJUST?
NEW TECHNOLOGY: FULL OPTIMAL PROPENSITY SCORE MATCHING
USE AND ABUSE OF PROPENSITY MODELS: WHAT'S OUT THERE?
PROPENSITY SCORES AND MISSING DATA
USING PROPENSITY SCORES TO HELP DESIGN OBSERVATIONAL STUDIES
WHAT ABOUT INSTRUMENTAL VARIABLES?
WHAT SHOULD ALWAYS BE DONE, AND OFTEN ISN'T?

Using Propensity Score Methods Effectively

Part One: Fundamentals

Thomas E. Love, Ph. D.
Fall Workshop
ASA Cleveland Chapter
October 11, 2004
thomaslove@case.edu

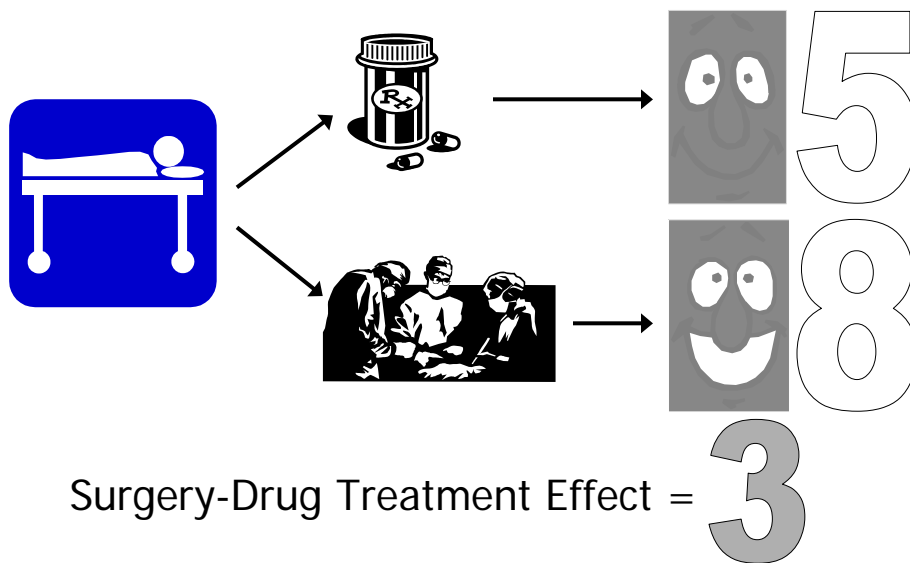
Acknowledgments

- Center for Health Care Research and Policy, including ...
- Randy Cebul, MD
- Neal Dawson, MD
- Charles Thomas
- Al Connors, MD
- Scott Husak
- 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
- Bo Lu, PhD
- Ben Hansen, PhD
- Michael Posner, MS
- Ali Ahmed, MD
- Society for Medical Decision Making
- ASA Cleveland Chapter

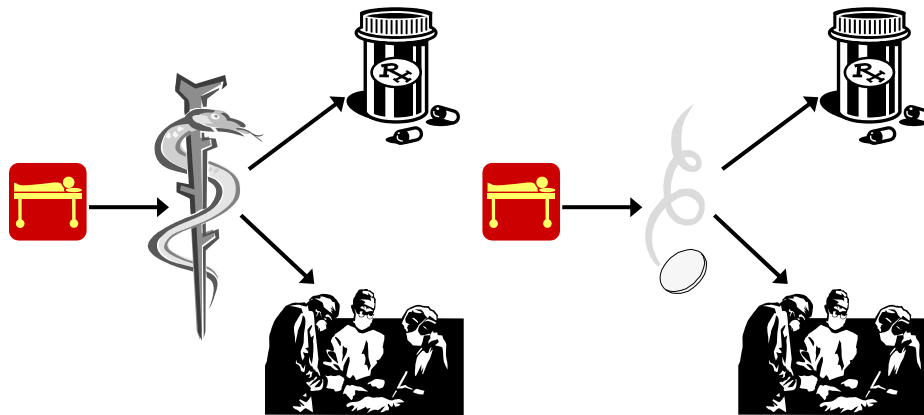
Outline of the Course

- Part One (Fundamentals)
 - What are propensity scores, and why do I need to know about them?
- Part Two (Mechanics)
 - How do I use a propensity score to build up a causal analysis of an observational study?
- Part Three (Strategies)
 - Best practices: How do I use propensity scores well? For what should I use them?

Looking for Causal Treatment Effects



Observational vs. Randomized Studies



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

Randomization ensures
that subjects receiving
different treatments
are comparable.

Observational Studies, Simply

- We have an outcome measured on two groups of subjects (treated and control).
- We want to make a fair comparison between the treated group and the control group in terms of the outcome.
- We can obtain covariates that describe the subjects before they received treatments, but we can't ensure that the groups will be comparable in terms of the covariates.

Importance of Randomization

- Randomization tends to produce relatively comparable or “balanced” treatment groups in large experiments.
 - The covariates aren’t used in assigning treatments in an experiment.
 - There’s no deliberate balancing of the covariates – it’s just a nice feature of randomization.
- We have some reason to hope and expect that other (unmeasured) variables will be balanced, as well.

A Randomized Clinical Trial of Coronary Surgery (VA Trial)

Covariate (measured pre-treatment)	Medical %	Surgery %
NY Heart Association Class II & III	94.2	95.4
History of myocardial infarction (MI)	59.3	64.0
Definite / possible MI (electrocardiogram)	36.1	40.5
Duration of chest pain > 25 mos.	50.0	51.8
History of hypertension	30.0	27.6
History of congestive heart failure	8.4	5.2
Cardiothoracic ratio > 0.49	10.4	12.2
Serum cholesterol > 249 mg / 100 ml *	31.6	20.6

* Difference between medical and surgical groups significant (p < .05)

Results of the VA Coronary Surgery Trial

- The VA study compared survival in the two groups three years after treatment.
 - Survival in the medical group was 87%
 - Survival in the surgical group was 88%
 - Both had a standard error of 2%, so the 1% difference in mortality was not significant
- Evidently, when comparable groups of patients received medical and surgical treatment at VA hospitals, outcomes were quite similar.

An Observational Study of Vitamin C and Treatment of Advanced Cancer

- Cameron and Pauling gave vitamin C to 100 patients believed terminally ill from cancer.
- 10 historical controls selected for each patient: same age, gender, cancer site and tumor type.
- Outcome: Time from “untreatability by standard therapies” to death.
- Result: Patients receiving vitamin C survived about 4 times longer than controls ($p < .0001$)

Cameron and Pauling (1976)

No Randomization Means ...?

- We want to compare groups who looked similar before they were exposed to treatments.
- We don't control the assignment of treatments, thus we can't use randomization to ensure comparability.
- How do we make fair comparisons?
 - Analytical adjustments required to account for baseline (covariate) differences in the groups.
 - A study is biased if the treatment groups differ in ways that matter for the outcomes under study.

Observational Study \neq RCT

- Mayo Clinic conducted an RCT on advanced cancer patients – no indication that vitamin C prolonged survival.
- What happened?
 - Controls in the OS evidently differed from treated patients in ways important to survival.
 - In the OS, controls were dead, while treated patients were alive at the start of the study – diagnosis (of terminal illness) may have been incorrect.
 - Time of “untreatability” is ambiguous.
- Must carefully design an OS, just like an experiment.
Moertel et al. (1985)

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

Non-randomly assigned exposures: What are their true effects?

- Treatments: There are reasons that provides select Rx over no Rx; Rx A over Rx B; dose 1 of Rx A vs. dose 2...
 - If these reasons affect outcomes, failing to account for them will bias treatment effect estimates
- A study exhibits selection bias if treated and control groups differ before treatment in ways that matter for outcome(s).

Using Matched Sets / Strata to Adjust for Overt (Selection) Bias

- Observe a set of P covariates, collected in \mathbf{X}
- Even if each covariate is binary, there are 2^P possible values of \mathbf{X} – many subjects are likely to have unique values of \mathbf{X} .
- Realistic goal: compare treated and control groups with similar distributions of \mathbf{X} , even if matched individuals have differing values of \mathbf{X}
- Key tool for doing this well: **propensity score**

The Propensity Score is a Conditional Probability

- Goal: compare the effect of two exposures: treatment and control, on some outcome.
- Allocation of Exposures was done on the basis of baseline (covariate) information (we don't have a precise formulation)
- We are essentially modeling this treatment allocation process.

Propensity Scores in the News

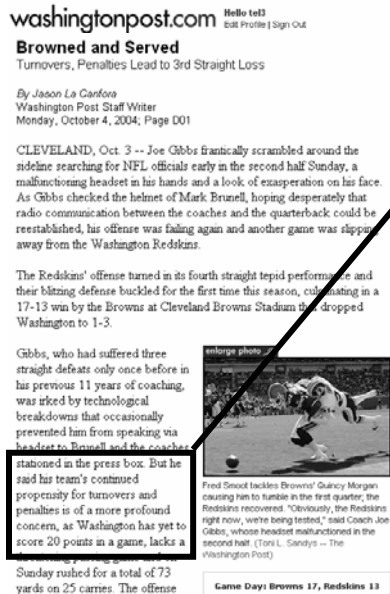


"Kerry pushes issue of stem cell research"
 Chicago Tribune article by Jill Zuckman
 Reprinted in the San Jose Mercury News
 October 4, 2004

With polls showing that the desire for stem cell research crosses party lines, Kerry has sought to tie Bush's decision to place limitations on the research to what he considers the president's broader **propensity** for making "wrong decisions."

In New Hampshire, people with one heartbreaking story after another talked about their desire for a cure as Kerry offered up empathy, a **score** of "God bless you's," hugs and pats on the back.

Propensity Scores in the News



But he said his team's continued **propensity** for turnovers and penalties is of a more profound concern, as Washington had yet to **score** 20 points in a game, ...

