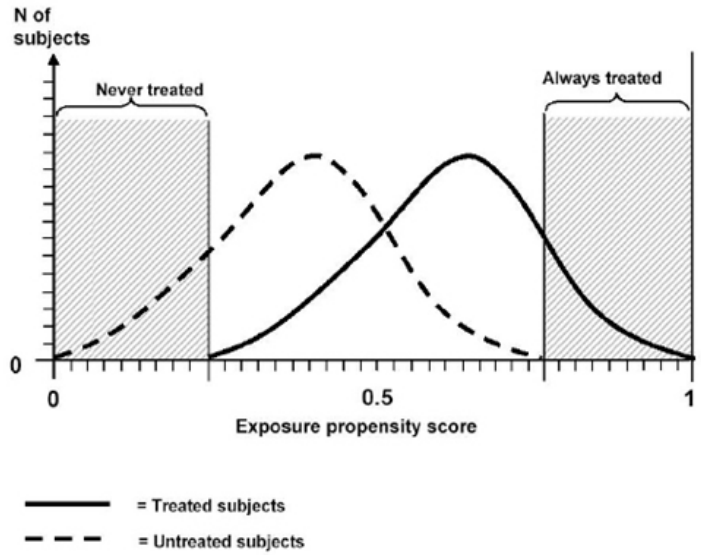


Introduction to Propensity Scores

OCTOBER 19, 2018

SCOTT QUINLAN, PHD

ASSISTANT TEACHING PROFESSOR

DEPARTMENT OF EPIDEMIOLOGY AND BIOSTATISTICS

Outline

The challenges of observational, real-world research

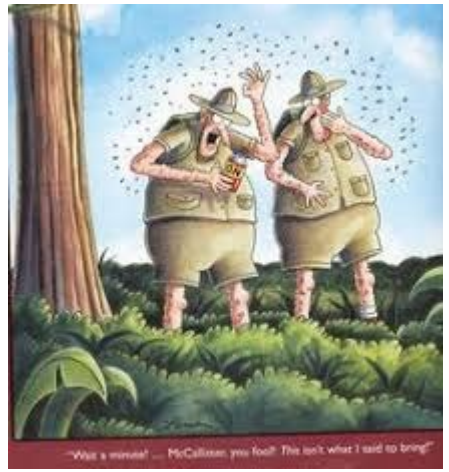
What is a propensity score?

How do we create a propensity score?

How can we use a propensity score?

What do we report when using propensity scores?

The strengths and limitations of propensity scores



Does AA actually work?

The effectiveness of Alcoholics Anonymous (AA) has been studied extensively, with sometimes mixed results.

Randomized trials are challenging.

- AA attendance is freely available and voluntary.

Observational research is an alternative.

- **But..**
- Are those who chose to attend AA similar to those who do not?

	Comparisons prior to stratification or matching			
	AA-attender (n = 336) mean (SD)	AA-nonattender (n = 233) mean (SD)	Standardized difference[b] in %	F-statistics
Demographics				
Male	0.59 (0.49)	0.56 (0.50)	6.93	0.66
Mean age	38.8 (10.1)	36.8 (11.5)	18.4	4.76*
Ethnicity: White	0.63 (0.48)	0.58 (0.49)	9.67	1.29
Black	0.26 (0.44)	0.26 (0.44)	−0.36	0.00
Others	0.11 (0.32)	0.16 (0.37)	−13.3	2.51
Marital: Married	0.34 (0.48)	0.45 (0.50)	−21.4	6.34*
Sep/div/widow	0.36 (0.48)	0.28 (0.45)	16.8	3.84
Single	0.30 (0.46)	0.27 (0.45)	5.72	0.45
Level of education	3.34 (1.02)	3.18 (0.98)	15.7	3.36
Motivation				
Readiness to change index	50.0 (6.7)	46.6 (7.5)	48.3	32.8***
Coercion				
# who pressure you to get treatment	1.85 (1.31)	1.58 (1.16)	22.6	6.88**
# who give you ultimatum	0.58 (0.80)	0.53 (0.72)	5.96	0.48
Problem severity				
ASI composite alcohol score	0.43 (0.32)	0.34 (0.30)	30.8	13.0***
# of dependence symptoms	5.20 (2.77)	3.69 (2.67)	55.8	42.5***
# of alcohol-rel. consequences	1.42 (1.42)	0.99 (1.15)	33.4	14.8***
Help-seeking				
# of AA meetings last year	36.6 (62.9)	8.08 (27.2)	58.8	42.3***
# of treatment episodes last year	4.24 (18.2)	1.96 (14.6)	13.9	2.55
Type of treatment				
Private	0.29 (0.46)	0.18 (0.38)	28.2	10.60**
HMO	0.33 (0.47)	0.67 (0.47)	−72.0	71.3***
Public	0.38 (0.48)	0.15 (0.36)	51.5	34.7***
Social influences				
Size of support network:				
# to talk to	4.25 (5.25)	3.62 (4.32)	13.1	2.29
# can get help from	4.15 (4.86)	4.15 (4.74)	0.01	0.00
# in regular contact with	5.61 (5.59)	6.20 (5.39)	−10.8	1.60
Drinking of network				
# of heavy or problem drinkers	0.85 (2.17)	0.86 (2.03)	−0.20	0.00
Prop. heavy/problem drinkers	0.13 (0.25)	0.15 (0.25)	−6.53	0.59
# who encourage you to drink	0.39 (2.37)	0.25 (1.20)	7.27	0.66
Prop. who encourage you to drink	0.038 (0.15)	0.046 (0.17)	−4.72	0.31

Drug and Alcohol Dependence 104 (2009) 56–64

Best way to have your appendix removed?

The choice between laparoscopic and open-wound appendectomy is often made based on patient characteristics and illness severity.

Can we then fairly compare outcomes?

