

Propensity Scores Workshop

github.com/THOMASELOVE/ichps2018

ICHPS: 2018-01-10 (revised)

Designing and Analyzing Observational Health Policy Studies Using Propensity Scores

All Materials: <https://github.com/THOMASELOVE/ichps2018>

This workshop describes and demonstrates effective strategies for using propensity score methods to address the potential for selection bias in observational studies comparing the effectiveness of treatments or exposures.

We review the main analytical techniques associated with propensity score methods (focusing on **matching**, **weighting** and **double robust** techniques) and describe key strategic concerns related to effective propensity score estimation, assessment and display of covariate balance, choice of analytic technique, stability and sensitivity analyses, and communicating results effectively.

Wednesday 2018-01-10: 8 - 10 AM in Crystal Ballroom CD

Overview

All Materials: <https://github.com/THOMASELOVE/ichps2018>

① Fundamentals

- Exposure Selection Bias in an Observational Study
- What is a propensity score and how should we build one?

② Using the propensity score to deal with selection bias

- Via matching
- Via weighting
- Double robust approach (weighting + regression)

③ Strategies for effective design and analyses of health policy studies

- Assessing and Displaying Covariate Balance
- Stability and Sensitivity Analysis

Fundamentals

Designing A Study: ASA and Mortality in Heart Subjects

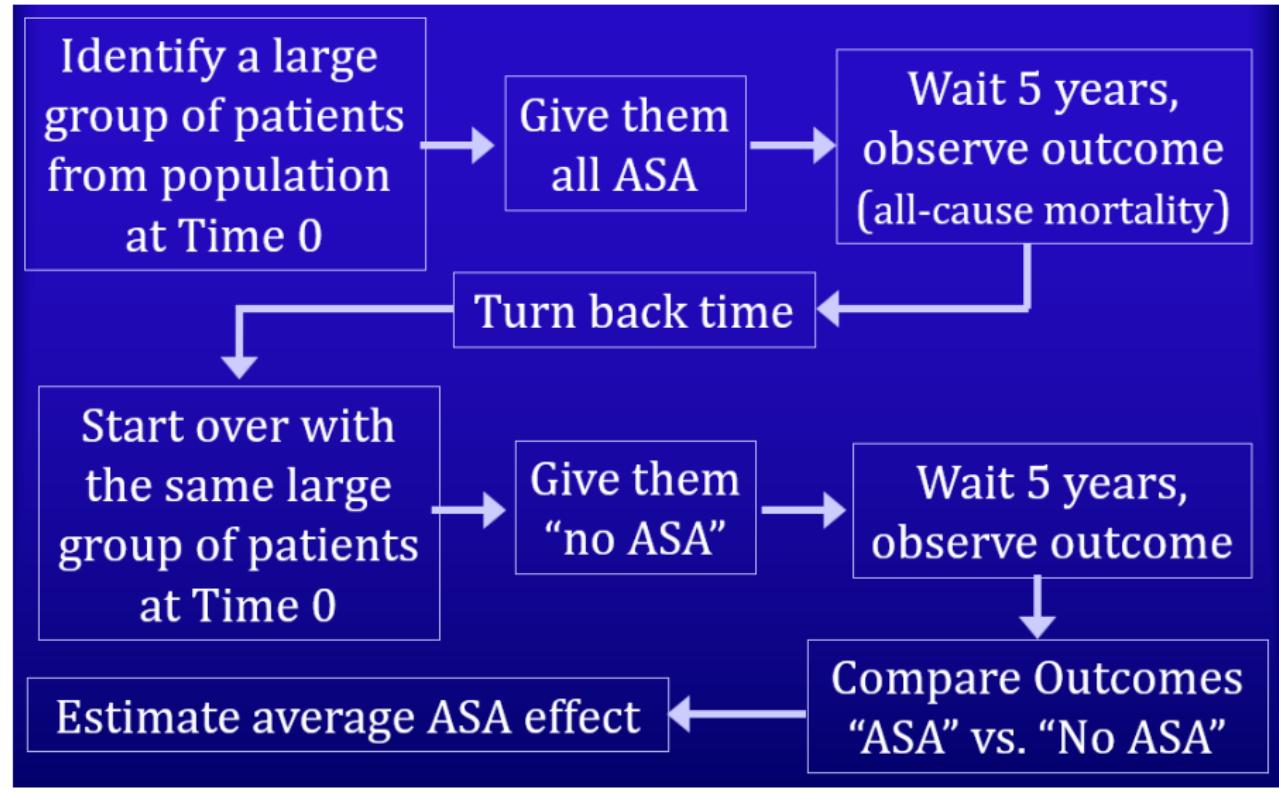
- Suppose you want to study the effect of aspirin (acetylsalicylic acid: ASA) on all-cause mortality.
- You identify an interesting group of Subjects as those undergoing stress echocardiography.
 - Your goal is to compare ASA Subjects to “no ASA” Subjects

What would be the **ideal** study?

Step 1. Identify a large group of Subjects from the population at Time 0.

Step 2?

ASA and Mortality: Ideal Study



ASA and Mortality: Best Practical Study?

Identify a large group of patients from population at Time 0

Divide into two groups, at random

Give one group ASA and the other “no ASA”

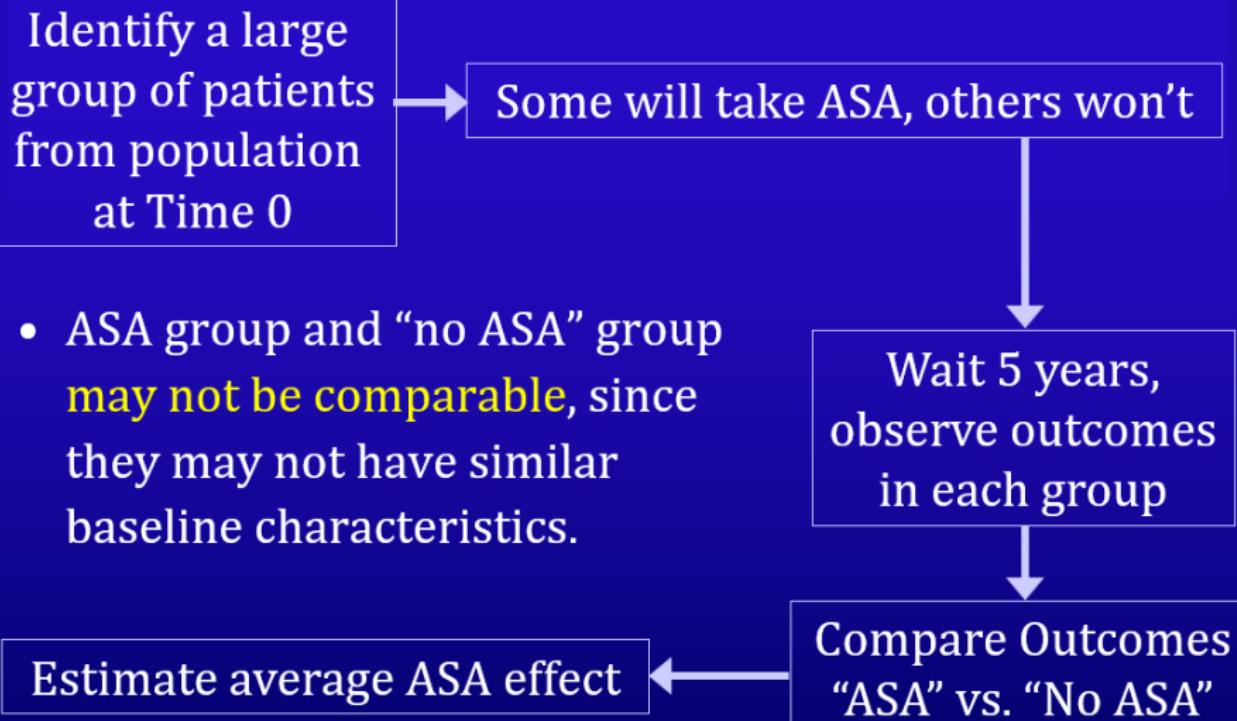
- ASA group and “no ASA” group should be comparable, since randomizing yields groups with similar baseline characteristics.

Wait 5 years, observe outcomes in each group

Compare Outcomes “ASA” vs. “No ASA”

Estimate average ASA effect

ASA and Mortality: Observational Study?



Simple Observational Studies

Specify a population, outcome, exposure and covariates.

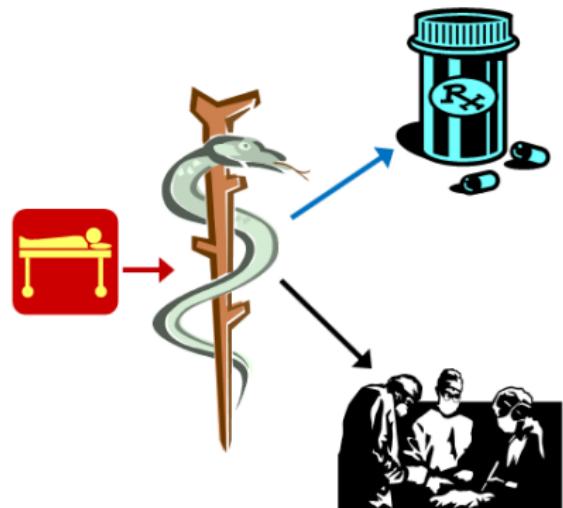
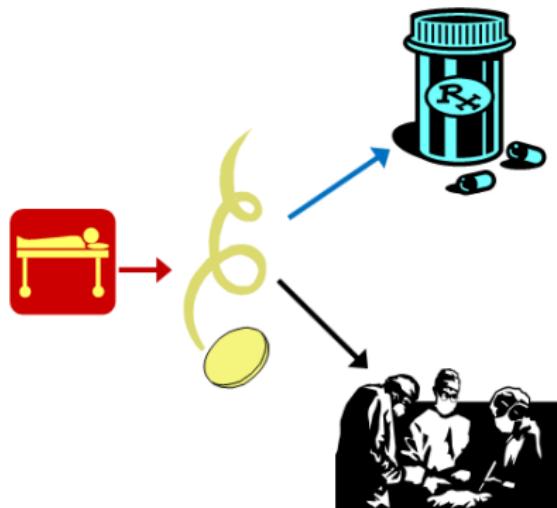
- We want to make a fair comparison between the exposed group and the control group within a population, in terms of an outcome of interest.
- We can obtain covariates that describe the subjects before they were allocated to the exposed and control groups.