Jason La Canfora
 Washington Post
 October 4, 2004

PS Weighting and Online Polling



Press Release

Source: Harris Interactive

Doctors' Interpersonal Skills Valued More Than Their Training or Being Up-to-Date

Friday October 1, 4:18 pm ET

ROCHESTER, N.Y., Oct. 1 /PRNewswire/ -- U.S. adults believe it is extremely important for their doctors to have strong interpersonal skills such as being respectful (85%) and listening carefully to health care concerns and questions (84%), though they also value highly good medical judgment (80%), according to the results of a new Harris Interactive® poll of 2,267 U.S. adults conducted online between September 21 and 23, 2004 for the Wall Street Journal Online's Health Industry Edition.

In addition, adults feel it is important for a doctor to be easy to talk to (84%), to take their concerns seriously (83%) and truly care about them and their health (81%).

The biggest "gap" in what people want from doctors vs. what they actually get is related to how up-to-date their doctors are on the latest medical research and treatment, where 78 percent feel this knowledge is extremely important for their doctors to have, but only 54 percent actually described their doctors as being up-to-date.

With interpersonal skills being of so much value to patients, it is no surprise that some have changed doctors due to interpersonal failures. Fourteen percent changed because they didn't feel their doctors listened to them carefully, 12 percent felt as though their doctors didn't spend enough time with them, and 11 percent felt that they weren't treated with respect.

The survey also showed that a majority of patients prefer to communicate with their doctors by telephone (71%) when they have a non-urgent question rather than in person (21%) or via email (8%).

WSJ Online Poll conducted by Harris Interactive

This poll was conducted online in the U.S. between September 21-23, 2004 among a nationwide cross section of 2,267 adults. Figures for age, sex, race/ethnicity, education, income and region were weighted where necessary to align with population proportions. Propensity score weighting was also used to adjust for respondents' propensity to be online.

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.

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

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

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

Value-Added Education Assessment

- Causal effects are comparisons of potential outcomes measured at the same time point on a common set of units
 - Y = test scores at the end of fifth grade in a specific classroom of students
 - Estimate the effect of being in school A vs. school B for a particular student, Q
 - Compare the potential outcomes $Y_Q(A)$ vs. $Y_Q(B)$
- Fundamental Problem of Causal Inference: Only one of these outcomes is observed for student Q .
 - Student Q is in either school A or school B

Causal analysis is a missing data problem!

Rubin, Stuart, Zanutto (2004)

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
 - \mathbf{X} = vector of covariates
 - Z = treatment assignment variable (1 if treated, 0 if control)

Rubin (1997), Rosenbaum (2002)

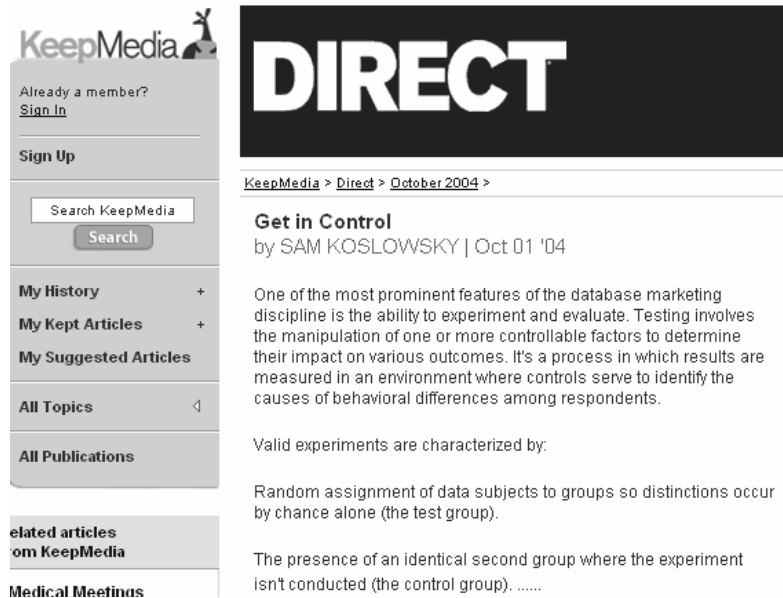
Making Inference About Treatment Effects Possible

- 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 \mathbf{X}
 - This is a “no hidden bias” assumption
 - We assume that \mathbf{X} contains all relevant information about treatment assignment

Estimating Causal Effects of Schools: Ideal Setting

- Randomized experiments = gold standard
 - What would we do if we could randomly assign?
 - Estimating causal effect of school A vs. school B where individual students are the units
 - Assign kids to the two schools at random, ensuring similar mixes of students in the classes?
 - Difference in observed outcomes between students in A and in B is unbiased estimate of true A vs. B difference
 - Some complications: (1) Interference between Units is possible (2) Versions of Treatments (3) Missing data

“Leveraging” Marketing Data



The screenshot shows a web page from KeepMedia. On the left is a navigation sidebar with links for 'Sign In', 'Sign Up', a search bar, and various article categories. The main content area features a large black header with the word 'DIRECT' in white. Below the header is the article title 'Get in Control' by SAM KOSLOWSKY, dated Oct 01 '04. The article text begins with 'One of the most prominent features of the database marketing discipline is the ability to experiment and evaluate. Testing involves the manipulation of one or more controllable factors to determine their impact on various outcomes. It's a process in which results are measured in an environment where controls serve to identify the causes of behavioral differences among respondents.' It then lists characteristics of valid experiments, including random assignment of subjects and the presence of a control group.

What Does This Have to do with Sales and Marketing?

- Adjusting for baseline characteristics in data to address causal questions.
- Suppose we want to look at a response (expression of interest in purchasing) if the customer received a promotion, compared to the response if the customer did not receive the promotion.

What Questions Do You Want To Answer About the Promotion?

- Targeting: To whom should we promote?
- Response: Can we estimate the impact of the promotion? Can we estimate ROI?
- Predictors: Can we “mine” for attributes that help predict response to the promotion?
- Sales Force Evaluation: Can we fairly estimate the average sales impact of our employees?
- Target List Evaluation: Are these likely responders?

www.anabus.com

The Database You Wish You Had

Customer	Sales if received promotion	Sales if received no promotion	Promotion's Impact
A	12	8	4
B	7	4	3
C	7	3	4
D	12	9	3

What is Reality?

Customer	Sales if received promotion	Sales if received no promotion	Promotion's Impact
A	12	?	?
B	7	?	?
C	?	3	?
D	?	9	?

USING PROPENSITY SCORE METHODS EFFECTIVELY

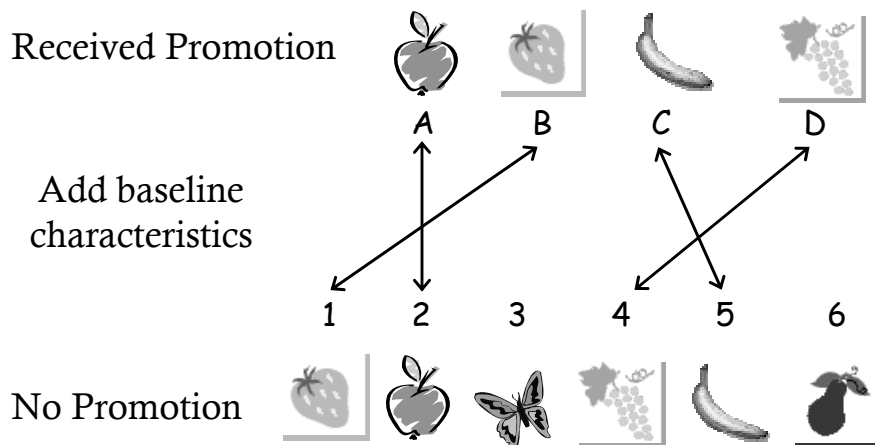
ASA CLEVELAND CHAPTER FALL WORKSHOP OCTOBER 11, 2004

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

So, the big question is...

- If we have someone who got the promotion in our database, how can we specify what that person's response would have been if they had not gotten the promotion?
- The promotion effect is the difference between the two responses (promotion – no promotion).
- Rubin Causal Model - "counterfactuals" and the potential outcomes framework

Causal Analysis "Comparing Apples to Apples"

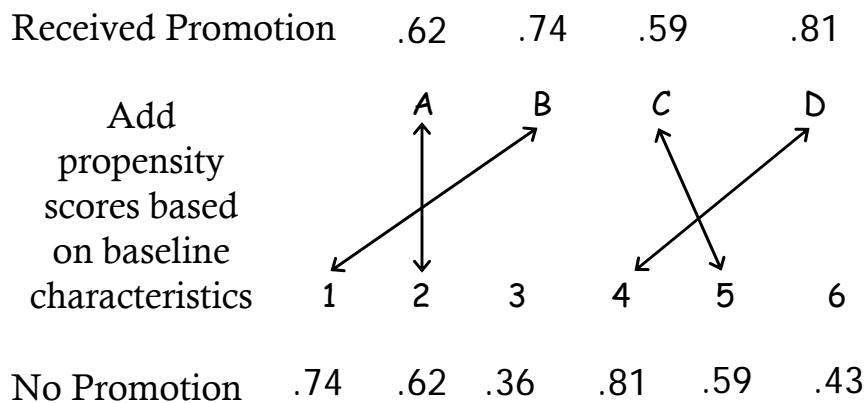


Propensity Scores - The Basic Notion

- We have a lot of information on each potential customer...
 - Demographics
 - Previous purchases
 - Geographic/socioeconomic information
- We can estimate the probability that a person with these background characteristics will receive the promotion

$$0 \leq \text{propensity score} \leq 1$$

Causal Analysis "Comparing Apples to Apples"



USING PROPENSITY SCORE METHODS EFFECTIVELY

ASA CLEVELAND CHAPTER FALL WORKSHOP OCTOBER 11, 2004

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

The New Database, Simply

Customer	Propensity to get promotion	Sales if received promotion	Sales if received no promotion
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...

Propensity Score Matching ⇒ New Database

Customer	Propensity to get promotion	Sales if got promotion	Sales if got no promotion	Impact
A	.80	12	9	3
B	.50	7	3	4
C	.50	7	3	4
D	.80	12	9	3

- Now we can estimate the impact of the promotion on each matched customer...