Arch Surg. 2010;145(10):939-945

Table 2. Patient Characteristics

Patient Characteristic	Aggregate Cohort		*P* Value
	Open Appendectomy	Laparoscopic Appendectomy	
No. (%)	6030 (28)	15 445 (72)	
Age, mean (SD), y	41 (17)	38 (16)	<.001
Female, No. (%)	2551 (42)	7458 (48)	<.001
Nonwhite race, No. (%)	2306 (38)	5451 (35)	<.001
ASA class, No. (%)			
1-2	5139 (85)	14 005 (91)	<.001
3-5	891 (15)	1440 (9)	
Emergency surgery, No. (%)	4938 (82)	11 884 (77)	<.001
Wound class, No. (%)			
Clean-contaminated	2198 (36.8)	6036 (39.7)	<.001
Contaminated	1742 (29.2)	6467 (42.5)	
Dirty/infected	2031 (34.0)	2707 (17.8)	
Evidence of rupture (*CPT* code 44960 or *ICD-9-CM* codes 540.0 and 540.1), No. (%)	1976 (33)	2136 (13.8)	<.001
Selected comorbid risk factors, No. (%)			
No diabetes	5703 (94.6)	14 833 (96.0)	<.001
Current smoker	1319 (21.9)	3426 (22.2)	.60
Ethanol use	187 (3.1)	302 (2.0)	<.001
No dyspnea	5883 (97.6)	15 216 (98.5)	<.001
DNR	30 (0.5)	40 (0.3)	.006
Independent functional status	5851 (97.0)	15 194 (98.4)	<.001
History of severe COPD	84 (1.4)	125 (0.8)	<.001
Ascites within 30 d	124 (2.1)	282 (1.8)	.30
History of MI	23 (0.4)	20 (0.1)	<.001
Hypertension	1187 (19.7)	2222 (14.4)	<.001
Acute renal failure	17 (0.3)	13 (0.08)	<.001
Currently undergoing dialysis	30 (0.5)	25 (0.2)	<.001
Sepsis			
SIRS	2188 (36.3)	5178 (33.6)	<.001
Sepsis	187 (3.1)	205 (1.3)	
Septic shock	31 (0.5)	21 (0.1)	
Pregnancy	76 (1.3)	144 (0.9)	.007

Does acupuncture work?

We want to see how well acupuncture works in people with chronic pain, but…

Those who choose acupuncture are often very different from those who do not.

	Started Acupuncture ($n = 952$)	Did Not Start Acupun ($n = 59,564$)
Propensity score characteristics[b]		
Opioid therapy plan	28.8%	17.8%
Physical therapy past 30 days	16.3%	15.1%
Physical therapy past 31–180 days	25.0%	11.1%
Physical therapy past 181–365 days	24.5%	12.1%
Nonspecific chronic pain	29.6%	14.4%
Substance abuse	4.6%	4.1%
Sleep problem	23.6%	14.6%
History of tobacco use	14.2%	12.9%
Anxiety	23.7%	15.6%
Pain treatment procedure	38.2%	22.5%
Pain diagnosis procedure	65.3%	52.5%
Pain medication	81.2%	65.0%
Age (years)	53.8 (14.0)	55.2 (15.0)
Number of outpatient visits	15.9 (10.8)	10.4 (10.1)
Months since cohort entry	29.1 (14.7)	25.2 (15.6)
Ambulatory Charlson score	1.8 (2.2)	1.9 (2.1)
Demographic Characteristics		
Female	72. 8%	62.0%
White	91.2%	91.9%
Hispanic	5.4%	7.7%
Medical and Psychiatric Comorbidities		
Depression	21.5%	15.8%

BMC Medical Research Methodology (2017) 17:42

The challenge

Randomized trials are the **gold standard** for comparing two different therapies, interventions, surgeries, etc.

- But, they may not be practical or feasible in all settings.

Observational studies are an alternative, but **exposure selection process** can lead to bias.

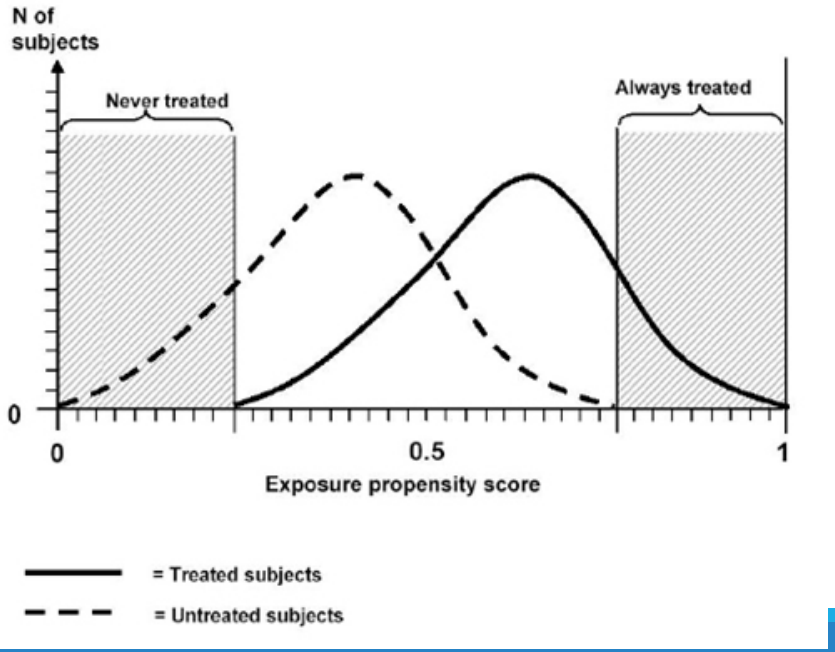
- Those exposed (i.e. treated) are sometimes quite different from those not exposed.

What can we do here?