But

- We **cannot** use randomization to ensure that the groups will be comparable in terms of the covariates.

No randomization forces the investigator to think hard about how the exposures are “assigned” . . .

Randomized Trials vs. Observational Studies



Randomization “ensures” that subjects receiving different exposures are comparable.

In **observational** studies, the researcher does not randomly allocate the exposures.

How Do We Avoid Being Misled by Observational Studies?

- What differentiates an observational study from a randomized controlled trial?
 - One key element: potential for selection bias.
- What is selection bias and what can we do about it?
 - Baseline characteristics of comparison groups are different in ways that affect the outcome.

We will often distinguish between overt and hidden bias.

- Overt Bias (seen in data - propensity scores can help)
- Hidden Bias (required data not collected - requires sensitivity analyses)

What do you want to know about an intervention?

- Response: Can we estimate the impact of the intervention? Can we estimate costs and benefits?
- Predictors: Can we “mine” for attributes that help predict response to the intervention?
- Evaluation: Can we fairly estimate the average health impact of our intervention?
- Target Evaluation: Can we identify likely responders?

This is based on a marketing list originally posted at anabus.com

The database you wish you had...

Subject	Outcome if Subject receives Aspirin	Outcome if Subject does not receive Aspirin
A	10	12
B	9	11
C	7	9
D	11	13
E	6	8
F	9	11

Harsh Reality

Subject	Outcome if Subject receives Aspirin	Outcome if Subject does not
A	10	
B	9	
C	7	
D	-	
E	-	
F	-	

This is based on a marketing list originally posted at anabus.com

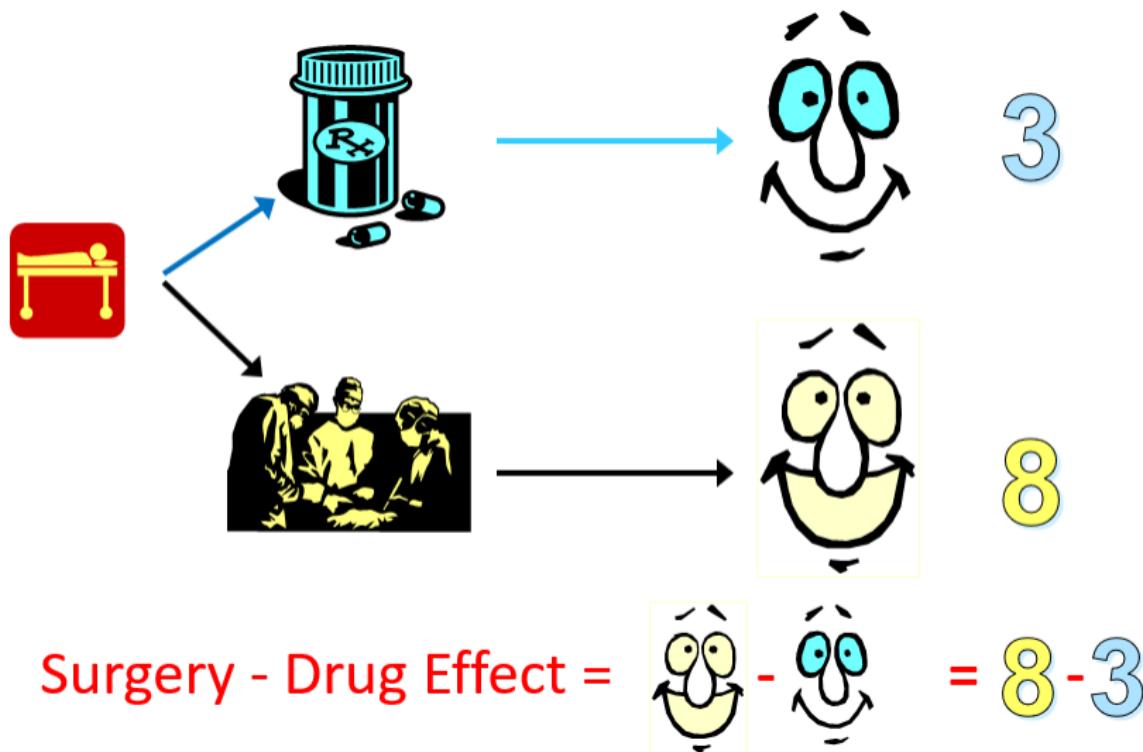
Rubin Causal Model

Builds on the potential outcomes framework¹.

- Key assumption: Strongly Ignorable Exposure Assignment
 - Potential outcomes Y_{ASA} , Y_{noASA} are assumed conditionally independent of exposure assignment, given covariates \mathbf{X} .
 - This is a “no hidden bias” assumption, where we assume \mathbf{X} contains all relevant information about exposure assignment.
- Data? Clinical-Researchers vs. Economists

¹Rubin 1997, Rosenbaum 2002

Causal Effects in terms of “Potential Outcomes”



Assessing the Causal Effect of an Exposure on an Outcome

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

The Propensity Score

$$PS = \Pr(\text{received exposure} | \mathbf{X})$$

The propensity score is...

- the conditional probability of receiving the exposure given a particular set of covariates
- a way of projecting meaningful covariate information for a given subject into a single composite summary score in (0, 1)
- a tool that lets us account for *overt* selection bias (things contained in \mathbf{X}) but not (directly) for the potential biasing effects of omitted/hidden covariates
- often, but not inevitably, fit with a “kitchen sink” logistic regression²

$$\ln\left(\frac{PS}{1 - PS}\right) = \beta_0 + \beta_1 X_1 + \dots + \beta_k X_k$$

²See McCaffrey et al 2004 on boosting, and see Brookhart 2006 on variable selection.

The New Database, Simply

Subject	Propensity to receive Aspirin	Outcome if Subject receives Aspirin
A	0.81	10
B	0.51	9
C	0.31	7
D	0.79	-
E	0.51	-
F	0.29	-

We could then match up the subjects, and plug in the estimates, for instance.

- So how do we model propensity for aspirin, or any other exposure?

Our Goal

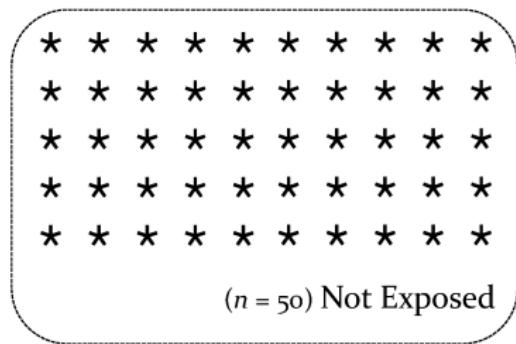
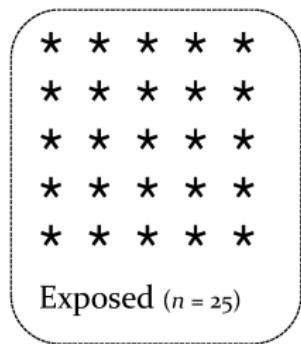
Fit a model that predicts the probability of exposure, given a set of covariates.

We anticipate that our exposure group will have different distributions of the covariates than our control group at baseline.

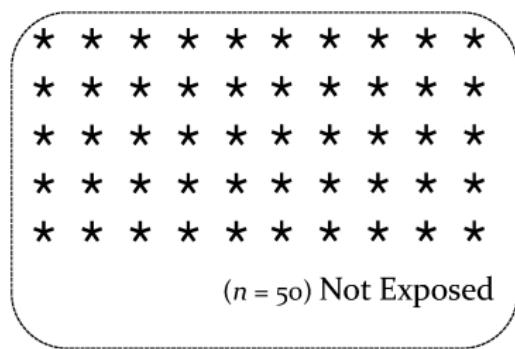
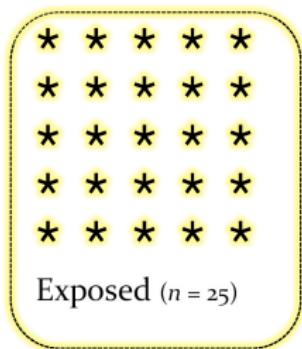
- Perhaps the subjects taking aspirin are older
- Perhaps they are more likely to be taking other medications
- Perhaps they are more likely to have certain comorbidities

We want to wind up with a fair basis for comparison between exposed and control subjects.