A Recurring Example: Security National Insurance

- We have a new life insurance product which we have marketed all over a Southern state.
- Two mailing strategies: “Fancy” and “Usual”
 - Fancy costs us three times as much as Usual
- We’d have liked to do a randomized experiment to compare Fancy to Usual.
- Instead, we let our managers use the database to decide who got Fancy and who got Usual.

Security National: Goal of Study

- The purpose of the marketing was to get people to call or write to us for more information on the product within some fixed period of time.
- The key outcome is this response rate.
- Key Question:

Pr(Response | Fancy) = Pr(Response | Usual)?

Security National: Details of Data

- We have 300 Fancy mailings and 1200 Usual mailings in our data base.
- Oversimplifying...
 - We have no missing data. If we have the value of a covariate for 1 person, we have it for all 1500.
 - There was no other information used to pick who got Fancy except that contained in our data base.
 - We have 6 baseline characteristics for each address.
 - The yes/no response was decided accurately after the same amount of time for all people.

Security National Data Base

- We have a subject number (1-300 for Fancy, and 301-1500 for Usual)
- We have six covariates, labeled Cov1 - Cov6.
- Cov1 – Cov4 are continuous variables
- Cov4 may have a quadratic impact, so that the addition of a squared term might help.
- Cov5 and Cov6 are binary variables
- Plausible interactions are [Cov5 with Cov1] and [Cov5 with Cov2].

So How Do We Build a Propensity Score Model?

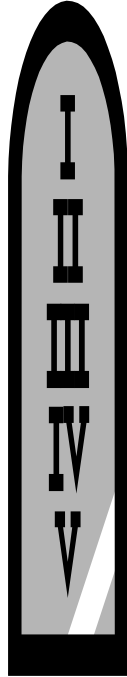
- Need to estimate $\Pr(\text{treatment} \mid \text{covariates})$
- Usual tool: Logistic regression model for the treatment allocation decision
 - We therefore want to consider including any variables that have a relationship to the treatment decision (i.e. precede it in time, and are relevant)
 - Clearly,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



USING PROPENSITY SCORE METHODS EFFECTIVELY

ASA CLEVELAND CHAPTER FALL WORKSHOP OCTOBER 11, 2004

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



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)

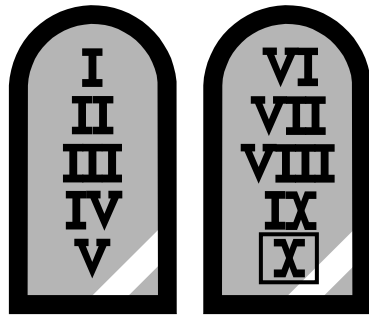


Thou Shalt Perform Bootstrap
Analyses To Assess Shrinkage

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

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 (in press [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. (in press [2004b])

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.

Using Propensity Score Methods Effectively

Part Two: Mechanics

Thomas E. Love, Ph. D.
Fall Workshop
ASA Cleveland Chapter
October 11, 2004
thomaslove@case.edu

For a nice introduction, try D'Agostino (1998)

Mechanics Part One: Multivariate Matching

- Close but inexact PS matching on a large pool of covariates removes most of the bias due to those covariates
- Assessing the Quality of the Matching
- Checking for Covariate Balance
- Examples
 - Aspirin Use and Mortality (CCF)
 - Security National Insurance

Seminal paper: Rosenbaum and Rubin (1983)

Multivariate Matching with the Propensity Score

- 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").

Aspirin Use and Mortality

- 6174 consecutive adults at CCF 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 et al. (2001)

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 propensity score model work?
- Are the covariates balanced?

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.

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.

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 \pm .03).
- Match patient within calipers of same gender.
- Repeat until no more matches are possible.

	Patient	Exposure	PS	Gender
{ .80 .79 .78 .77 .76 .75 .74 .73 .72	A	Treated	.76	Male
	B	Not Treated	.77	Female
	C	Not Treated	.74	Male
	D	Not Treated	.80	Male

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>

Excel Spreadsheet to Match Subjects by Propensity Score

	A	B	C	D	E	F	G	H	I	J	K
		Treated? (0=Control, 1=Treated)	Propensity (Probability)	Linear Propensity (Logit)	<i>Please do not write in purple section.</i>						
1					<div style="text-align: center; font-weight: bold; background-color: #333; color: white; padding: 2px;">Propensity Matching Program</div> <div style="font-size: small; margin-top: 5px;">Starting Information</div> <div style="display: flex; justify-content: space-between;"> <div style="width: 60%;"> <p>Starting Method</p> <p><input checked="" type="radio"/> Easiest to Match</p> <p><input type="radio"/> Hardest to Match</p> <p><input type="radio"/> ID Randomly Selected</p> <p><input type="radio"/> ID Selected by User</p> <p>Starting ID:</p> </div> <div style="width: 35%; text-align: center;"> <p style="border: 1px solid gray; padding: 2px; margin-bottom: 5px;">Run Matching Procedure</p> <p style="border: 1px solid gray; padding: 2px; margin-bottom: 5px;">Clear Input Worksheet</p> <p style="border: 1px solid gray; padding: 2px;">Restore Original Data & Delete Worksheets</p> </div> </div> <div style="margin-top: 10px; font-size: x-small;"> <p>Matching Standard</p> <p>Matching Standard</p> <p><input type="radio"/> Match Everyone</p> <p><input checked="" type="radio"/> Within Limit</p> <p>Select Percent of SE: <input style="width: 40px;" type="text" value="60"/></p> </div>						

Spreadsheet built by Love TE & Husak SS – not ready for primetime

Propensity Matcher Input Sheet

ID	Treated? (0=Control, 1=Treated)	Propensity (Probability)	Linear Propensity (Logit)
1	1	0.2	-1.386
2	1	0.3	-0.847
3	1	0.4	-0.405
4	1	0.6	0.405
5	1	0.7	0.847
6	0	0.24	-1.153
7	0	0.28	-0.944
8	0	0.31	-0.800
9	0	0.34	-0.663
10	0	0.39	-0.447
11	0	0.42	-0.323
12	0	0.45	-0.201

- Needed input file has ID, treatment status and Propensity Score
- The spreadsheet calculates the linear propensity scores automatically.

Propensity Matcher Control Panel

Propensity Matching Program

Starting Information

Starting Method

Easiest to Match

Hardest to Match

ID Randomly Selected

ID Selected by User

Starting ID:

Matching Standard

Matching Standard

Match Everyone

Within Limit

Select Percent of SE: ▲ ▼

Run Matching Procedure

Clear Input Worksheet

Restore Original Data & Delete Worksheets

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

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

- Standardized Differences are appropriate summaries of Covariate Balance for both Continuous and Categorical Variables

$$d = \frac{100(\bar{x}_{Treatment} - \bar{x}_{Control})}{\sqrt{\frac{s_{Treatment}^2 + s_{Control}^2}{2}}} \text{ for continuous variables}$$

$$d = \frac{100(p_{Treatment} - p_{Control})}{\sqrt{\frac{p_T(1-p_T) + p_C(1-p_C)}{2}}} \text{ for binary variables}$$

|Standardized Differences| > 10% Indicate Serious Imbalance

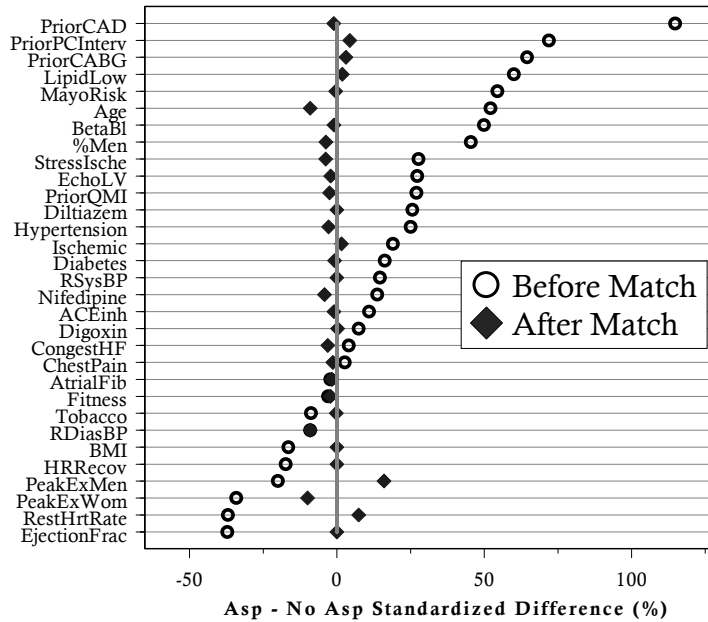
Before Match:

- 811/2310 (35.1%) Aspirin users used β -blockers
- 550/3864 (14.2%) non-Aspirin users used β -blockers
- Standardized Difference is 49.9%
- P value for difference is < .001

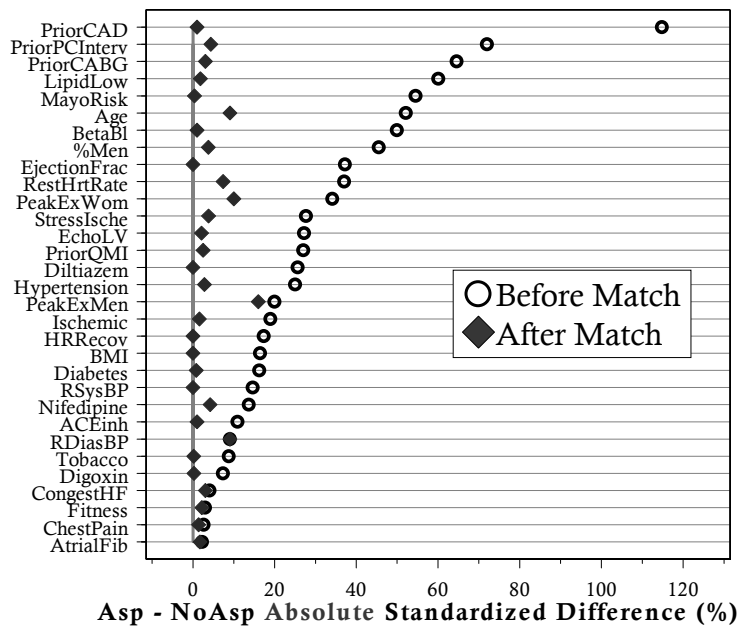
After Match:

- 352/1351 (26.1%) Aspirin users used β -blockers
- 358/1351 (26.5%) non-Aspirin users used β -blockers
- Standardized Difference is –1.0%
- P value for difference is .79

Covariate Balance for Aspirin Study



Absolute Standardized Differences



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.