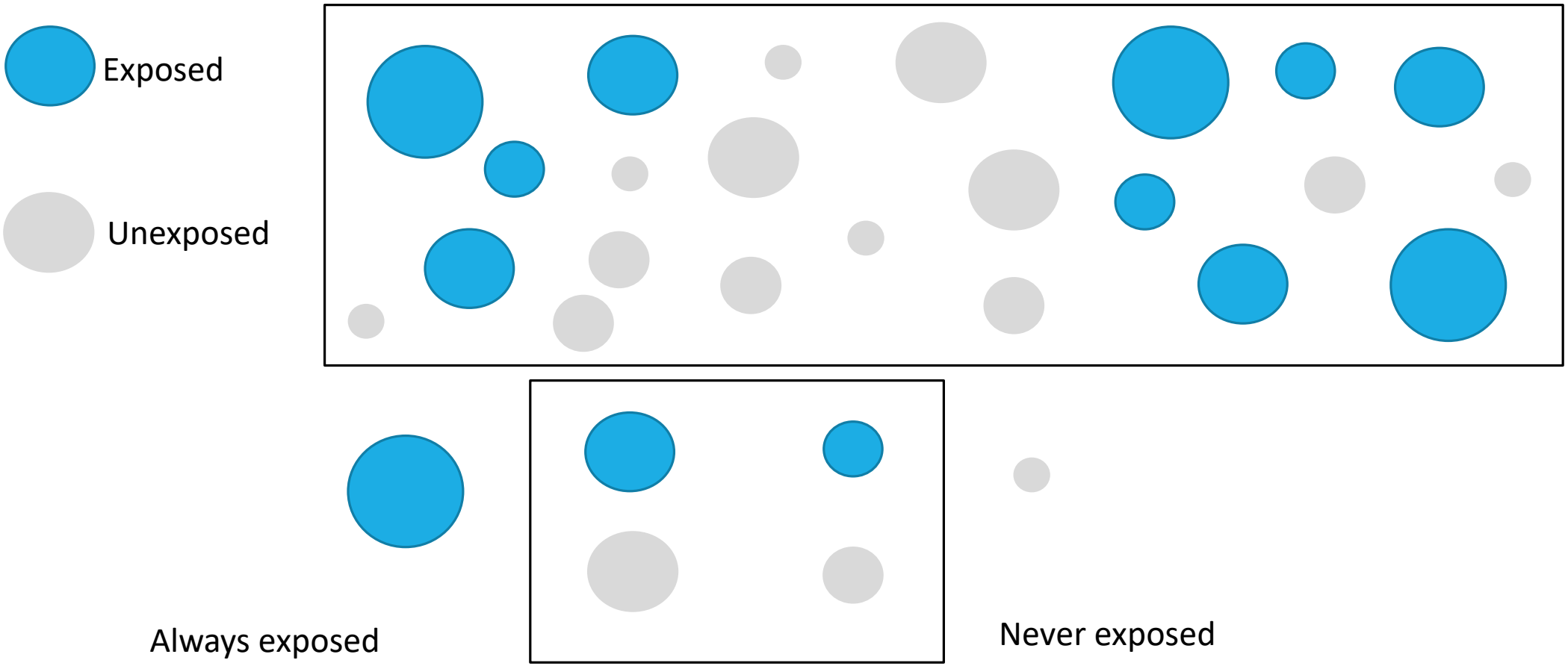
The basic idea

We measure (or have information on) a number of characteristics for each person for the time period **before** someone is exposed.

We use this information to create a model that **predicts the probability** of receiving the exposure (compared to an alternative of interest).



Propensity score methods



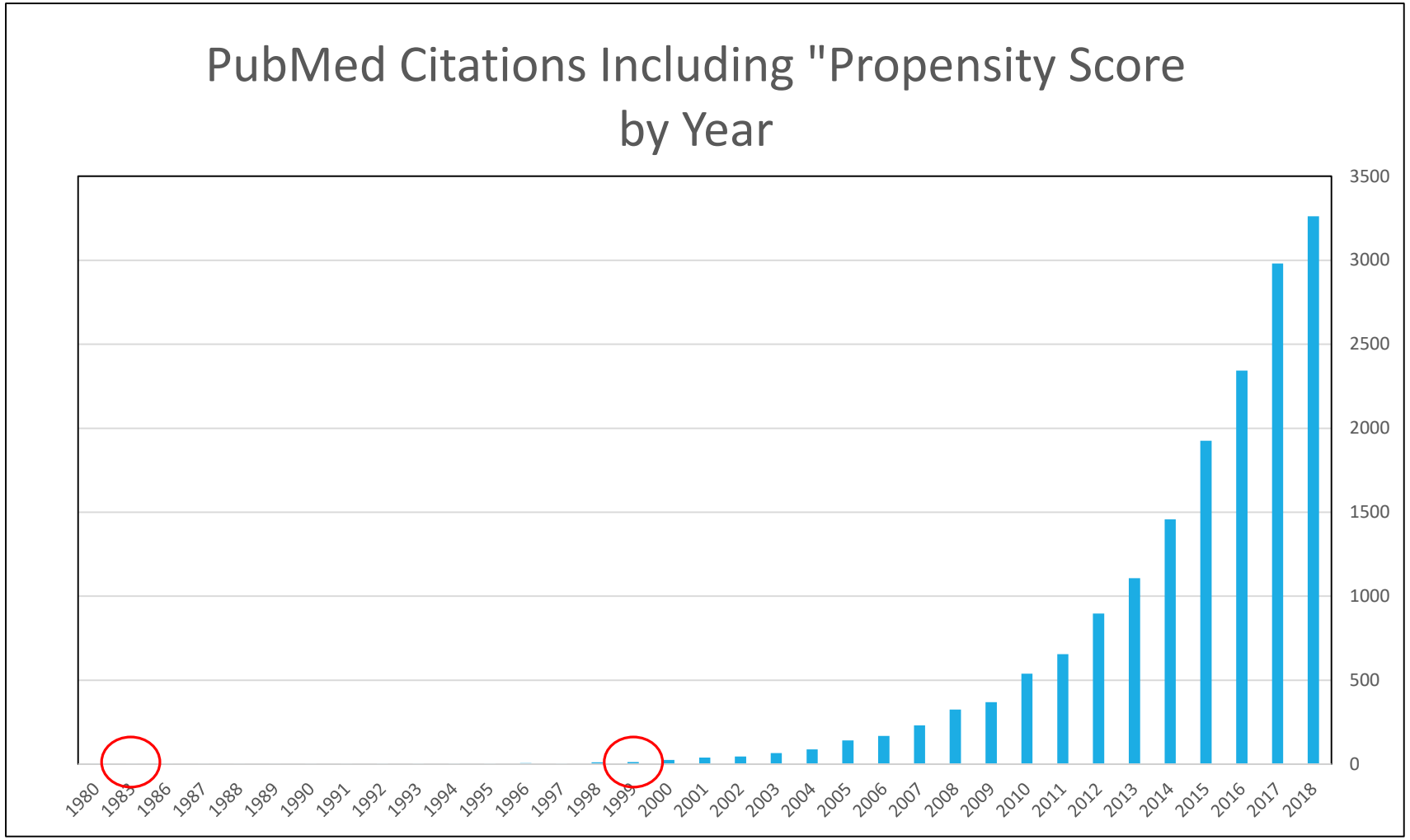
What is a propensity score?

A **probability** of being exposed (treated, vaccinated, etc.) based on characteristics that are present **before** exposure occurs.

Each person in our study is assigned a score that ranges from 0 (never exposed) to 1 (always exposed).