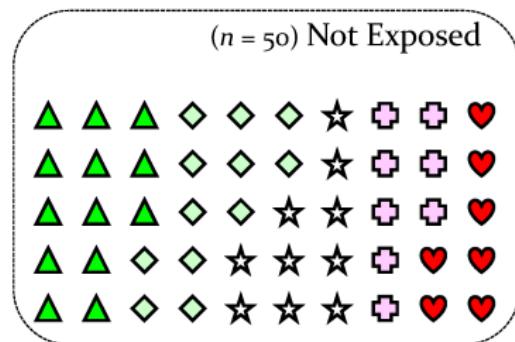
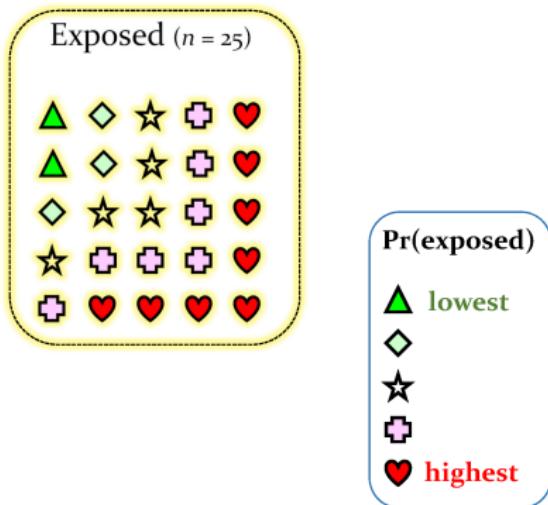
Simple Observational Study



Apply the Exposure



Characterize by propensity to receive the exposure...



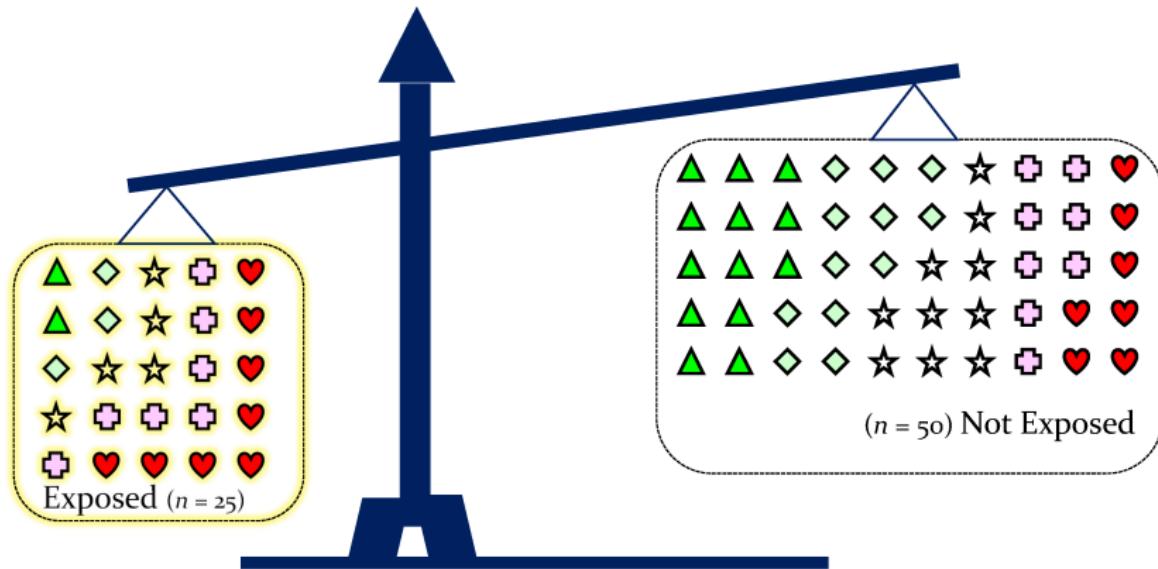
Estimating Propensity for Exposure Given Covariates

Usual Tool: Logistic Regression Model for Exposure Allocation

- Consider including any variables that have a relationship to the exposure decision
 - Precede the exposure in time
 - Relevant to exposure assignment
 - Relationship to outcome? (some controversy here)
- No information included on the actual exposure received, or on the outcome(s)
- In early stages, always err on the side of inclusion

$$\text{logit}[\Pr(\text{aspirin})] = \beta_0 + \beta_1 \text{Age} + \beta_2 \text{Statin} + \dots + \beta_k \text{Diabetes}$$

Are baseline characteristics in balance?



Model Without the Propensity Score

Outcome = $\beta_0 + \beta_1 * \text{Exposure}$,

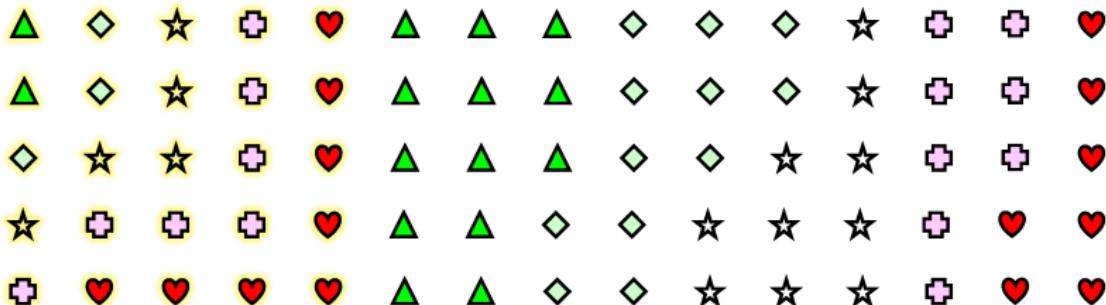
for pool of 75 subjects

We interpret β_1 as the exposure's effect.



Direct Adjustment for Propensity Score

Outcome = $\beta_0 + \beta_1 * \text{Exposure} + \beta_2 * \text{Propensity Score}$,
Again, across entire pool of 75 subjects



Propensity Score Models: What to Worry About...

- ① Do you have a reasonable sample size to build a logistic regression model, e.g., at least 96 subjects + some function of the number of candidate predictors³?
- ② Is your logistic regression model parsimonious?
- ③ Are your predictors correlated with one another?
- ④ Are your predictors statistically significant?
- ⑤ Have you performed appropriate diagnostic checks?
- ⑥ Have you done bootstrap analyses to assess shrinkage?
- ⑦ Have you used cross-validation to aid in model selection?
- ⑧ Have you done external validation of your model on new data?
- ⑨ Does an ROC-curve analysis suggest your model does well in terms of rank-order discrimination?
- ⑩ Have you determined that your model's predictions are well-calibrated?

³see Frank Harrell reference

What to Actually Worry About

None of those things.

Instead, we simply ensure that the fitted propensity scores (when used in matching, weighting, etc.) adequately balance the distribution of covariates across the exposure groups.

Again, we want to wind up with a **fair basis for comparison** between exposed and control subjects.

Matching Using The Propensity Score

Multivariate Matching using the Propensity Score

Goal: Emulate a randomized clinical trial in matching⁴, then use standard analyses to compare matched sets.

Design: Exposed subjects matched to people who didn't receive exposure but who had similar propensity to receive exposure (matching exposed to unexposed “clones”)

- Match subjects on a scalar (the propensity score) so that the **groups** are similar in terms of distributions of multiple covariates simultaneously.

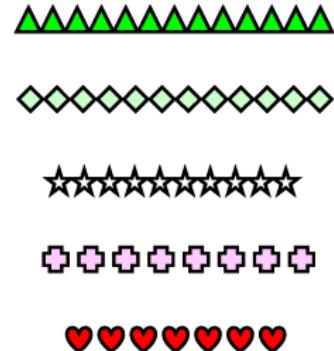
⁴Seminal paper: Rosenbaum & Rubin 1985. See also Rosenbaum 2010 and Rosenbaum 2017

Propensity Score Matching (1:1)

Exposed Pool



“Not Exposed” Pool



Propensity Score Matching (1:1)

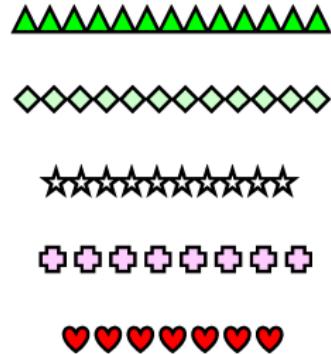
Exposed Pool



Select an **exposed** subject,
perhaps at random



"Not Exposed" Pool



Propensity Score Matching (1:1)

Exposed Pool



“Not Exposed” Pool



Find a matching subject
from the **not exposed** pool
(match on propensity score)



Propensity Score Matching (1:1)

Exposed Pool



Form a matched pair



We're matching without replacement.

"Not Exposed" Pool



Propensity Score Matching (1:1)

Exposed Pool



Select another **exposed** subject



"Not Exposed" Pool



Propensity Score Matching (1:1)

Exposed Pool



Find a good match
among the subjects
not exposed.



"Not Exposed" Pool



Propensity Score Matching (1:1)

Exposed Pool



A second matched pair!