Incomplete vs. Inexact Matching

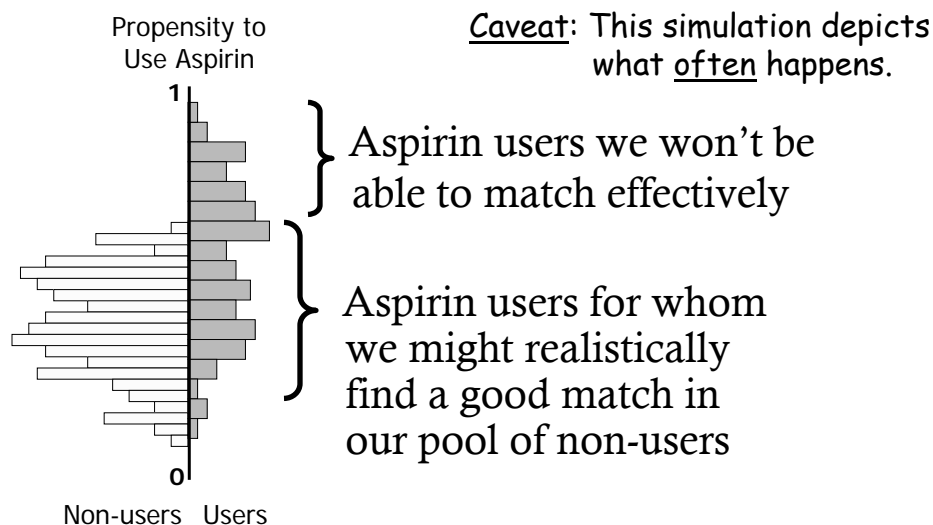
- Trade-off between
 - Failing to match all treated subjects (incomplete)
 - Matching dissimilar subjects (inexact matching)
- Severe bias due to incomplete matching – it's usually better to match all treated subjects, then follow with analytical adjustments for residual imbalances in the covariates.
- In practice, concern has been inexactness.
- Certainly worthwhile to define the comparison group and carefully explore why subjects match.

For more, see Rosenbaum (1991, 2002)

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 pts had higher propensity for aspirin.
 - 99.8% of unmatched aspirin users had prior CAD.
- Likely that unmatched users tended towards larger propensity scores than matched users.

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



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%, $p = .002$).

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.

Back to Our Recurring Example: Security National Insurance

- Two mailing strategies: “Fancy” and “Usual”
 - Fancy costs us three times as much as Usual
- Response rate (yes/no) is our key outcome of interest – key comparison to be made is:
 $\Pr(\text{Response} \mid \text{Fancy}) = \Pr(\text{Response} \mid \text{Usual})?$
- Propensity Scores = Predicted $\Pr(\text{treatment})$ for each subject, i.e. fitted values from a logistic regression model

Logistic Regression Model

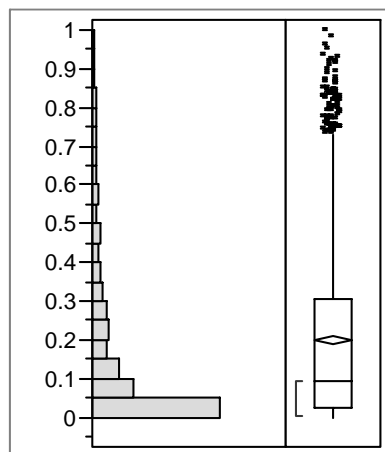
$$\text{logit}(PS) = -4.7 - .10Cov_1 + .19Cov_2 + .45Cov_3 - .12Cov_4 + .012(Cov_4)^2 - 1.84Cov_5 - 1.55Cov_6 + .014Cov_3Cov_4 - .092Cov_3Cov_5$$

Subject	Fancy?	Cov1	Cov2	Cov3	Cov4	Cov5	Cov6
1	1	9.5	18.1	8	5.4	0	0

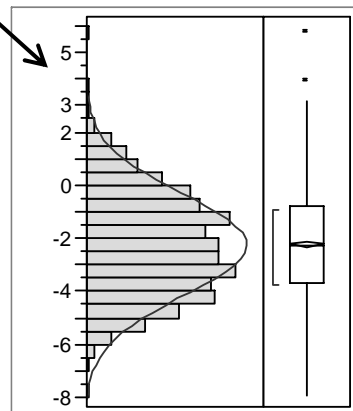
- For subject 1, $\text{logit}(PS) = 1.117$
- $\exp(1.117) = 3.056$, so $PS = 3.056 / 4.056$
- Subject 1 propensity for fitted = 0.75

Propensity Score Distribution

Max	75th	Med	25th	Min	Mean	SD	95% CI
.997	.306	.092	.024	.0004	.200	.235	(.188, .212)



$$\text{logit}(PS) = \ln(PS/[1-PS])$$



USING PROPENSITY SCORE METHODS EFFECTIVELY

ASA CLEVELAND CHAPTER FALL WORKSHOP OCTOBER 11, 2004

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

Covariate Balance (Before Matching)

	Usual Mean	Usual SD	Fancy Mean	Fancy SD	P value	Stdzd Diff (%)
Cov1	10.48	4.62	9.02	4.40	<.001	32
Cov2	12.98	4.49	13.71	5.60	.017	-14
Cov3	7.04	1.65	7.86	1.63	<.001	-50
Cov4	-0.13	4.86	0.51	6.22	.057	-11
(Cov4) ²	23.63	33.41	38.84	50.40	<.001	-36
Cov5*Cov1	6.10	6.16	1.09	3.24	<.001	102
Cov5*Cov2	7.81	7.44	1.55	4.69	<.001	101

Covariate Balance (Before Matching)

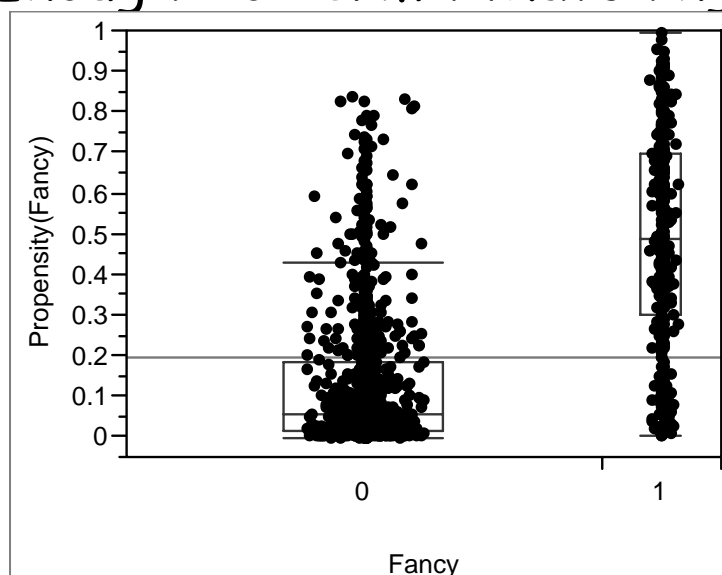
	Usual %	Fancy %	P value	Stdzd Diff (%)
Cov5*Cov1	59.3	12.7	< .001	111
Cov5*Cov2	73.6	48.3	< .001	54

	Usual Mean	Usual SD	Fancy Mean	Fancy SD	P value	Stdzd Diff (%)
logit(PS)	-2.73	1.62	-0.17	1.58	<.001	-160

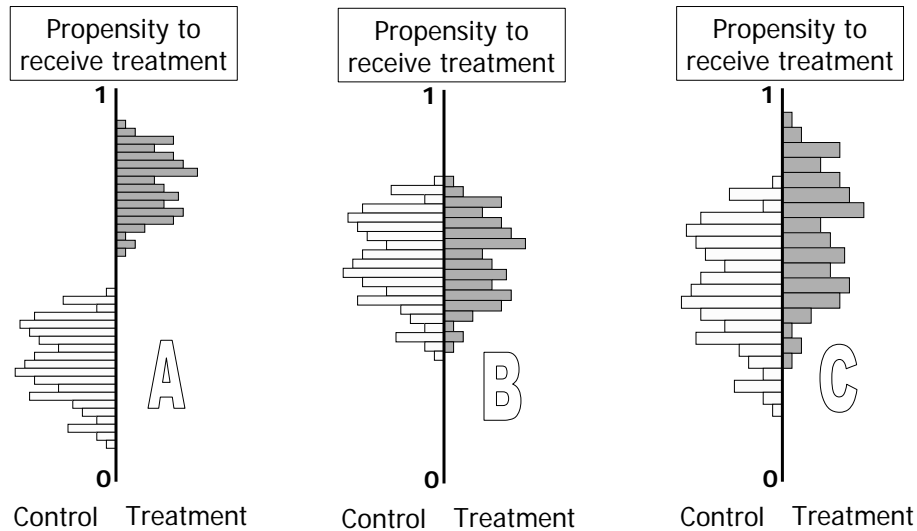
Security National: Running and Checking the Match

- Match each Fancy subject to a Usual subject with a similar propensity score.
- We will only use the PS to match, and stop matching when we no longer find a Usual subject with $\text{logit}(\text{PS})$ within .6 standard error of the $\text{logit}(\text{PS})$ of our Fancy subject.
- We will then assess the quality of the match in terms of the resulting covariate balance.

Do Propensity Scores Overlap Enough To Permit Matching?



How Much Propensity Score Overlap Do We Want?