The score can then be used to do a number of things:

- Matching
- Stratification
- Adjustment
- Weighting


PubMed Citations Including "Propensity Score by Year
3500
3000
2500
2000
1500
1000
500
0
1980
1983
1986
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018

Where do we get propensity scores?

Recall that these are just **probabilities** of being exposed, given a person's characteristics.

Logistic regression is most commonly used.

$$Logit\ (probability\ exposed) = Characteristics\ \boldsymbol{BEFORE}\ exposure$$

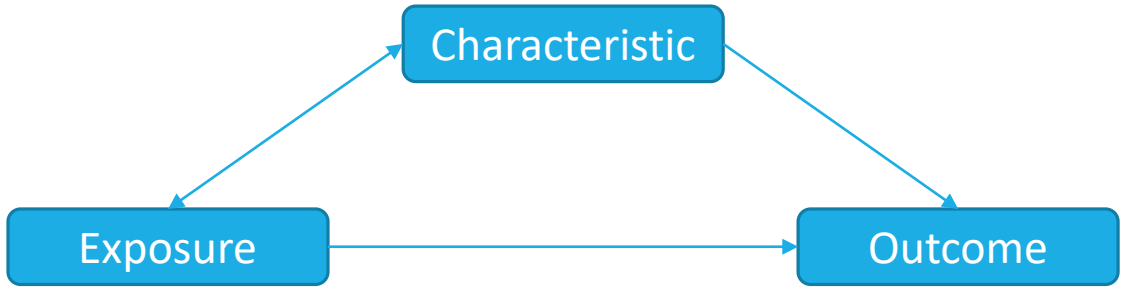
More complex methods are also being studied.

- Neural networks
- Machine learning
- Boosting methods

Model selection

Three different characteristics to consider:

- Those related to the **outcome only** YES!
- Those related to the **exposure only** NO!
- Those related to both the **outcome and the exposure** YES!



Model selection

Model selection techniques **not as effective** here.

Parsimonious not as important as **thoroughness**.

Statistical significance not as much of a concern.

Multicollinearity not as much of a concern.

Balance is our goal!

Remember our goal

One measure of the quality of a logistic regression model is the **c-statistic**, values closer to 1.0 indicate better discriminatory ability.

- How well can the model predict the probability of an outcome?

Our goal is to create **balanced** groups to allow for a fair comparison.

- The c-statistic (and related measures) are of secondary importance here.

Example: People take statin medications to control their cholesterol levels. People who do not do well on a statin medication alone (such as simvastatin) may have other therapies added on (such as ezetimibe). But, this decision is driven by LDL levels, such that (for example):

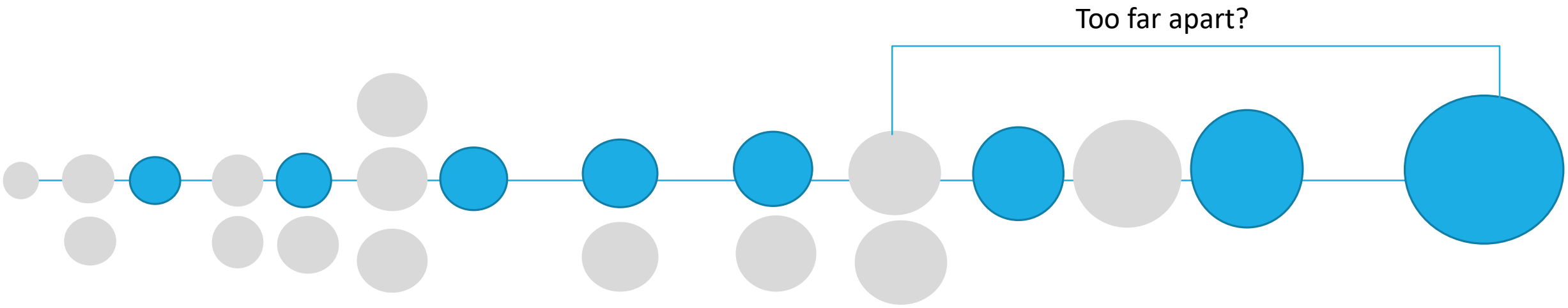
- LDL > 180 → prescribe combination therapy
- LDL ≤180 → stick with simvastatin alone
- If we know LDL, we can likely predict exposure almost perfectly, but is this what we want?

Now that I have a propensity score, what can I do with it?

There are several approaches to consider:

- Matching
- Stratification
- Adjustment
- Inverse probability of treatment weights (IPTW)

Matching

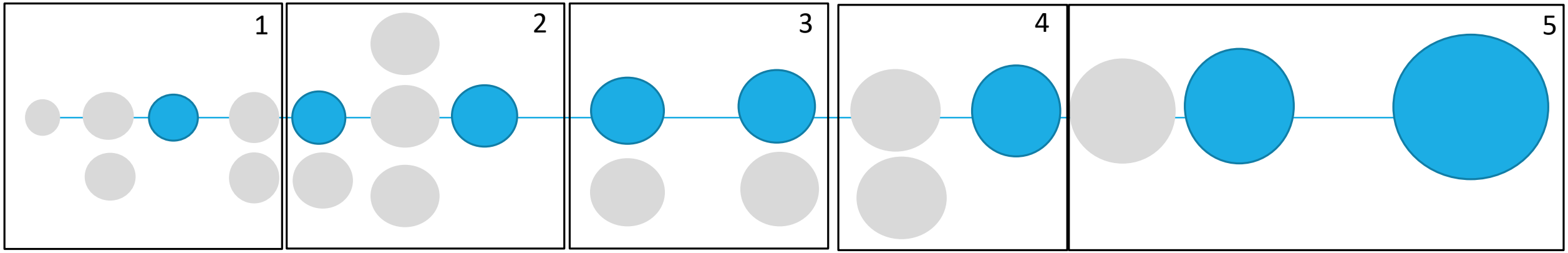


With our without replacement.

Greedy versus **optimal**.

Apply **caliper** requirement?

Stratification



Choose number of strata, but 5 is usually most common.

Analyze **within strata** and then pool.

Be wary of: Imbalances, residual confounding, effect modification

Consistent effect?