"Not Exposed" Pool



Propensity Score Matching (1:1)

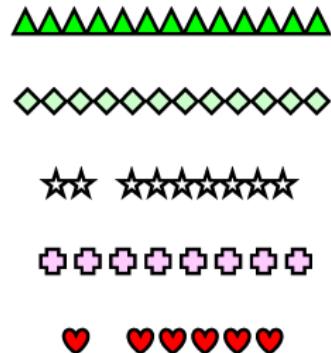
Exposed Pool



Keep matching, until
we can find no more
acceptable matches



"Not Exposed" Pool



Propensity Score Matching (1:1)

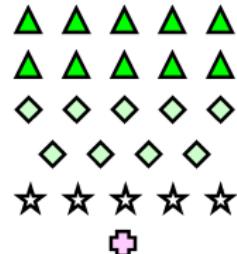
Exposed Pool
(unmatched)



Matched Set
(24 pairs)

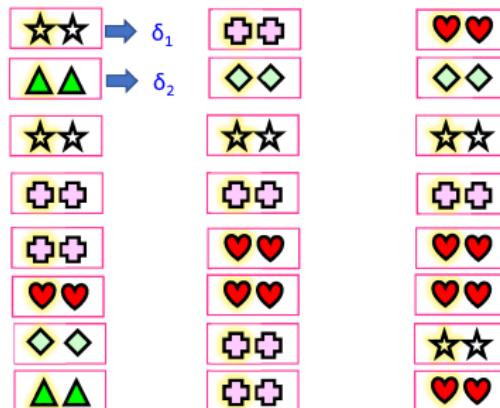
★★	♥♥	⊕⊕
▲▲	◊◊	◊◊
★★	★★	★★
⊕⊕	⊕⊕	⊕⊕
⊕⊕	♥♥	♥♥
♥♥	♥♥	♥♥
◊◊	★★	⊕⊕
▲▲	♥♥	⊕⊕

“Not Exposed” Pool
(unmatched)



Propensity Score Matching (1:1)

Matched Set
(24 pairs)

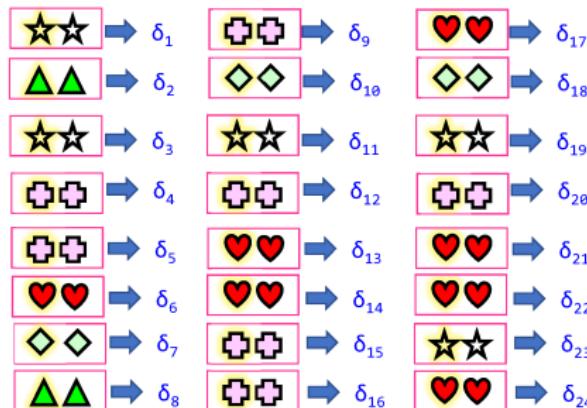


Within each matched pair,
compare outcome in exposed
subject to outcome in “not
exposed” subject.

Estimated outcome effect
Within a specific pair j is
estimated by δ_j

Propensity Score Matching (1:1)

Matched Set
(24 pairs)



Within each matched pair,
compare outcome in exposed
subject to outcome in “not
exposed” subject.

Use standard methods for
matched samples (e.g., paired t
tests) to estimate the causal
effect of the exposure on the
outcome based on the δ
estimates from the pairs

The Aspirin Use and Mortality Study

Aspirin Use and All-Cause Mortality Among Patients Being Evaluated for Known or Suspected Coronary Artery Disease

A Propensity Analysis

Patricia A. Gum, MD

Maran Thamilarasan, MD

Junko Watanabe, MD

Eugene H. Blackstone, MD

Michael S. Lauer, MD

Context Although aspirin has been shown to reduce cardiovascular morbidity and short-term mortality following acute myocardial infarction, the association between its use and long-term all-cause mortality has not been well defined.

Objectives To determine whether aspirin is associated with a mortality benefit in stable patients with known or suspected coronary disease and to identify patient characteristics that predict the maximum absolute mortality benefit from aspirin.

Aspirin Use and Mortality

Study of 6,174 adults at Cleveland Clinic (1990-1998) undergoing stress echocardiography to evaluate coronary disease⁵.

- 2,310 (37%) were taking aspirin (the exposure)
- 31 covariates are reported, including demographics, clinical history, medications, cardiovascular assessments, and exercise capacity
- Outcome: all-cause mortality (median follow-up 3.1 years)

Analysis without covariates:

- 4.5% of the aspirin and 4.5% of the non-aspirin patients died.
- The unadjusted hazard ratio was 1.08 (0.85, 1.39).

⁵ see Gum et al. 2001. The study began with 9,954 consecutive adults.



Gum (2001) Table 1

Table 1. Baseline and Exercise Characteristics According to Aspirin Use*

Variable	Aspirin (n = 2310)	No Aspirin (n = 3864)	P Value	Δ_{A-No}	Δ_{Std}
Demographics					
Age, mean (SD), y	62 (11)	56 (12)	<.001	6.0	52.1
Men, No. (%)	1779 (77)	2167 (56)	<.001	20.9	45.5
Clinical history					
Diabetes, No. (%)	388 (17)	432 (11)	<.001	5.6	16.2
Hypertension, No. (%)	1224 (53)	1569 (41)	<.001	12.4	25.0
Tobacco use, No. (%)	234 (10)	500 (13)	.001	-2.8	-8.8
Prior coronary artery disease, No. (%)	1609 (70)	778 (20)	<.001	49.5	114.8
Prior coronary artery bypass graft, No. (%)	689 (30)	240 (6)	<.001	23.6	64.6
Prior percutaneous coronary intervention, No. (%)	667 (29)	148 (4)	<.001	25.0	72.0
Prior Q-wave MI, No. (%)	369 (16)	285 (7)	<.001	8.6	27.0
Atrial fibrillation, No. (%)	27 (1)	55 (1)	.04	-0.3	-2.3
Congestive heart failure, No. (%)	127 (6)	178 (5)	.12	0.9	4.1
Medication use					
Digoxin use, No. (%)	171 (7)	216 (6)	.004	1.8	7.4
β -Blocker use, No. (%)	811 (35)	550 (14)	<.001	20.9	49.9
Diltiazem/verapamil use, No. (%)	452 (20)	405 (10)	<.001	9.1	25.6
Nifedipine use, No. (%)	261 (11)	283 (7)	<.001	4.0	13.7
Lipid-lowering therapy, No. (%)	775 (34)	380 (10)	<.001	23.7	60.1
ACE inhibitor use, No. (%)	349 (15)	441 (11)	<.001	3.7	10.9

Using Standardized Differences to Quantify Covariate Imbalance

For continuous variables,

$$\Delta_{Std} = \frac{100(\bar{x}_{ASA} - \bar{x}_{No})}{\sqrt{\frac{s_{ASA}^2 + s_{No}^2}{2}}}$$

For binary variables,

$$\Delta_{Std} = \frac{100(p_{ASA} - p_{No})}{\sqrt{\frac{p_{ASA}(1-p_{ASA}) + p_{No}(1-p_{No})}{2}}}$$

Beta-Blocker	Aspirin	No Aspirin	Δ_{Std}
Before Match	35.1% (811/2310)	14.2% (550/3864)	49.9%
After Match	26.1% (352/1351)	26.5% (358/1351)	-1.0%

Gum (2001) Table 1 (continued)

Table 1. Baseline and Exercise Characteristics According to Aspirin Use*

Variable	Aspirin (n = 2310)	No Aspirin (n = 3864)	P Value	Δ_{A-No}	Δ_{Std}
Cardiovascular assessment and exercise capacity					
Body mass index, mean (SD), kg/m ²	29 (5)	30 (7)	<.001	-1	-16.4
Ejection fraction, mean (SD), %	50 (9)	53 (7)	<.001	-3	-37.2
Resting heart rate, mean (SD), beats/min	74 (13)	79 (14)	<.001	-5	-37.0
Resting blood pressure, mean (SD), mm Hg					
Systolic	141 (21)	138 (20)	<.001	3	14.6
Diastolic	85 (11)	86 (11)	.04	-1	-9.1
Purpose of test to evaluate chest pain, No. (%)	300 (13)	468 (12)	.31	0.9	2.6
Mayo Risk Index ≥1, No. (%)†	2021 (87)	2517 (65)	<.001	22.3	54.5
Peak exercise capacity, mean (SD), METs					
Men	8.6 (2.4)	9.1 (2.6)	<.001	-0.5	-20.0
Women	6.6 (2.0)	7.3 (2.1)	<.001	-0.7	-34.1
Heart rate recovery, mean (SD), beats/min	28 (11)	30 (12)	<.001	-2.0	-17.4
Ischemic ECG changes with stress, No. (%)	430 (24)	457 (14)	<.001	6.8	19.0
Echocardiographic left ventricular ejection fraction ≤40%, No. (%)	321 (14)	226 (6)	<.001	8.0	27.2
Stress-induced ischemia on echocardiography, No. (%)	495 (21)	436 (11)	<.001	10.1	27.7
Fair or poor physical fitness for age and sex, ¹³ No. (%)	714 (31)	1248 (38)	.26	-1.4	-3.0

*MI indicates myocardial infarction; ACE, angiotensin-converting enzyme; MET, metabolic equivalent task; and ECG, electrocardiogram.

†The Mayo Risk Index is described in the "Methods" section.

Pre-Matching Characteristics by Aspirin Use

Do the aspirin and non-aspirin groups show important differences in distribution at baseline?

- At baseline, aspirin patients display higher risk of mortality, in general
 - they are older, more likely to be male, and more likely to have a clinical history
 - they are more likely to be on other medications than non-aspirin subjects
 - their cardiovascular assessments are (generally) worse and have worse exercise capacity
- The table reports on 31 characteristics prior to matching
 - 24 of 31 have p values below 0.001, one more is $p = 0.001$, and two more are $p = 0.04$
 - 25 of 31 have standardized differences of more than 10%, and six are more than 50%

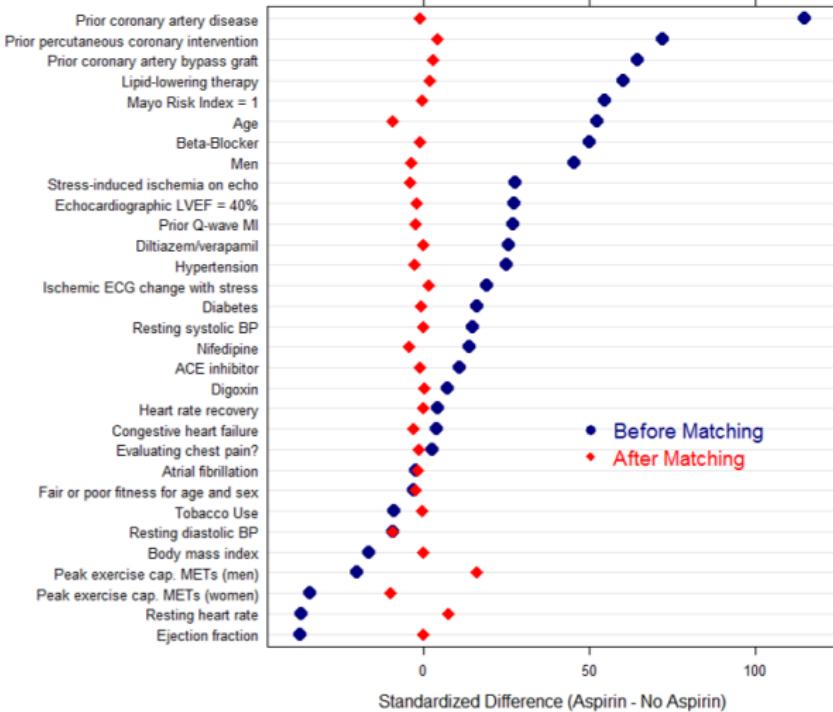
Propensity Score Matching

A logistic regression model was used to estimate the propensity for aspirin use (31 covariates, main effects) for each of the 6,174 subjects. $C = 0.83$.