Details of Propensity Matching

- 197 matched pairs were formed.
- For all pairs, difference in $\text{logit}(\text{PS}) < .03$

Quality	ID	PS	Logit	C ₁	C ₂	C ₃	C ₄	C ₅	C ₆
Worst	130	.527	.108	1.5	5.5	10	-8.1	1	0
Match	527	.520	.079	6.5	16.8	9	4.1	0	1
Median	161	.275	-.968	12.6	7.1	9	-6.2	0	1
Match	929	.274	-.972	9.4	19.5	6	7.2	0	1
Best	17	.130	-1.898	16.2	8.6	8	-4.6	0	1
Match	376	.130	-1.898	20.4	11.7	9	-1.4	0	1

USING PROPENSITY SCORE METHODS EFFECTIVELY

ASA CLEVELAND CHAPTER FALL WORKSHOP OCTOBER 11, 2004

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

Are The Covariates Balanced Now?

Variable	Standardized Difference Before Match	Stdzd. Difference After Match
Cov1	32.4	-1.9
Cov2	-14.4	-0.4
Cov3	-49.6	7.2
Cov4	-11.4	-0.5
Cov5	111.0	6.4
Cov6	53.7	0
(Cov4) ²	-35.6	-0.9
C5*C1	101.7	10.5
C5*C2	100.6	3.0

Security National Life Insurance: Producing the New Data Base

- We will use the matched Usual subject as a proxy for what would have happened had each Fancy subject received the Usual treatment instead.
- We will develop an appropriate contingency table and compute relevant test statistics to assess whether $\Pr(\text{Response} | \text{Fancy}) = \Pr(\text{Response} | \text{Usual})$.

USING PROPENSITY SCORE METHODS EFFECTIVELY

ASA CLEVELAND CHAPTER FALL WORKSHOP OCTOBER 11, 2004

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

Unadjusted Result

Group	Mailings	Responses	Rate
Fancy	300	120	40.0%
Usual	1200	245	20.4%
Total	1500	365	24.3%

- Odds Ratio = 2.60 95% CI (1.96, 3.43)
- “Fancy” recipient is 2.60 times as likely to respond to the mailing as a “Usual” recipient.
- Since $\text{Cost}(\text{Fancy}) = 3 \text{ Cost}(\text{Usual})$...

PS-Matched Sample Results

Group	Mailings	Responses	Rate
Fancy	197	78	39.6%
Usual	197	25	12.7%
Total	394	95	24.1%

- Odds Ratio = 4.51 95% CI (2.65, 7.81)
- Accounting for matching using a McNemar-style procedure, Odds Ratio = 4.76 (3.07, 7.65)
- “Fancy” recipient is more than 4.5 times as likely to respond to the mailing as a “Usual” recipient.
- CI for difference in response rates: (18.1%, 29.6%)

Mechanics Part Two: Subclassification / Stratification

- Propensity scores permit subclassification on multiple covariates simultaneously.
 - Permits the use of the whole sample of data (not just matched sets), without relying (as in regression adjustment) on a functional form
- Examples
 - Surgery vs. Medicine for Coronary Artery Disease
 - Security National Insurance
 - Weighting as a form of Stratification

Seminal paper: Rosenbaum and Rubin (1984)

Cochran's Subclassification Example: U.S. Male Death Rates per 1000 person-years

Smoking Group	Mean age in years	Unadjusted death rate	Adjusted for age (Two subclasses)
Non-smokers	57.0	13.5	13.5
Cigarettes only	53.2	13.5	16.4
Cigars, pipes	59.7	17.4	14.9

- Combine “low age” mortality rate in each smoking group with “high age” mortality rate in that group, weighting by population proportions of “low age” and “high age” U.S. males.

Cochran (1968, 1983), Rosenbaum and Rubin (1984)

How Many Strata Do We Need?

- Here, we've used only two strata – “low age” and “high age”.
- Cochran used 3 and 12 strata – similar results.
- Age is sort of continuous – we could have stratified much more finely – worth doing?
- Cochran presents a theoretical argument concluding that five strata, each containing 20% of the subjects, will remove about 90% of the bias in a single continuous covariate.

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?

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.

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.
 - 74 observed covariates describing hemodynamic, angiographic, lab and exercise test results, as well as patient histories and demographics.
 - Each of the 74 covariates was significantly different comparing surgical to medical patients.

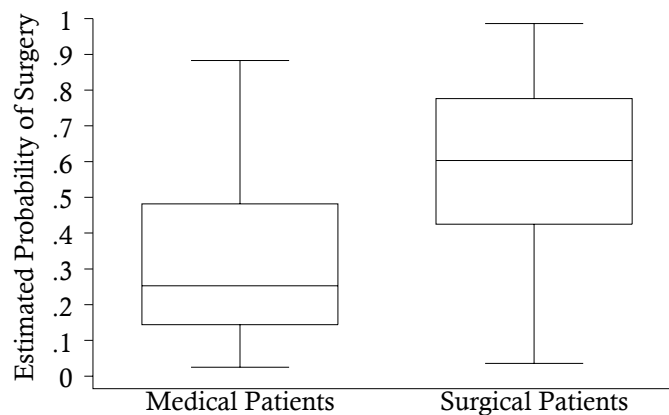
Rosenbaum and Rubin (1984)

Propensity Model for CAD Study

- Logistic regression used to predict treatment assignment for each of the 1515 subjects on the basis of...
 - The 74 covariates themselves
 - Interactions between some covariates
 - Quadratic terms for some covariates
 - Model selection process was sequential – described in the paper...

Rosenbaum and Rubin (1984)

Overlap of Treatment Groups



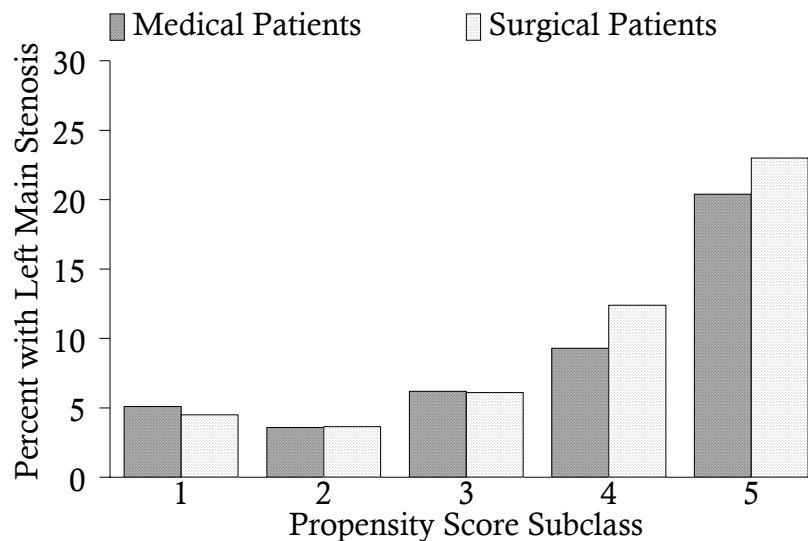
- For almost every surgical patient, there is a comparable medical patient in terms of having a similar estimated $\Pr(\text{surgery})$.

Propensity Score Subclassification

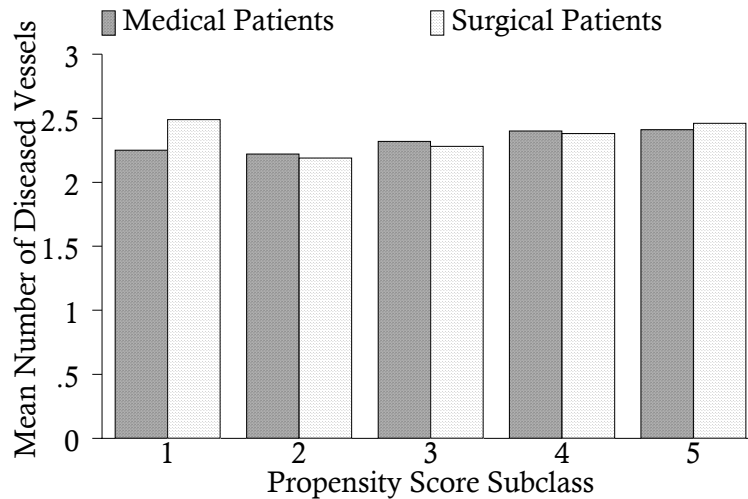
- 1515 patients were divided into five strata of 303 patients each, using estimated propensity scores.

PS Strata	Propensity Score \approx Prob(surgery covars.)	Actually got surgery	Actually got medical
5	Highest 303 scores	234 (77%)	69 (23%)
4	2 nd highest	164 (54%)	139 (46%)
3	Middle	98 (32%)	205 (68%)
2	2 nd lowest	68 (22%)	235 (78%)
1	Lowest 303 scores	26 (9%)	277 (91%)

Balance Within Subclasses: % with Left Main Stenosis



Balance within Subclasses: Number of Diseased Vessels



Subclass Specific Estimates: Outcomes: Survival at 6 mo & Uninterrupted Improvement at 6 mo

PS Subclass	Group	Patients	P(Survived)	P(Improved)
1	Med	277	.892	.351
	Surg	26	.846	.538
2	Med	235	.953	.402
	Surg	68	.926	.705
3	Med	205	.922	.351
	Surg	98	.898	.699
4	Med	139	.941	.303
	Surg	164	.933	.706
5	Med	69	.924	.390
	Surg	234	.914	.696

USING PROPENSITY SCORE METHODS EFFECTIVELY

ASA CLEVELAND CHAPTER FALL WORKSHOP OCTOBER 11, 2004

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

Directly Adjusted Probabilities of Survival and Uninterrupted Improvement (and Standard Errors)

	6 months Prob (SE)	3 years Prob (SE)
Survival		
Medical	.926 (.022)	.790 (.040)
Surgical	.903 (.039)	.846 (.049)
Improvement		
Medical	.359 (.042)	.126 (.036)
Surgical	.669 (.059)	.298 (.057)

Propensity Score Subclassification

- PS stratification allows assessment of whether sufficient overlap exists with respect to covariates among groups to allow the effect to be estimated.
- PS methods lead to more reliable estimates of association than multiple regression, especially if there is a substantial selection or other overt bias.
- Benefits of matching while using all the data.
- Can perform separate additional adjustments on each of the PS strata, if they are of interest