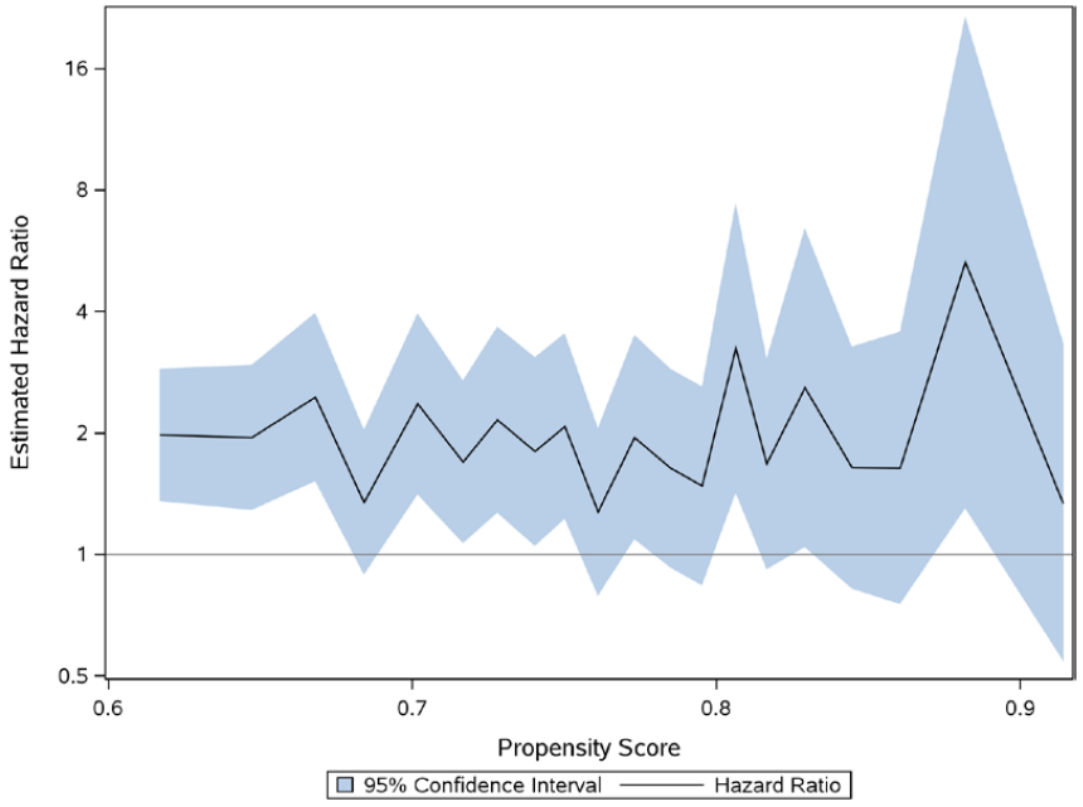


Figure 2. Estimated treatment effects and 95% confidence interval within deciles of the estimated propensity scores.

***Circ Cardiovasc Qual Outcomes*. 2013;6:604-611**

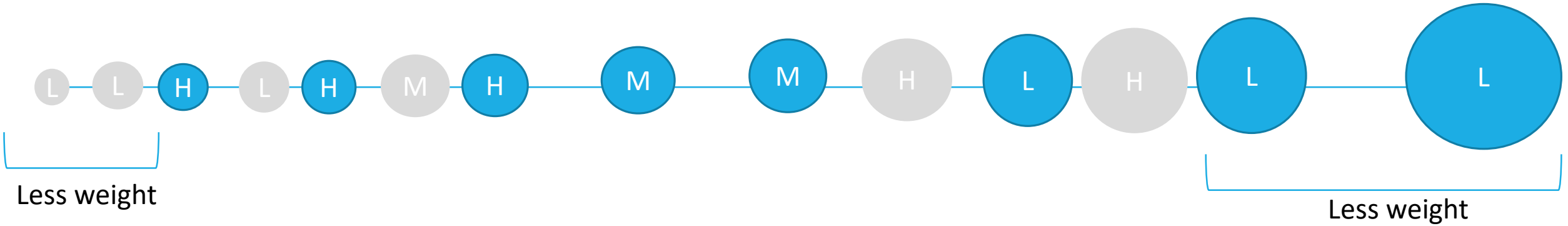
Inverse probability of treatment weighting (IPTW)

For each person in the original sample we assign a weight based on the**inverse** probability of the treatment (or exposure) received.

$$Weight_i = \boxed{\frac{z_i}{p_i}} + \boxed{\frac{1-z_i}{1-p_i}}$$

Exposed Unexposed

$$z_i = treatment\ (1 = yes, 0 = no)$$
$$p_i = probability\ of\ treatment$$



Regression adjustment

One last technique is to use the propensity score in the analysis phase as an **adjustment**, just like we would for any covariate of interest.

$$Y(Outcome) \quad = Exposure + Covariates + Propensity\ Score$$

Used quite frequently, but has limitations:

- Need to understand the relationship between propensity score and outcome!

How well did it work?

Our goal is to create two **balanced groups**, one exposed and the other not exposed.

Before moving to our analysis, we should consider how well the process worked.

Assessing balance:

- **P-values are discouraged** since they are impacted by the difference between groups AND sample size.
- **Plots** can be a helpful starting point.
- The **standardized difference** is the preferred method of assessing balance.
- More complex (less intuitive) methods.