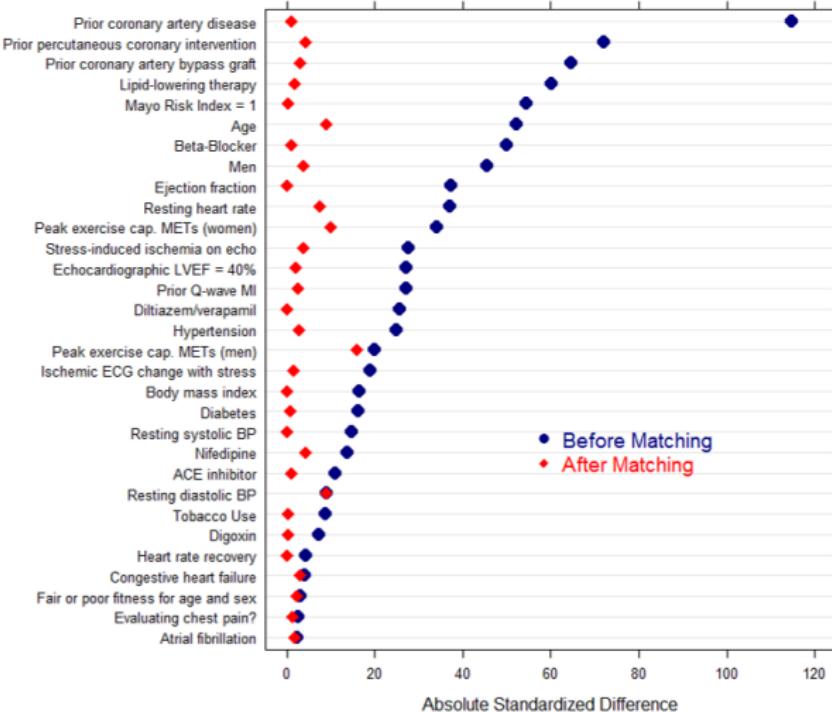
Matching Approach (Greedy and Incomplete):

- Tried to match each aspirin user to a unique non-user with a propensity score that was identical to five 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 1,351 (58%) of the 2,310 aspirin patients to 1,351 unique non-users.

Standardized Difference Plot (Aspirin - No Aspirin)



Absolute Standardized Differences (Aspirin vs. No Aspirin)



Matching with Propensity Scores

1,351 aspirin subjects matched well to 1,351 unique non-aspirin subjects

- Big improvement in covariate balance
- Table 1 for matched group looks like an RCT
- Can analyze the resulting matched pairs with standard methods (stratified Cox models, etc.)

Matching still incomplete (lots of possible bias here) and this isn't the best algorithm for matching, either...

Estimating the Hazard Ratios

During follow-up, 153 (6%) of the 2,702 matched patients died.

- In the matched group, aspirin use was associated with a lower risk of death (4% vs. 8%, $p = 0.002$)

Approach	n	Est. HR	95% CI
Full sample, no adjustment	6174	1.08	(0.85, 1.39)
Full sample, no PS, adj. for all covariates	6174	0.67	(0.51, 0.87)
PS-matched sample	2702	0.53	(0.38, 0.74)
PS-matched, adj. for PS and all covariates	2702	0.56	(0.40, 0.78)

Aspirin Conclusions / Caveats

- Subjects included in this study *may* be a more representative sample of real world patients than an RCT would provide.
 - On the other hand, they were getting cardiovascular care at the Cleveland Clinic.
 - And there are some inclusion and exclusion criteria here, too.
- PS matching still isn't randomization, we can only account here for the factors that were measured, and only as well as the instruments can measure them.
- There's no information here on aspirin dose, aspirin allergy, duration of treatment or medication adjustments.

Statistical Concerns

- This isn't the best way to match, certainly.
- There's no formal assessment of sensitivity to hidden bias.
- Looks like they avoided the issue of missing data.

Dealing with Missing Data

What if we have missing covariate values⁶?

- The pattern of missing covariates is easy to balance
 - Add a missingness indicator variable for all covariates with NA
 - Then fill in values for those cases in the original variable before estimating PS
- Matching on this augmented PS will tend to balance the observed covariates and the **pattern** of missingness, but yields no guarantee that the missing values themselves are actually balanced.

⁶For more on these issues, try D'Agostino 1998 and D'Agostino and Rubin 2000



Matching, 1

Matching is a fundamental part of the toolbox. For a book-length treatment, I recommend Rosenbaum 2010.

- Propensity scores facilitate matching on multiple covariates at once.
 - Matching is especially attractive when covariates classify subjects into many small categories.
- Matching on a multivariate distance within PS calipers often beats matching on the PS alone, especially if you can pre-specify pivotal covariates.
 - Matching within PS calipers followed by additional matching on key prognostic covariates is an effective method for both reducing bias and understanding the effects of specific covariates.
 - Matching on $\text{logit}(\text{PS})$ rather than on raw PS can often improve yield.

Matching, 2

- If match is incomplete, it's especially useful to consider both matching and non-matching analyses
- Optimal matches, full matches, cardinality matches, genetic matches and other more sophisticated matching approaches can be fruitful.
- Matching can be especially attractive if data are costly - we can match on what we have first, and then collect new data only on the pre-matched subjects.

Weighting using the Propensity Score

Propensity Score Weighting

Adjusting for the propensity score removes the bias associated with differences in the observed covariates in the exposed and control groups.

One way to implement this is to **reweight** exposed and control observations (or just controls, sometimes) to make them representative of the population of interest.

- PS methods generally lead to more reliable estimates of association than multiple regression, especially if there is a substantial selection or other overt bias.
- We can get the benefits of matching while still using all of the collected data.
- We can incorporate propensity weighting along with survey weighting, when oversampling is done, for instance.
- We can incorporate weighting with regression adjustment on the propensity score, producing a double robust estimate.

Propensity Score Weighting (“ATT”)

All Exposed
get **weight 1**



Propensity Score Weighting (“ATT”)

All Exposed
get weight 1



“Not Exposed”
unweighted



Propensity Score Weighting (“ATT”)

All Exposed
get weight 1



“Not Exposed”
weighted



Propensity Score Weighting (“ATT”)

All Exposed
get weight 1



“Not Exposed”
weighted



“Not Exposed”
unweighted



Propensity Score Weighting (“ATT”)

All Exposed
get weight 1



Average Outcome
with Exposure

“Not Exposed”
weighted



Outcome without Exposure
(weighted)



“Weighted Average”
Effect of Exposure on
Outcome

ATT Weighting using the Propensity Score

ATT = average treatment effect on the treated

- Let every exposed (treated) subject's weight be 1.
- A control subject's weight is a function of its propensity for exposure

$$w_j = \frac{PS_j}{1 - PS_j}$$

ATT estimate = Average outcome for treated group - PS weighted outcome for control group

ATE Weighting using the Propensity Score

Alternatively, we can reweight both exposed and control patients to obtain an average treatment effect estimate⁷.

- An exposed (treated) subject's weight is the inverse of its propensity score.

$$w_j = \frac{1}{PS_j}$$

- A control subject's weight is the inverse of one minus its propensity for exposure.

$$w_j = \frac{1}{1 - PS_j}$$

⁷For more, see Rubin 2001, and Lunceford and Davidian 2004

The Support / RHC Study

Studying Right Heart Catheterization in SUPPORT

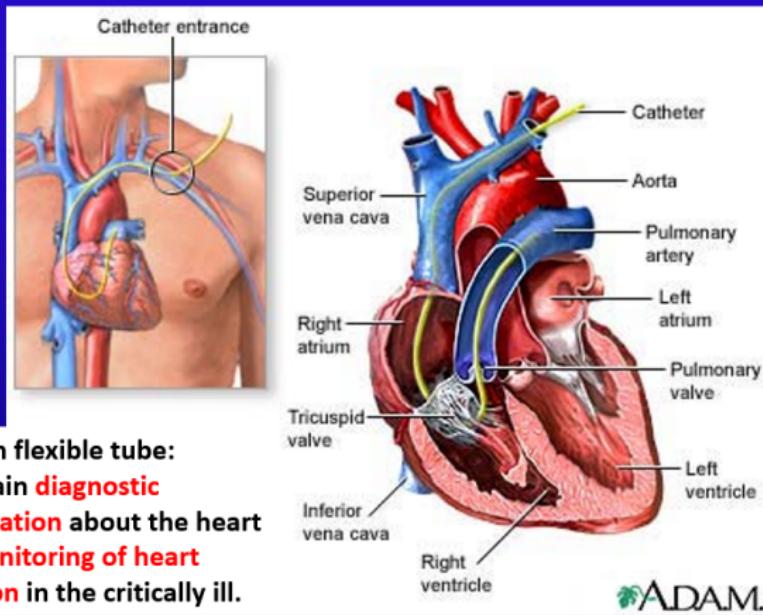
SUPPORT: Study to Understand Prognoses and Preferences for Outcomes and Risks of Treatments⁸

- Goal: Examine the association between the use of RHC during the first 24 hours in the ICU and outcomes
- Outcomes: survival, length of stay, intensity and costs of care
- Sample: 5,735 critically ill adult ICU patients in nine disease categories

Study was prospective!