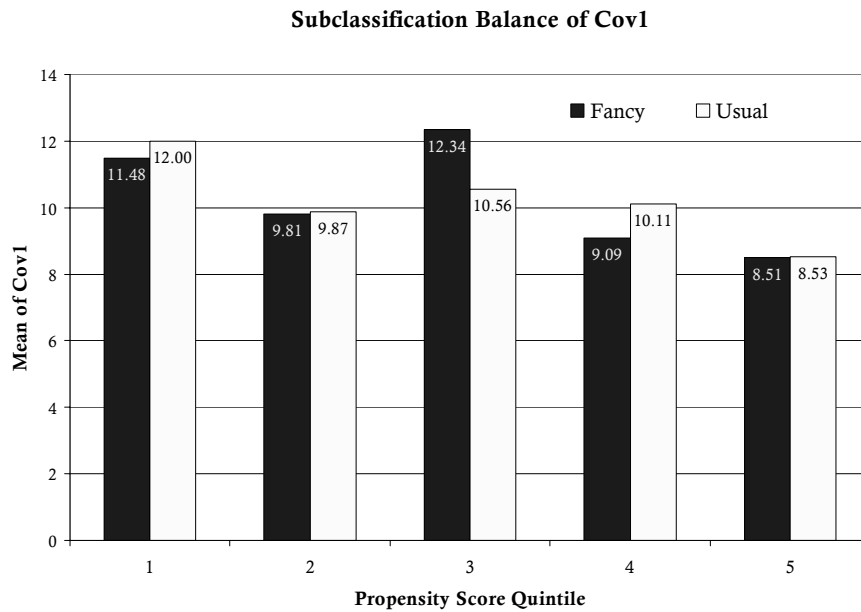
Security National Example: Using Stratification with the PS

- Instead of matching, we will subclassify the 1500 subjects into 5 groups of 300 each by their propensity scores.
- We will ensure that the covariate balance remains acceptable under such a scheme.
- We will calculate appropriate response rates within each subclassification, and then combine the results to assess whether $\Pr(\text{Response} | \text{Fancy}) = \Pr(\text{Response} | \text{Usual})$

What the Subclassification Looks Like

PS Subclass	Propensity Scores	Actually Got Fancy
5	.381 and above	199 of 300 (66%)
4	.1499 - .3804	60 of 300 (20%)
3	.05301 - .1496	19 of 300 (6%)
2	.01761 - .05299	12 of 300 (4%)
1	.01756 and below	10 of 300 (3%)

Is Cov1 Balanced by Subclassification?



Odds Ratios Within Each Subclassification

Subclass	5	4	3	2	1
Odds Ratio	5.09	2.78	2.80	3.65	1.95
“Fancy” Response	41%	35%	42%	50%	40%
“Usual” Response	12%	16%	21%	22%	26%
Overall Response	31%	20%	22%	23%	26%

Mantel-Haenszel Combined Odds Ratio
 Estimate = 3.51

M-H $p < .0001$, 95% CI (2.43, 5.08)

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). Also see Lunceford and Davidian (2004)

Weighting and Right Heart Catheterization

- 5735 patients in the study: 2184 treated (RHC) and 3551 controls (no RHC).
- Outcome: indicator of 30 day survival
- Treatment status: 1 if RHC applied within 24 hours of admission to the hospital
- PS model estimated by Hirano and Imbens using 57 of 72 available covariates (those with $t > 2.0$).
- Reweight each treated patient by $1/PS$, and each control patient by $1/(1-PS)$.
- PS weighting has attractive efficiency properties for estimating average treatment effects

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

USING PROPENSITY SCORE METHODS EFFECTIVELY

ASA CLEVELAND CHAPTER FALL WORKSHOP OCTOBER 11, 2004

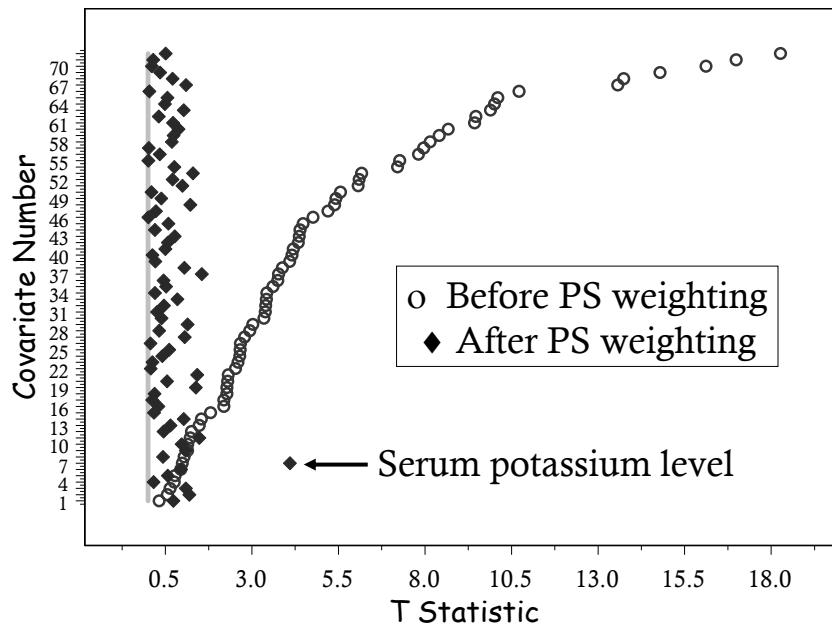
Thomas E. Love, Ph. D. thomaslove@case.edu www.chrp.org/love

RHC Covariate Balance Before/After Weighting on the Propensity Score

Variable	Before Weighting			After Weighting		
	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)

Absolute T Statistics for RHC vs. No RHC Group Means



Mechanics Part Three: Multivariate Adjustment using Propensity Scores

- We sometimes use treatment indicator and the propensity score itself as predictors, along with other important covariates and risk adjusters.
- How do we present the results effectively?
- Examples
 - Prostate Cancer Treatment Comparison
 - Security National Insurance

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 on outcomes.

Potosky et al (2000)

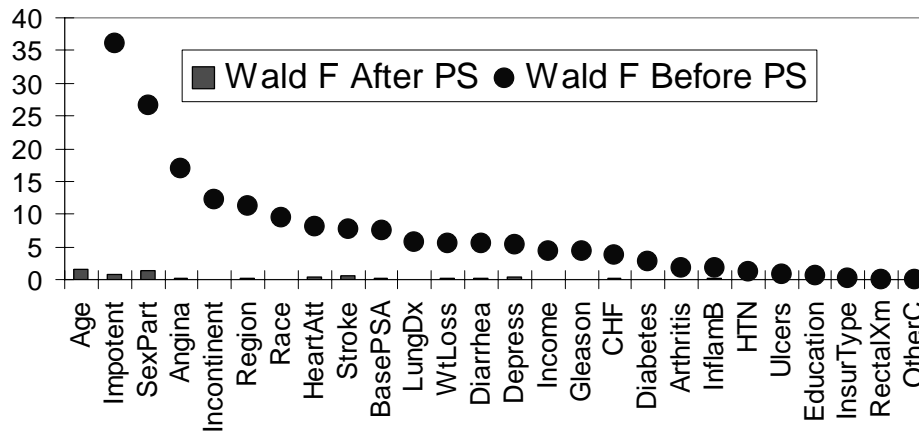
USING PROPENSITY SCORE METHODS EFFECTIVELY

ASA CLEVELAND CHAPTER FALL WORKSHOP OCTOBER 11, 2004

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

131 ●

Are the Covariates Balanced? Before/After PS Stratification



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.

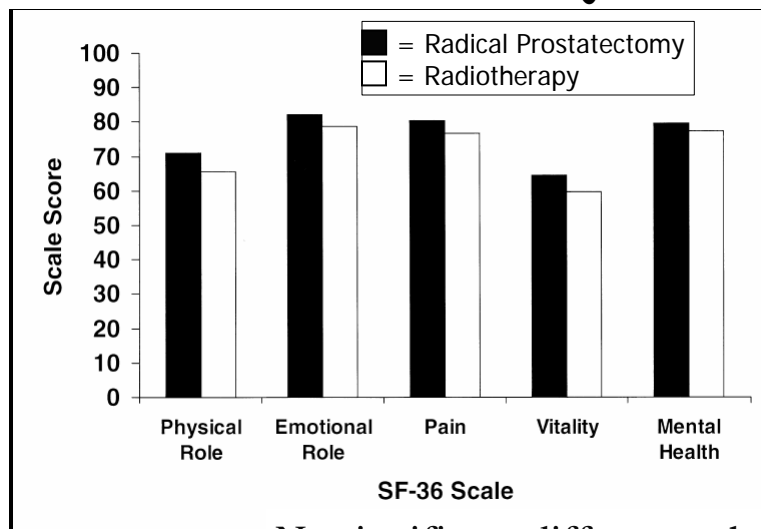
Multivariate Adjustment in Prostate Cancer Study

- Multivariate Logistic or Least Squares Regression Outcome Models...
- For instance, if $p = \text{Pr}(\text{Any Sexual Activity}) \dots$

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 \text{Treatment} + \beta_2 \text{PS} + \beta_3 \dots \text{Covs}$$

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

Comparing Outcomes Directly After PS and Covariate Adjustment



No significant differences here.

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.

Conclusions for Prostate Cancer Study

- Treatment choice, baseline function, and age are the main determinants of changes in disease-specific outcomes.
- RP has a greater effect on urinary incontinence and sexual function than RT.
- RT has a greater bowel function effect.
- How strong is the evidence 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 we could identify.
- Treatment effects on health outcomes were generally quite large, consistent with earlier studies, and clinically plausible.

Security National Example: Multivariate Adjustment with the PS

- Instead of matching or stratification, we will use the PS directly in a multivariate logistic regression model (along with an indicator of the treatment, and some of the key covariates) designed to predict response to the marketing.
- We will then assess whether $\Pr(\text{Response} | \text{Fancy}) = \Pr(\text{Response} | \text{Useful})$.