Patient Characteristic	Aggregate Cohort			Propensity-Matched Cohort		
	Open Appendectomy	Laparoscopic Appendectomy	*P* Value	Open Appendectomy	Laparoscopic Appendectomy	*P* Value
No. (%)	6030 (28)	15 445 (72)		5666 (50)	5666 (50)	
Age, mean (SD), y	41 (17)	38 (16)	<.001	40.1 (16.8)	41.4 (17.2)	<.001
Female, No. (%)	2551 (42)	7458 (48)	<.001	2425 (43)	2495 (44)	.20
Nonwhite race, No. (%)	2306 (38)	5451 (35)	<.001	2123 (37)	2132 (38)	.90
ASA class, No. (%)						
1-2	5139 (85)	14 005 (91)	<.001	4944 (87)	4854 (86)	.02
3-5	891 (15)	1440 (9)		722 (13)	812 (14)	
Emergency surgery, No. (%)	4938 (82)	11 884 (77)	<.001	4597 (81)	4571 (81)	.50
Wound class, No. (%)						
Clean-contaminated	2198 (36.8)	6036 (39.7)	<.001	2207 (39.0)	2163 (38.2)	.40
Contaminated	1742 (29.2)	6467 (42.5)		1746 (30.8)	1706 (30.1)	
Dirty/infected	2031 (34.0)	2707 (17.8)		1711 (30.2)	1794 (31.7)	
Evidence of rupture (*CPT* code 44960 or *ICD-9-CM* codes 540.0 and 540.1), No. (%)	1976 (33)	2136 (13.8)	<.001	1628 (29)	1733 (31)	.03
Selected comorbid risk factors, No. (%)						
No diabetes	5703 (94.6)	14 833 (96.0)	<.001	5399 (95.3)	5345 (94.3)	.03
Current smoker	1319 (21.9)	3426 (22.2)	.60	1247 (22.0)	1347 (23.8)	.03
Ethanol use	187 (3.1)	302 (2.0)	<.001	158 (2.8)	182 (3.2)	.20
No dyspnea	5883 (97.6)	15 216 (98.5)	<.001	5541 (97.8)	5540 (97.8)	.90
DNR	30 (0.5)	40 (0.3)	.006	25 (0.4)	24 (0.4)	.90
Independent functional status	5851 (97.0)	15 194 (98.4)	<.001	5535 (97.7)	5519 (97.4)	.40
History of severe COPD	84 (1.4)	125 (0.8)	<.001	67 (1.2)	73 (1.3)	.60
Ascites within 30 d	124 (2.1)	282 (1.8)	.30	105 (1.9)	122 (2.2)	.30
History of MI	23 (0.4)	20 (0.1)	<.001	16 (0.3)	16 (0.3)	>.99
Hypertension	1187 (19.7)	2222 (14.4)	<.001	1012 (17.9)	1189 (21.0)	<.001
Acute renal failure	17 (0.3)	13 (0.08)	<.001	8 (0.1)	8 (0.1)	>.99
Currently undergoing dialysis	30 (0.5)	25 (0.2)	<.001	15 (0.3)	14 (0.3)	.90
Sepsis						
SIRS	2188 (36.3)	5178 (33.6)	<.001	2037 (36.0)	2078 (36.7)	.80
Sepsis	187 (3.1)	205 (1.3)		138 (2.4)	145 (2.6)	
Septic shock	31 (0.5)	21 (0.1)		14 (0.3)	14 (0.3)	
Pregnancy	76 (1.3)	144 (0.9)	.007	14 (0.3)	13 (0.2)	.80

Plots

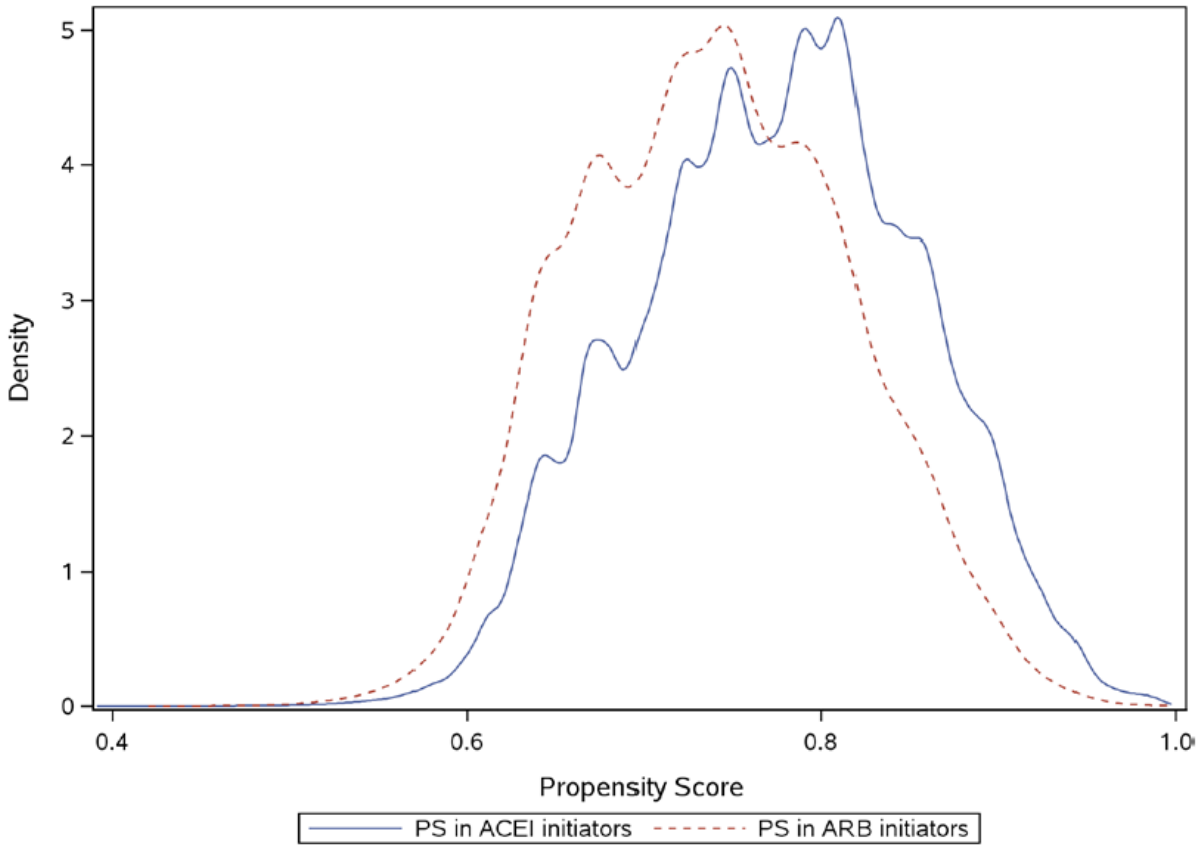
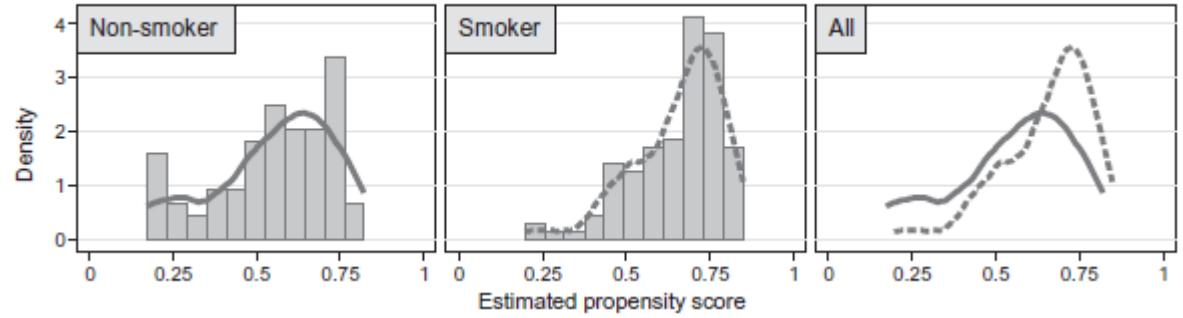