⁸Connors et al. 1996

Right Heart / Swan-Ganz / Pulmonary Artery Catheterization



Pass a thin flexible tube:

1. to obtain **diagnostic information** about the heart
2. for **monitoring of heart function** in the critically ill.

ADAM.

<http://www.nlm.nih.gov/medlineplus/ency/imagepages/18087.htm>

Does the RHC do more harm than good?

Prior (small) observational studies comparing RHC to non-RHC patients:

- RR of death higher in RHC elderly patients than non-RHC elderly
- RR of death higher in RHC patients with acute MI than non-RHC patients with MI
- Patients with higher than expected RHC use had higher mortality

Big Problem: Selection Bias. Physicians (mostly) decide who gets RHC and who doesn't.

Why not a RCT?

- RHC directly measures cardiac function
- Some MDs believe RHC is necessary to guide therapy for some critically ill patients
- Procedure is very popular - existing studies haven't created equipoise

81 Characteristics used to predict PS(RHC usage)

- Age, Sex, Race
- Education, Income, Insurance
- Primary and Secondary Disease category
- Admission diagnosis category (12 levels)
- ADL and DASI 2 weeks before admission
- DNR status on day 1
- Cancer (none, local, metastasized)
- 2 month survival model
- Weight, temperature, BP, heart rate, respiratory rate
- Comorbid illness (13 categories)
- Body chemistry (pH, WBC, PaCO₂, etc.)

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

RHC vs. Non-RHC patients

RHC patients were more likely to

- Be male, have private insurance, enter the study with ARF, MOSF or CHF

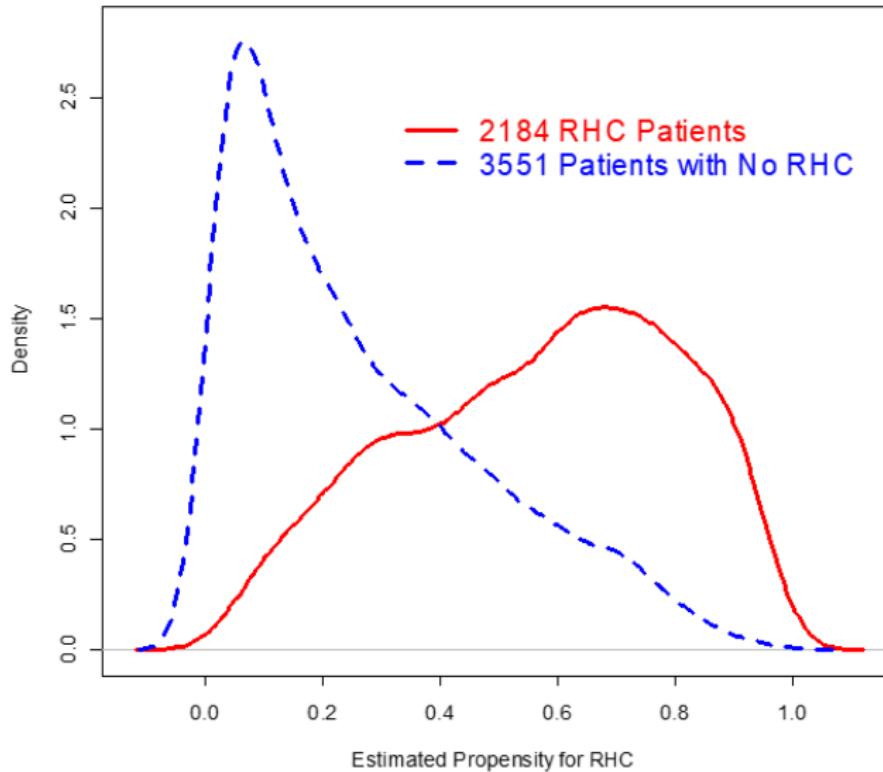
RHC patients were less likely to

- Be over 80 years old, have cancer, have a DNR order in the first 24 hours of hospitalization

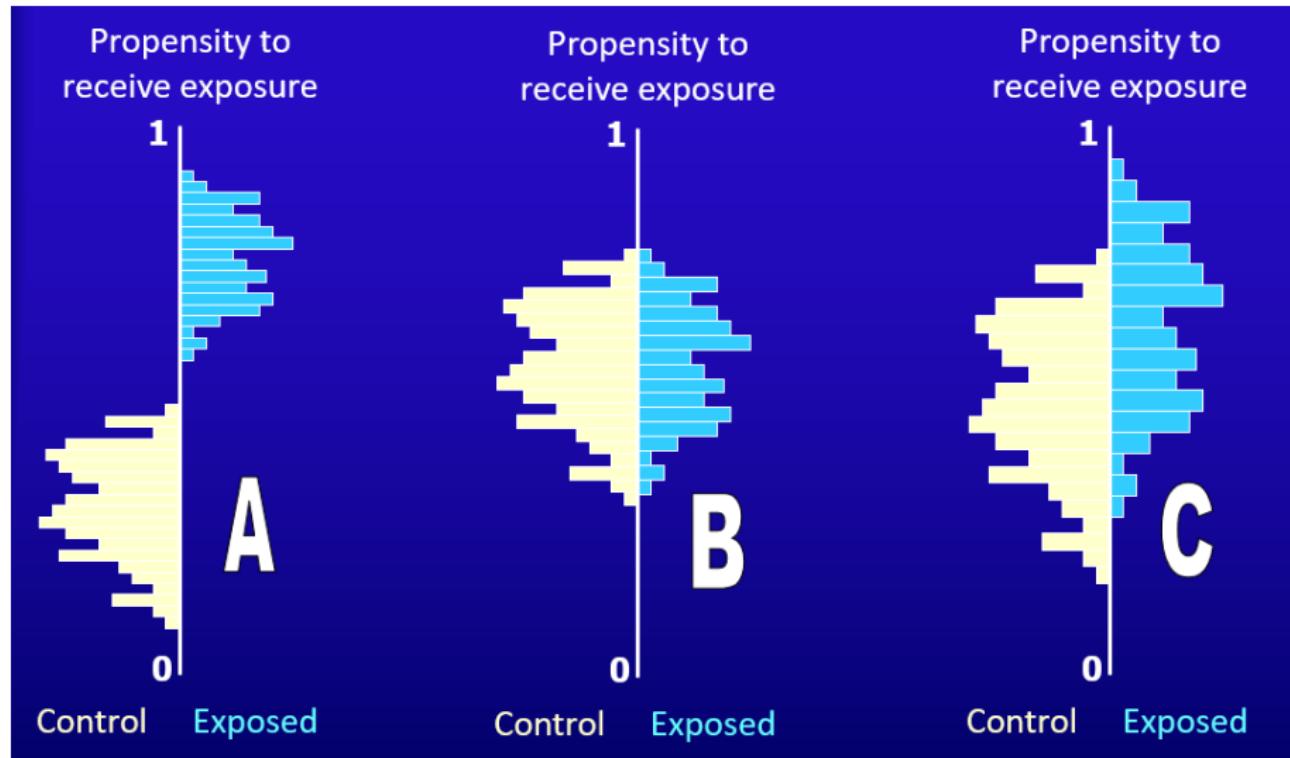
RHC patients had significantly

- Fewer comorbid conditions,
- More abnormal results of vital signs, WBC count, albumin, creatinine, etc.
- Lower model probability of 2-month survival

How Much Overlap do we see in the RHC data?



How Much Overlap do we want?



Right Heart Catheterization and the Perils of Selective Weighting

- 5,735 hospitalized patients in SUPPORT study
 - 2,184 treated (RHC) and 3,551 controls (no RHC).

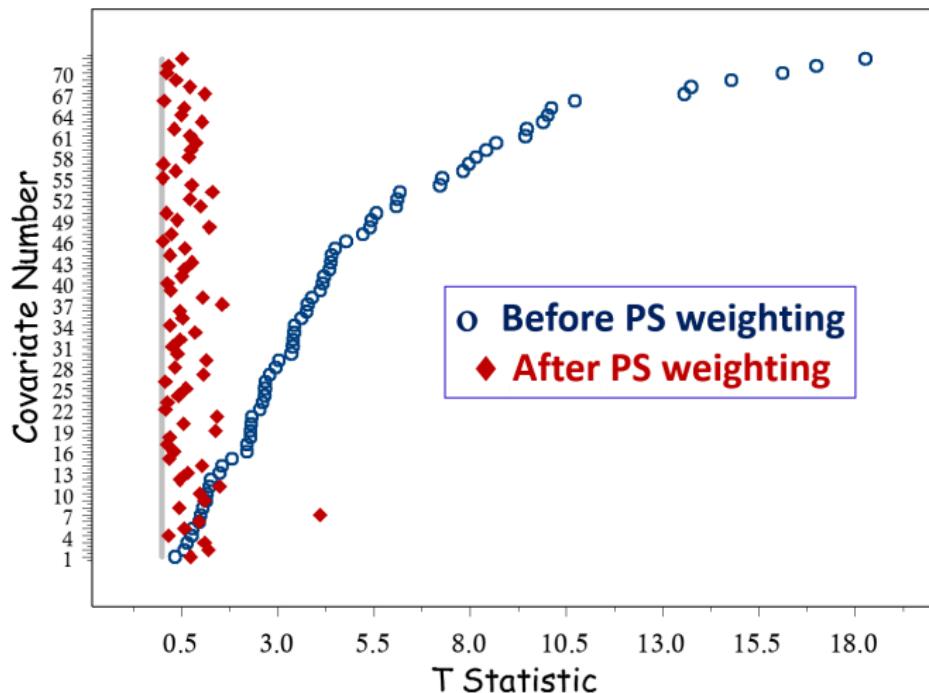
Reweighting each treated patient by $1/\text{PS}$, and each control patient by $1/(1-\text{PS})$.