Model 1: Adjusting for the PS

$$\Pr(\text{Response}) = \frac{\exp(-0.647 + 0.618 \text{ Fancy} - 0.783 \text{ Propensity Score})}{1 + \exp(-0.647 + 0.618 \text{ Fancy} - 0.783 \text{ Propensity Score})}$$

- n = 1500 observations in a logistic regression model.
- Odds Ratio for Fancy = 3.44
- 95% CI for Odds Ratio is (2.41, 4.94)
- Area under ROC Curve = .624

More Multivariate Models

Model	2	3	4
# Observations	1500	1500	1500
Adjusting for	logit(PS)	PS, C ₃ , C ₅ , (C ₄) ²	C ₁ -C ₆ , (C ₄) ² , C ₅ C ₁ , C ₅ C ₂
Odds Ratio	3.49	3.47	3.56
ROC Area	.624	.622	.625

- All models show similar and highly significant (p < .001) “Fancy” effect.
- No model discriminates well.

Is Covariate Adjustment Not As Good As Matching/Stratification?

- If the unadjusted variance in the control group is much larger than the variance in the treated group, it is likely that covariate adjustment will not be as effective as matching or subclassification on the PS.
- Whether or not matched sampling is used, further analytical adjustments may be desirable to control residual bias and to increase efficiency.
- Rubin found that regression adjustment of matched-pair differences is a robust technique.

Rubin (1997)

Model 5: Using PS-Matched Data

$$\text{Pr}(\text{Response}) = \frac{\exp(-1.086 + 0.753 \text{ Fancy} - 0.245 \text{ Propensity Score})}{1 + \exp(-1.086 + 0.753 \text{ Fancy} - 0.245 \text{ Propensity Score})}$$

- Using only the 394 matched cases.
- Propensity Score no longer significant.
 - Dropping PS from model has no real impact on odds ratio for Fancy, but keeping it improves C.
- Odds Ratio for Fancy is 4.51
- 95% CI is (2.75, 7.62)
- ROC Area is .681 – best model so far, but of course we have a different sample here.

Hidden Bias and Sensitivity Analysis

What can we do about hidden bias?

How can we assess the potential impact of
things we didn't measure?

How much hidden bias would have to be
present in order to change my conclusions?

When is it sufficient to adjust for the observed covariates?

- Main Statistical Assumption: Strongly Ignorable Treatment Assignment
- After adjustment for observed covariates, we assume that the different treatment groups are comparable.
- No hidden bias.

The Potential Impact of Hidden Bias

- Overt Bias – resolved with adjustment techniques using measured covariates
- Hidden Bias – assess potential for unmeasured covariates to affect results with sensitivity analysis. These can:
 - Create a false exposure effect
 - Mask a true exposure effect
 - Cause mis-estimation of effect size / direction

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.

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.
- 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)

Sensitivity Analysis for CAD Surgery vs. Medicine

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

- Conclusion: $\Pr(\text{improved} \mid \text{surgery})$ far exceeds $\Pr(\text{improved} \mid \text{medicine})$.
- 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 and Rubin (1983)

What Can Sensitivity Analysis Do?

- With a suitable test, Cameron & Pauling's study of vitamin C and colon cancer is insensitive to extremely large biases (10-fold increases in odds of exposure to vitamin C).
- Yet the findings were contradicted in a RCT.
- Sensitivity analysis cannot indicate what biases are present, it can only indicate the magnitude needed to alter the conclusion.
- Conclusion: Selection bias was probably huge.

Cameron and Pauling (1976), Rosenbaum (2002)

Security National Example: Sensitivity to Hidden Bias

- We have essentially assumed away this problem in our setup, but there are several possible methods for assessing the potential impact of hidden bias on our conclusions.
- All of these include an assessment of the characteristics required of an unmeasured binary covariate to have it change our conclusions substantially.

Could Our Matched Samples Conclusion Be Due To A Hidden Bias?

- McNemar's test: $p < .0001$ – strong evidence that “Fancy” mailing causes more responses.
- To attribute the higher “Fancy” response rate to an unobserved binary covariate instead of to the difference between “Fancy” and “Usual”, the unobserved covariate would have to produce a 3.43-fold increase in the odds of getting the Fancy mailing, and the covariate would also need to be a near perfect predictor of response to the solicitation.

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)

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.

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.

Using Propensity Score Methods Effectively

Part Three: Strategies

Thomas E. Love, Ph. D.
Fall Workshop
ASA Cleveland Chapter
October 11, 2004
thomaslove@case.edu

What Propensity Scores Can and Cannot Do, in a Slide

- If we match 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 any baseline differences on selection to treatment or control.
- But if our propensity model misses an important reason why subjects are selected to treatment or control, we'll be in trouble.

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?

Rare Outcomes, Common Treatments

- Propensity score models are most likely to be helpful in settings where an outcome is relatively rare, but treatment allocation is reasonably varied across patients.
- Largest bang for buck: more spread in exposures than in outcomes
- Is it ever a BAD idea to use propensity scores? Can you get into trouble?

Also see Joffe and Rosenbaum (1999). If you must, see Cepeda (2003)

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...

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.

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)

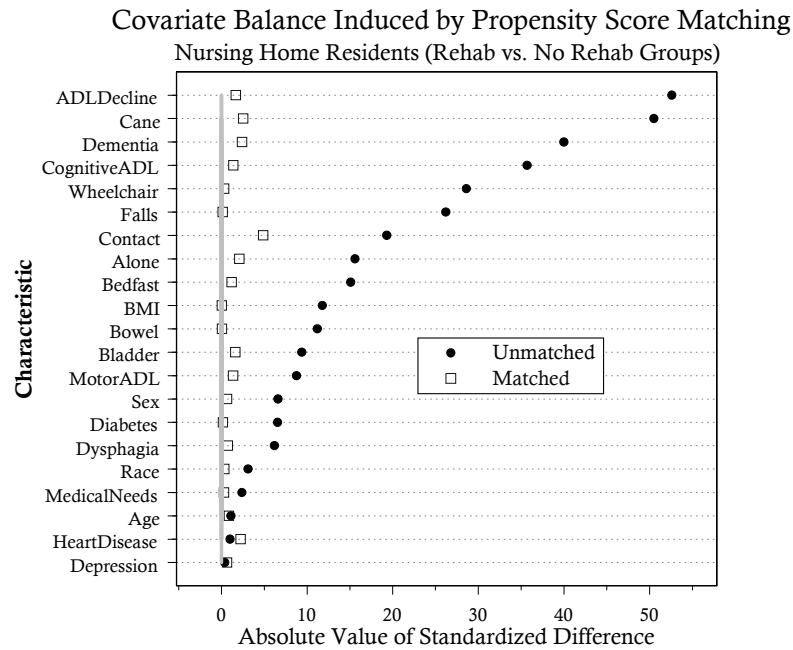
Variable	Rehab	No Rehab
Mean Age	78.1	78.0
% female	67	64
Mean BMI	24.2	23.6
Mean Cog ADL	79.6	71.0
% ath. ♥ disease	15	16
% dementia	20	38
% live alone	40	33
% with a cane	54	30
% recent ADL ↓↓	66	41

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

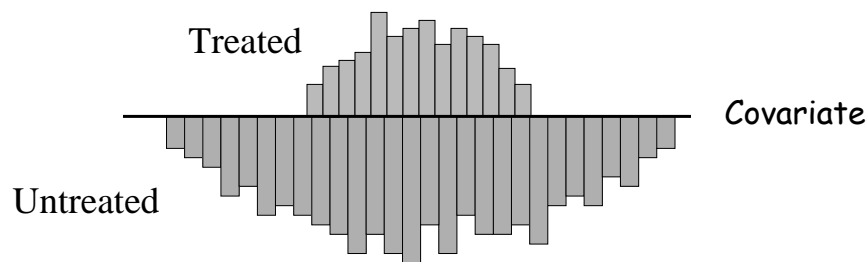
Choices To Be Made

- 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?



Is Covariate Adjustment Not As Good As Matching/Stratification?

- If the unadjusted variance in the non-treated group is much larger than the variance in the treated group, it is likely that covariate adjustment alone will not be as effective as matching or subclassification on the PS.



When Is Matching Your Best Choice?

- Certain covariates are more easily controlled through matching in the design than through other types of analytical adjustments.
- Typically these are covariates that classify subjects into many small categories.
- If matching isn't used, some categories may wind up with treated subjects and no controls, or vice versa.

Cost and Matching

- Cost is an important consideration.
- If some covariate information is readily available, but other data are difficult to obtain or expensive, matching becomes more attractive.
- If data come with negligible costs, matching during the design is less attractive.
- Why? Suppose some controls are so different (at baseline) from the treated subjects that they will be of little use. Matching may stop you from collecting data on such controls.

Rubin (2001), Rosenbaum (2002)

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?