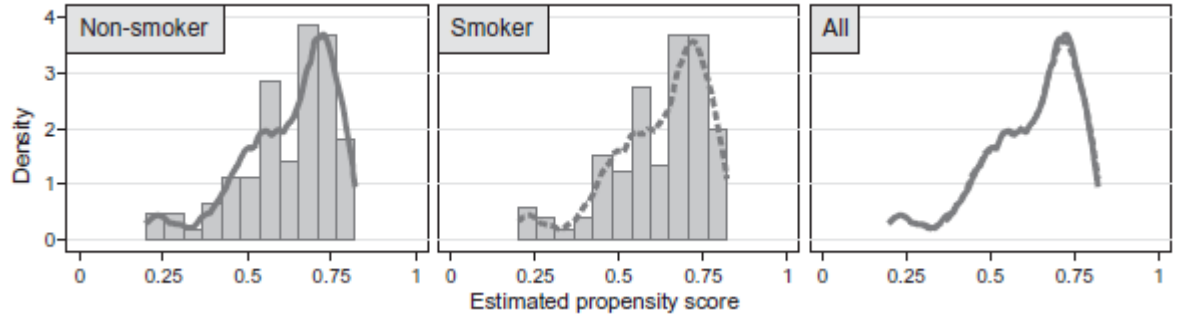
Figure 1. Estimated density of the propensity scores (PSs) among new users of angiotensin-converting enzyme inhibitors (ACEis) and angiotensin receptor blockers (ARBs).

***Circ Cardiovasc Qual Outcomes*. 2013;6:604-611**

Our hope

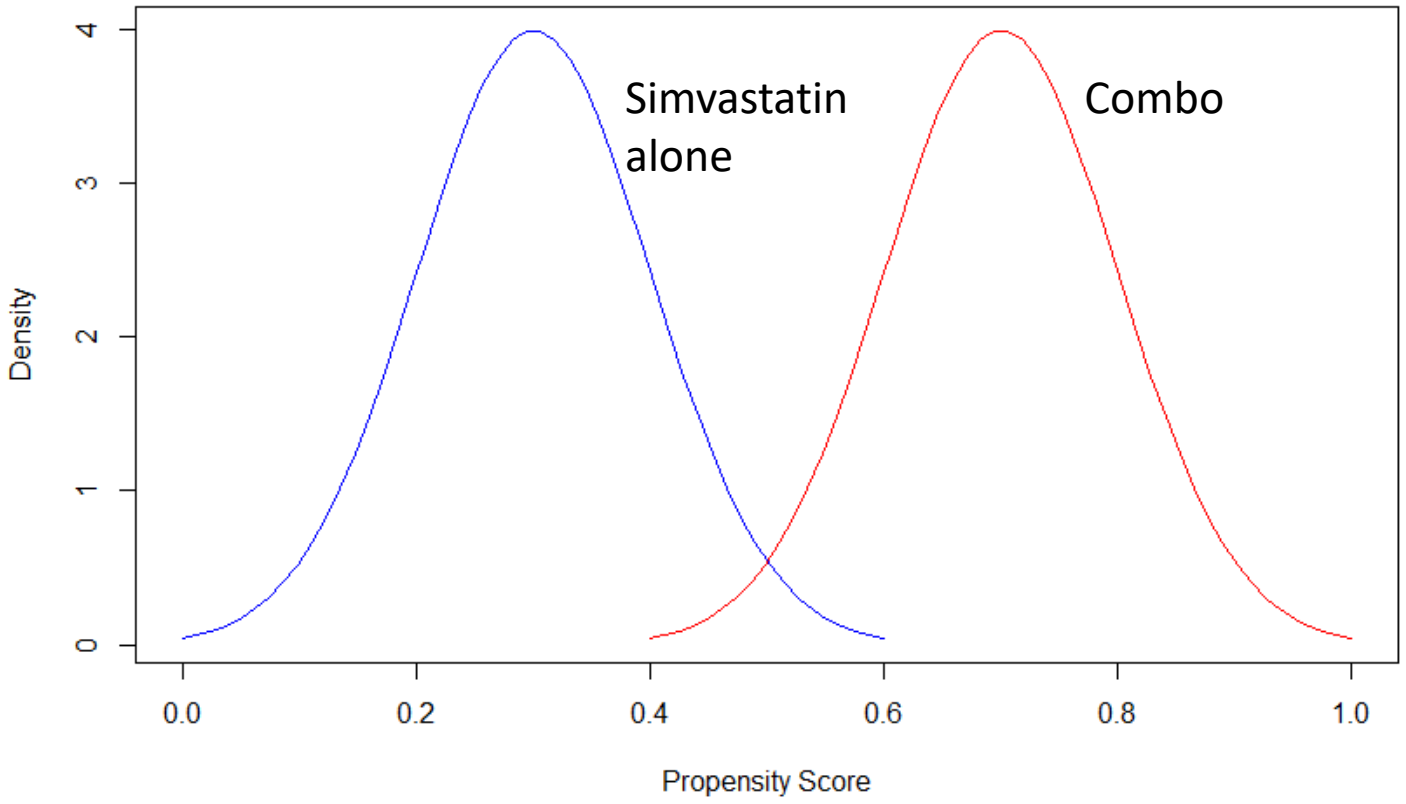


Before



After

To add ezetimibe or not?



Is this the right comparison group?

People with manageable cholesterol are very unlikely to receive combination therapy.

Using the standardized difference

Measures the difference between the two groups in terms of **standard deviations**.

Does not depend on sample size.

$$d = \frac{(\overline{x}_{treatment} - \overline{x}_{control})}{\sqrt{\frac{s^2_{treatment} + s^2_{control}}{2}}}$$

Continuous covariates

$$d = \frac{(\hat{p}_{treatment} - \hat{p}_{control})}{\sqrt{\frac{\hat{p}_{treatment}(1 - \hat{p}_{treatment}) + \hat{p}_{control}(1 - \hat{p}_{control})}{2}}}$$

Categorical covariates

A standardized difference of **0.10 (or 10%) or lower** is considered good balance.