- PS model estimated by Hirano and Imbens⁹ using 57 of 72 available covariates
 - Selected only those with $|t| > 2.0$
 - Serum potassium, for instance, prior to weighting showed a mean of 4.04 in the RHC group and 4.07 in the “No RHC” group, for a $t = -0.99$, so it was not included in the propensity model.

⁹Hirano and Imbens 2001, Connors 1996, Hirano, Imbens & Ridder 2003

Results of this Weighting Approach

Absolute T Statistics for RHC vs. No RHC Group Means



Effectiveness of RHC Propensity Score Weighting

- The weighting is based on a propensity model including 57 of the 72 covariates.
- Serum potassium not included in this PS.
- Most means are much closer, although six variables become less balanced (larger absolute t statistic) after weighting. None of these six were in the 57-variable PS model.
- Weighting by the propensity score appears to balance control and treatment groups well.

A “Double Robust” Estimator

- ① Fit propensity score model
 - ② Weight the individual subjects (ATT, commonly) by the propensity score.
 - ③ Directly adjust (via regression) for the propensity score in estimating the treatment effect.
-
- Forces you to think hard about selection.
 - You don't care about parsimony in the PS, so you can maximize predictive value.
 - Can fit a very complex PS model, and a smaller outcome model.
 - Some hope that if PS model or weighting is helpful, the combination will be helpful.

What Propensity Scores Can and Cannot Do

- If we match exposed subjects to controls with similar propensity scores, we can behave as if they had been randomly assigned to exposures.
- Or, if we use weighting with or without additional regression to adjust for propensity to get treatment, we can compare exposed subjects to controls without worrying about the impact of baseline differences we've measured on selection to exposure.
- But if our propensity model misses an important reason why subjects are selected to an exposure, we'll be in trouble, and we'll never know it.

Sensitivity Analysis

The Role of Assumptions

Scenario	Analytic Goal	Role of Assumptions
Randomized Experiments	Testing H_0_: No effect	None
Randomized Experiments	Estimating effects, CIs	Minor
Observational Studies	Anything at all	Massive

- **All** observational studies are potentially affected by hidden bias.
 - Sensitivity analyses are needed in any such study.

Stability Analyses

- Are our conclusions mere artifacts of analytic decisions, or are they stable over analyses that differ in (apparently) innocuous ways?
- Examine a series of discrete decisions, hoping conclusions mostly don't change if the decision is changed.

Sensitivity Analyses

- How much do plausible changes in assumptions change conclusions?
- How much hidden bias would have to be present to alter the study's conclusions?

We want to assess the potential for **unmeasured** covariates to change our conclusions.

- To maximize trouble, the unmeasured covariate must be independent of the variables in our propensity model.
 - We missed a dimension of the problem, or our measure is terribly weak.
 - This is part of the motivation to be inclusive in building the PS model.

Idealized Standards for a Sensitivity Evaluation

- Logic, Theory and Empirical Evidence
- “It is unlikely that a non-huge hidden bias would substantially change our conclusions”
 - Measured and incorporated every major known factor that we could identify.
 - Effects on health outcomes were generally large, consistent with earlier work and clinically plausible.

An omitted variable is most likely to change our conclusions about the exposure if it is

- closely related to the outcome,
- seriously imbalanced by exposure,
- uncorrelated with the propensity score.

Does PS Matching Balance “Omitted” Covariates?

No.

- We fit¹⁰ a published PS model to data from the RHC study, using 82 covariates.
- Then we obtained data on 17 other covariates, not included in the PS model.

Corr. with PS	Covariates	Balance Improved	Median Bias Reduction
Sig. ($p < .05$)	10	9 (90%)	45%
Not Sig.	7	2 (29%)	-36%

¹⁰Love et al 2003 abstract

Sensitivity Analysis Approach

- 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 Γ .

The Role of Sensitivity Analysis

- Cameron and Pauling (1976, 1978) concluded Vitamin C increased survival time of colon cancer patients.

The Role of Sensitivity Analysis

- Cameron and Pauling (1976, 1978) concluded Vitamin C increased survival time of colon cancer patients.
- The result was 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 the outcome¹¹.

¹¹ See Rosenbaum 2002, 2017

The Role of Sensitivity Analysis

- Cameron and Pauling (1976, 1978) concluded Vitamin C increased survival time of colon cancer patients.
- The result was 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 the outcome¹¹.
- Sensitivity analysis looks great, yet the findings were contradicted in a Mayo Clinic RCT.

¹¹ See Rosenbaum 2002, 2017

The Role of Sensitivity Analysis

- Cameron and Pauling (1976, 1978) concluded Vitamin C increased survival time of colon cancer patients.
- The result was 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 the outcome¹¹.
- Sensitivity analysis looks great, yet the findings were contradicted in a Mayo Clinic RCT.
- Conclusion: Sensitivity analyses cannot indicate what biases are present, it can only indicate the magnitude needed to alter the conclusion.

¹¹ See Rosenbaum 2002, 2017

Surgery vs. Medicine for Coronary Artery Disease

CAD: Surgery or Medicine

Coronary bypass surgery or medical/drug therapy for CAD¹²?

- 1,515 subjects
 - 590 surgical patients (39%), the rest medical
- PS included 74 observed covariates
 - Hemodynamic, Angiographic, Lab and Exercise Test Results
 - Patient histories, and demographics
- Stratified into PS quintiles, then combined estimates of $\text{Pr}(\text{sustained improvement at 6 mo})$.

¹²See Rosenbaum and Rubin 1983, 1984

Surgery vs. Medicine Study: Results + Sensitivity

Substantial Improvement	Prob (SE)
Medical	0.359 (0.042)
Surgical	0.669 (0.059)

Conclusion: $\text{Pr}(\text{improved} \mid \text{Surgery}) > \text{Pr}(\text{improved} \mid \text{medicine})$

- 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 conclusion¹³.

¹³ See Rosenbaum and Rubin 1983, 1984

Goal of a Formal Sensitivity Analysis

To replace a general qualitative statement that applies in all observational studies . . .

“The association we observe between treatment and outcome does not imply causation.”

“Hidden biases can explain observed associations.”

with a quantitative statement that is specific to what is observed in a particular study

“To explain the association seen in this particular study, one would need a hidden bias of this particular magnitude.”

Hidden Bias

- Two units (patients, subjects, whatever) with the same observed covariates have different chances of receiving the exposure.
- If the association is strong, the hidden bias needed to explain it would be large.
- If a study is free of hidden bias (one example: RCT), this means that any two units that appear similar in terms of their observed covariates actually have the same chance of assignment to exposure.

How would inferences about treatment effects be altered by hidden biases of various magnitudes? - How large would these differences have to be to alter the qualitative conclusions of the study?

The Design Sensitivity Parameter Γ

Γ measures degree of departure from a study that is free of hidden bias. A sensitivity analysis will consider possible values of Γ and show how the inference might change.

- Γ describes the odds ratio comparing the odds of being selected for treatment for two units who are similar on all observed covariates.
 - If $\Gamma = 1$, this means that the study is free of hidden bias
 - Subjects with the same observed covariates have the same odds (hence same probability) of exposure.
 - If $\Gamma = 2$, then two units who appear similar, who have the same set of observed covariates \mathbf{X} , could differ in their odds of receiving the treatment by as much as a factor of 2, so that one could be twice as likely as the other to receive the exposure.

Relating Γ to Sensitivity Statements

A study is **sensitive** if values of Γ close to 1 could lead to inferences that are very different from those obtained assuming the study is free of hidden bias.

- A study is *insensitive* if extreme values of Γ are required to alter the inference.

"To attribute the (observed significant) outcome to an unobserved covariate rather than to the treatment, that unobserved covariate has to increase the odds of treatment by a factor of Γ , and also predict our outcome quite well."

Getting the Message Across

Straightforward, Spreadsheet-Based Formal Sensitivity Analysis for Matched Samples

- Separate tabs for Outcomes: Binary, Continuous, Survival (w/censoring)
- All calculations based on base case formulas using sign-score tests as described in Rosenbaum 2002 [some nuances ignored (dealing with ties, etc.)]
- Available documents provides three examples tested in my course

Demonstration: A Survival Outcome

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

Spreadsheet Demonstration

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

Communicating the Results

In the absence of hidden bias, a sign-score test for matched data with censoring provides strong evidence ($p < .001$) that low (vs. normal) potassium decreases survival time, even after adjustment via PS matching.

To attribute the lower survival time to an unobserved binary covariate unrelated to our propensity model rather than the effect of low potassium, that covariate would need to both

- increase the odds of low potassium more than 2.5-fold and
- be an excellent predictor of mortality.

Summary: Sensitivity Analysis

Hidden bias is the great problem with observational studies, and with PS models.

- Sensitivity analysis after matching can be applied in many scenarios.
- We hope to find that an unobserved covariate would have to be very powerful to alter our conclusions.
- That doesn't mean that such a covariate (or set of them) doesn't exist.