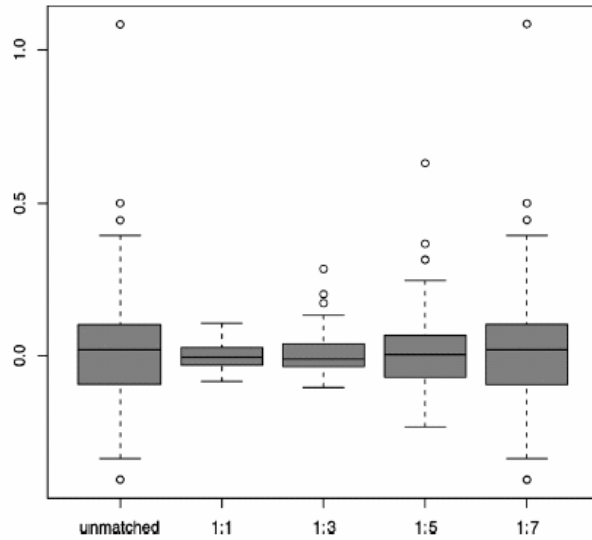


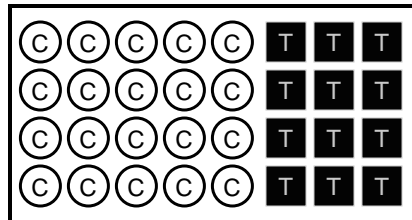
Figure 1. Covariate Imbalances in 1:k Matching. Each boxplot represents standardized biases in the 99 categories of the 27 categorical covariates along with standardized bias in the propensity score (which in each plot is the uppermost outlier). Strictly speaking, the matching represented at far right 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

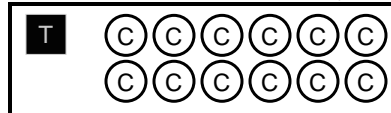
- In all of these cases, we're using less data
- Still some imbalance

Optimal Full Matching

ORIGINAL SAMPLE



MATCHED SET 5: Discrepancy = D_5



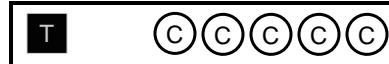
MATCHED SET 1: Discrepancy = D_1



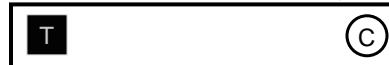
MATCHED SET 2: Discrepancy = D_2



MATCHED SET 3: Discrepancy = D_3



MATCHED SET 4: Discrepancy = D_4



- 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

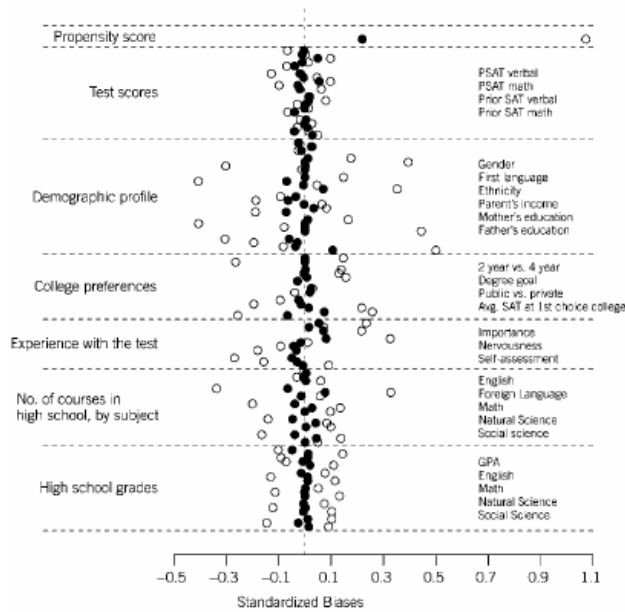


Figure 3. Standardized Biases Without Stratification or Matching, Open Circles, and Under the Optimal [5, 2] Full Match, Shaded Circles.

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

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 (in press [2004])

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 (in press [2004a])

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

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)

Dealing with Missing Data: A Typical Approach

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

Designing Observational Studies

- Big danger in observational studies: bias
- Value of Propensity Scores: reduce bias
- Repeated analyses attempting to balance covariate distributions across treatment groups do not bias estimates of the treatment’s effect on outcome variables.
- So we can (and should) use propensity scores to try to balance covariates.

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.

Techniques for Initial Design of an OS Using Propensity Scores

- Matching
- Subclassification / Stratification
- Weighting
- Goal: Assemble groups of treated and control units such that within each group the distribution of covariates is balanced.
- Allows us to attribute outcome differences to the effect of treatment vs. control.

When Can We Move On?

- Three conditions which must all apply for regression adjustment to be trustworthy:
- Difference in the means of $\text{logit}(\text{PS})$ in the two groups being compared must be small.
- Ratio of variances of $\text{logit}(\text{PS})$ 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 $\text{logit}(\text{propensity scores})$ 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 $\text{logit}(\text{PS})$ 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 $\text{logit}(\text{PS})$ 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 $0.8 \leq \text{resid ratio} \leq 1.2$
- 5% of covariates had $\text{resid ratio} > 2$ or < 0.5

USING PROPENSITY SCORE METHODS EFFECTIVELY

ASA CLEVELAND CHAPTER FALL WORKSHOP OCTOBER 11, 2004

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

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	$\leq .5$.5 - .8	.8 - 1.2	1.2 - 2	> 2	Total
1.09	1.00	3	9	57	26	5	100

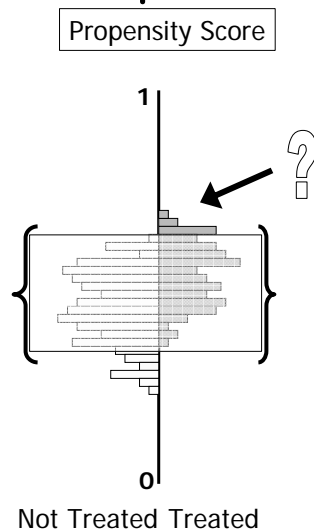
- After The Match

B	R	$\leq .5$.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.

USING PROPENSITY SCORE METHODS EFFECTIVELY

ASA CLEVELAND CHAPTER FALL WORKSHOP OCTOBER 11, 2004

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

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 $\text{logit}(\text{PS})$.
- 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

Subclasses	B	R	Residual Variance Ratios				
			$\leq .5$.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 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).

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?

- Collect data so as to be able to model selection
- Demonstrate selection bias – need for PS
- Ensure covariate overlap for comparability
- Evaluate covariate balance after PS application
- Specify relevant post-adjustment population carefully
- Model or estimate treatment effect in light of PS adjustment / matching / stratification
- Estimate sensitivity of results to potential hidden bias

REFERENCES USED IN BUILDING THE COURSE

1. Rosenbaum PR (2002) *Observational Studies*, 2nd Edition. New York: Springer.
2. Cameron E Pauling L (1976) Supplemental ascorbate in the supportive treatment of cancer: Prolongation of survival times in terminal human cancer. *Proceedings of the National Academy of Sciences (USA)* **73**, 3685-3689.
3. Cepeda MS 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**, 280-287.
4. Cochran WG (1968) The effectiveness of adjustment by subclassification in removing bias in observational studies, *Biometrics* **24**, 295-313.
5. Cochran WG (1983) *Planning and Analysis of Observational Studies*. New York: Wiley.
6. Connors AF Speroff T Dawson NV Thomas CL et al. (1996) The effectiveness of right heart catheterization in the initial care of critically ill patients (with Editorial). *Journal of the American Medical Association*, **276**, 889-897, 915-918.
7. D'Agostino RB (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.
8. D'Agostino RB 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**, 317-329.
9. D'Agostino RB Rubin DB (2000) Estimating and using propensity scores with partially missing data. *Journal of the American Statistical Association*, **95**, 749-759.
10. Earle CC Tsai JS et al. (2001) Effectiveness of chemotherapy for advanced lung cancer in the elderly: Instrumental variable and propensity analysis. *Journal of Clinical Oncology*, **19**, 1064-1070.
11. Gum PA Thamilarasan M et al. (2001) Aspirin use and all-cause mortality among patients being evaluated for known or suspected coronary artery disease (with Editorial). *Journal of the American Medical Association*, **286**, 1187-1194, 1228-1230.
12. Hansen BB (2004) Full matching in an observational study of coaching for the SAT. *Journal of the American Statistical Association*, **467**, 609-618.
13. Hirano K Imbens GW (2001) Estimation of causal effects using propensity score weighting: An application to data on right heart catheterization, *Health Services & Outcomes Research Methodology*, **2** (3-4), 259-278.
14. Hirano K Imbens GW Ridder G (2003) Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, **71**, 1161-1189.
15. Joffe MM Rosenbaum PR (1999) Propensity scores. *American Journal of Epidemiology*, **150**, 327-333.
16. Landrum MB Ayanian JZ (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.

USING PROPENSITY SCORE METHODS EFFECTIVELY

ASA CLEVELAND CHAPTER FALL WORKSHOP OCTOBER 11, 2004

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

17. Love TE Cebul RD Thomas CL Dawson NV (2003) Effect of matching for propensity on the balance of “unmeasured” covariates. *Journal of Clinical Epidemiology*, **56**, 920.
18. Lunceford JK Davidian M (2004) Stratification and weighting via the propensity score in estimation of causal treatment effects : A comparative study. *Statistics in Medicine*, **23**, 2937-2960.
19. Moertel C et al. (1985) High-dose vitamin C versus placebo in the treatment of patients with advanced cancer who have had no prior chemotherapy: A randomized double-blind comparison. *New England Journal of Medicine*, **284**, 878-881.
20. Murray PK, Singer M, Dawson NV, Thomas CL, Cebul RD (2003) Outcomes of rehabilitation services for nursing home residents. *Archives of Physical Medicine and Rehabilitation*, **84**, 1129-1136
21. Posner MA Ash AS 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.
22. Potosky AL Legler J et al. (2000) Health outcomes after prostatectomy or radiotherapy for prostate cancer: Results from the Prostate Cancer Outcomes Study. *Journal of the National Cancer Institute*, **92**, 1582-1592.
23. Rosenbaum PR (1991) Discussing hidden bias in observational studies. *Annals of Internal Medicine*, **115**, 901-905.
24. Rosenbaum PR Rubin DB (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.
25. Rosenbaum PR Rubin DB (1984) Reducing bias in observational studies using subclassification on the propensity score, *Journal of the American Statistical Association*, **79**, 516-524.
26. Rosenbaum PR Rubin DB (1985) Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, **39**, 33-38.
27. Rubin DB (1997) Estimating causal effects from large data sets using propensity scores. *Annals of Internal Medicine*, **127**, 757-763.
28. Rubin DB (2001) Using propensity scores to help design observational studies: Application to the tobacco litigation. *Health Services and Outcomes Research Methodology*, **2**, 169-188.
29. Rubin DB (in press [2004]) On principles for modeling propensity scores in medical research. *Pharmacoepidemiology and Drug Safety*, preprint available online.
30. Rubin DB Stuart EA Zanutto EL (2004) A potential outcomes view of value-added assessment in education. *Journal of Educational and Behavioral Statistics*, **29**, 103-116.
31. Weitzen S Lapane KL et al. (in press [2004a]) Principles for modeling propensity scores in medical research: A systematic literature review. *Pharmacoepidemiology and Drug Safety*, preprint available online.
32. Weitzen S Lapane KL et al. (in press [2004b]) Weaknesses of goodness-of-fit tests for evaluating propensity score models: the case of the omitted confounder. *Pharmacoepidemiology and Drug Safety*, preprint available online.