	AA-attender (n = 336) mean (SD)	AA-nonattender (n = 233) mean (SD)	Standardized difference[b] in %	Standardized difference[b] in %
Demographics				
Male	0.59 (0.49)	0.56 (0.50)	6.93	0.00
Mean age	38.8 (10.1)	36.8 (11.5)	18.4	−11.1
Ethnicity: White	0.63 (0.48)	0.58 (0.49)	9.67	4.32
Black	0.26 (0.44)	0.26 (0.44)	−0.36	−4.02
Others	0.11 (0.32)	0.16 (0.37)	−13.3	−1.03
Marital: Married	0.34 (0.48)	0.45 (0.50)	−21.4	−0.72
Sep/div/widow	0.36 (0.48)	0.28 (0.45)	16.8	−9.09
Single	0.30 (0.46)	0.27 (0.45)	5.72	−10.1
Level of education	3.34 (1.02)	3.18 (0.98)	15.7	−1.40
Motivation				
Readiness to change index	50.0 (6.7)	46.6 (7.5)	48.3	−11.5
Coercion				
# who pressure you to get treatment	1.85 (1.31)	1.58 (1.16)	22.6	16.8
# who give you ultimatum	0.58 (0.80)	0.53 (0.72)	5.96	0.93
Problem severity				
ASI composite alcohol score	0.43 (0.32)	0.34 (0.30)	30.8	6.03
# of dependence symptoms	5.20 (2.77)	3.69 (2.67)	55.8	9.84
# of alcohol-rel. consequences	1.42 (1.42)	0.99 (1.15)	33.4	−3.81
Help-seeking				
# of AA meetings last year	36.6 (62.9)	8.08 (27.2)	58.8	8.15
# of treatment episodes last year	4.24 (18.2)	1.96 (14.6)	13.9	4.00
Type of treatment				
Private	0.29 (0.46)	0.18 (0.38)	28.2	2.51
HMO	0.33 (0.47)	0.67 (0.47)	−72.0	−5.98
Public	0.38 (0.48)	0.15 (0.36)	51.5	4.11
Social influences				
Size of support network:				
# to talk to	4.25 (5.25)	3.62 (4.32)	13.1	−8.34
# can get help from	4.15 (4.86)	4.15 (4.74)	0.01	−8.95
# in regular contact with	5.61 (5.59)	6.20 (5.39)	−10.8	1.92
Drinking of network				
# of heavy or problem drinkers	0.85 (2.17)	0.86 (2.03)	−0.20	6.71
Prop. heavy/problem drinkers	0.13 (0.25)	0.15 (0.25)	−6.53	7.46
# who encourage you to drink	0.39 (2.37)	0.25 (1.20)	7.27	9.36
Prop. who encourage you to drink	0.038 (0.15)	0.046 (0.17)	−4.72	8.55

HINT: Look for values more than 0.10 (or 10%) in absolute terms

Drug and Alcohol Dependence 104 (2009) 56–64

What to report?

Original pools of exposed and unexposed.

Sample size before and after matching.

The model used to create the propensity scores.

The algorithm used to match.

Diagnostics of match quality.

Information on those who did not match.

Summarize how the propensity score was determined

Variables for inclusion in the propensity models were chosen based on *a priori* considerations of clinical significance (i.e., might strongly predict a cardiovascular event or indicate underlying disease severity) and from an exploratory analysis of the 100 most common diagnoses, procedures, and outpatient prescription medications, including antidiabetic therapy and medications used to prevent or treat CHD, such as ACE inhibitors, beta blockers, statins and others, and dispensed in the 6-month baseline period. The claims data did not contain information on over-the-counter medications, such as aspirin, or those dispensed in a hospital setting. Variables were first

Pharmacoepidemiology and Drug Safety 2007; 16: 504–512

	1999			
	Matched		Unmatched	
	TZD	M+S	TZD	M+S
Number of subjects	1270	1270	648	648
Age (years)				
18–29	1.9	1.7	1.5	1.3
30–39	8.9	9.0	9.4	7.2
40–49	25.8	24.7	22.4	30.9
50–64	55.4	57.3	59.7	53.6
65+	8.0	7.2	6.9	7.0
Male (%)	53.7	52.0	48.0	57.3
Using insulin (%)	10.9	10.4	68.8	0.1
With any oral antibiotic drug (%)	69.5	69.7	47.1	86.5
Health utilization parameter (Mean)*				
Total health costs†	$2061	$2141	$3928	$2185
Drug costs‡	$522	$527	$1012	$403
No. of glycated hemoglobin tests 1-month prior to initiation	0.39	0.38	0.37	0.38
No. of glycated hemoglobin tests 2–6 months prior to initiation	0.59	0.57	0.86	0.47
No. of hypoglycemic episodes	0.06	0.05	0.15	0.04
No. of diagnosis codes	7.47	7.40	9.88	6.65
No. of pathology/laboratory codes (80048–88299)	7.58	7.81	10.78	6.98
No. of ER visits	0.17	0.16	0.22	0.19
No. of inpatient stays	0.10	0.09	0.13	0.12
No. of ambulatory visits	9.31	9.37	13.66	7.98
No. of drugs dispensed	6.46	6.37	10.12	5.56
No. of cardiovascular inpatient stays	0.08	0.07	0.12	0.09
No. of cardiovascular ambulatory visits	1.86	1.73	2.56	1.79
No. of cardiovascular drugs dispensed	1.02	0.95	1.36	0.99
Duration of health plan enrollment before initiation date (days)	302.98	301.34	311.00	288.14

Unmatched can tell us something too

How many were left unmatched?

How do they differ from those who matched?

Who are we analyzing? Not analyzing?

Pharmacoepidemiology and Drug Safety 2007; 16: 504–512

Advantage over traditional regression

Separates the design and the analysis phases.

Simpler to **determine balance** (or lack thereof).

More **flexible** when the outcome is rare and the exposure is common.

Consider **discontinuing** if there is no overlap.

Challenges

Propensity score methods can only account for **measured** characteristics.

Tradeoff between **closeness** of match/strata and **sample size**.

Missing data issues.

Rare exposures.

Propensity score is **specific to the outcome**.

Learn more

Austin PC. An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. *Comparative Behavioral Research* 2011; 46: 399-424

Brookhart MA, et al. Propensity score methods for confounding control in nonexperimental research. *Circ Cardiovasc Qual Outcomes*. 2013;6:604-611

Thank you!

SQUINLAN@GWU.EDU