Our PS “Formula” for the Heart Failure papers

- ① Identify available data related to selection for the exposure, and to risk for the outcome.
- ② Demonstrate need for PS modeling (imbalance in key characteristics), and evaluate PS balance after matching, usually through standardized difference plots (usually significance, too, unfortunately.)
- ③ Model exposure effect (Cox models stratified by matched pair identifiers, typically.)
- ④ Formal sensitivity analysis if effect is significant.

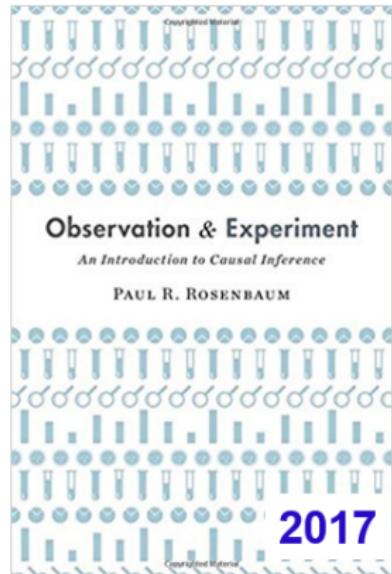
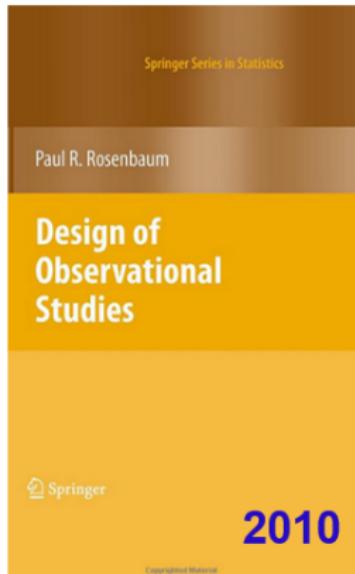
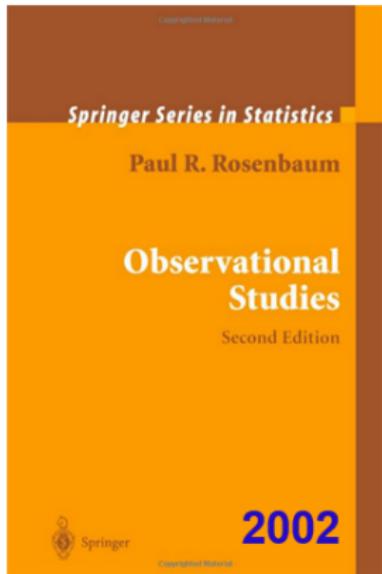
References

All links are posted to the web site.

- Elizabeth Stuart's **Propensity Score software page**
- Noah Griefer's **cobalt vignette**
- RAND corporation's **TWANG package materials**
- Luke Keele's **Rbounds package**
- My student, David Ngendahimana, is addressing the link between covariate balance and improved sensitivity in selecting a matching procedure. We hope to have something arXived and available for testing on github in the Spring. David is at dkn18 at case dot edu.
- My old spreadsheet software is available to you at our github page.

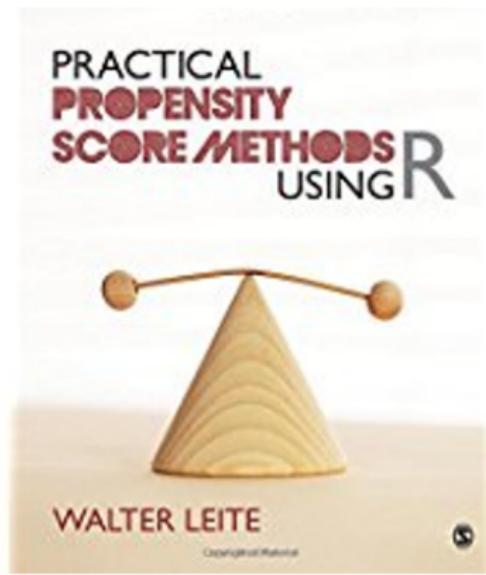
Books I Recommend

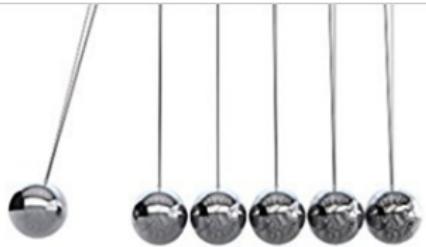
Paul Rosenbaum's Books



Walter Leite
2017
Sage Publications

<http://study.sagepub.com/leite>





CAUSAL INFERENCE IN STATISTICS

A Primer

Judea Pearl

Madelyn Glymour

Nicholas P. Jewell

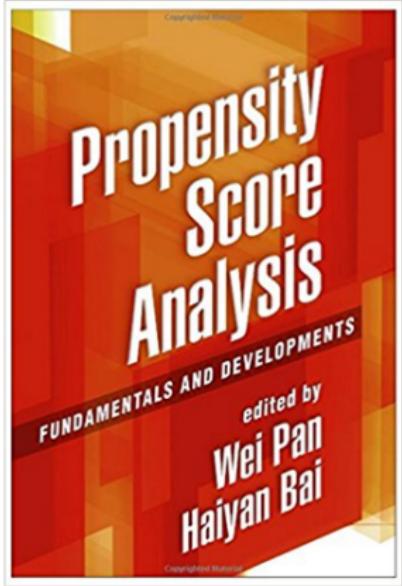


WILEY

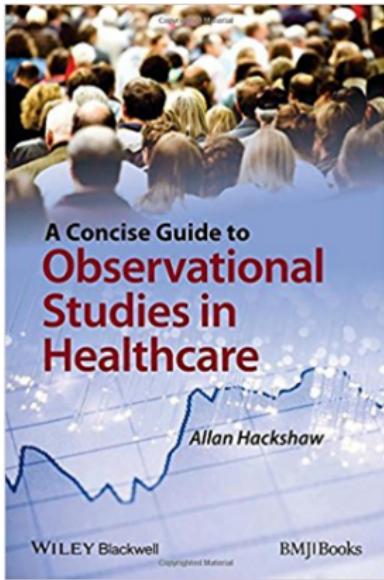
Pearl, Glymour
and Jewell
2016
Wiley

<http://bayes.cs.ucla.edu/PRIMER/>

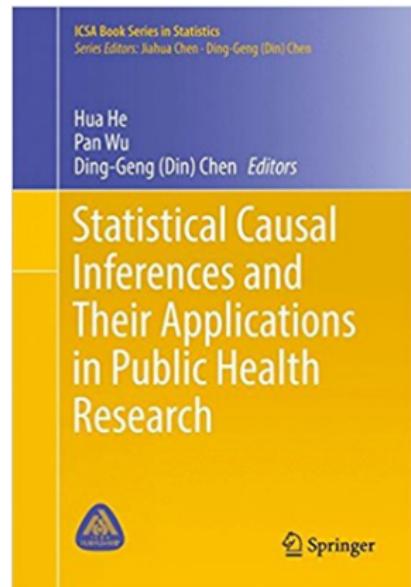
2015



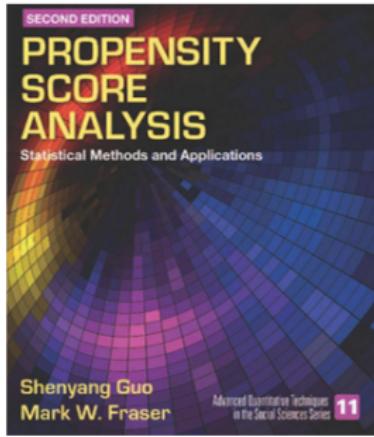
2015



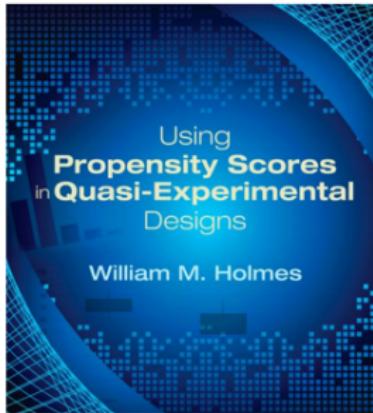
2016



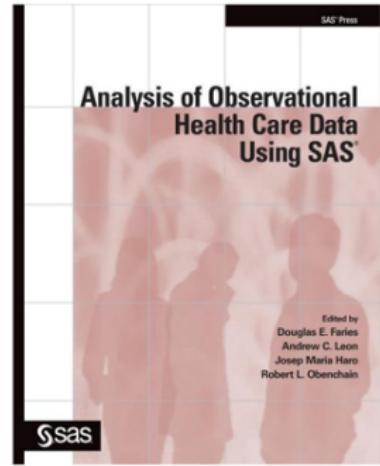
2014



2013



2010



Journal Articles Worth Your Time

- D'Agostino 1998 tutorial
- Austin and Mamdani 2006 case study
- Rubin 2001 HS&ORM (in fact, the entire Dec 2001 issue)
- Stuart 2010 and Ho et al 2007 on matching
- Joffe and Rosenbaum 1999
- Rosenbaum 2015 (on sensitivity analysis)
- King and Nielsen (working paper) argue that the use of propensity score matching to emulate an RCT isn't a good idea, even if using other propensity-score adjustments is a good idea.

Closing Thoughts

A Few Advantages of Propensity Scores

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

Strategic Issues

- How can we make our investigations compelling to our intended audience?
- Why is this hard?
 - Audience is not focused on statistical techniques
 - Audience may have limited training in statistics
- Why is this important?
 - Who makes key policy decisions?
 - Who needs to be convinced by the evidence?

Strategic Issues in Observational Studies

- 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

See Rosenbaum's 2002, 2010 and 2017 books for more.

Some Cautions and Limitations

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

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 intervention effect in light of adjustment
- ⑧ Estimate sensitivity of results to potential hidden bias