Comparing Apples and Oranges
An Introduction to Propensity Score Analysis

Liang Li, PhD

Associate Professor

Department of Biostatistics

Biostatistics Grand Rounds at MDACC, 1/2015

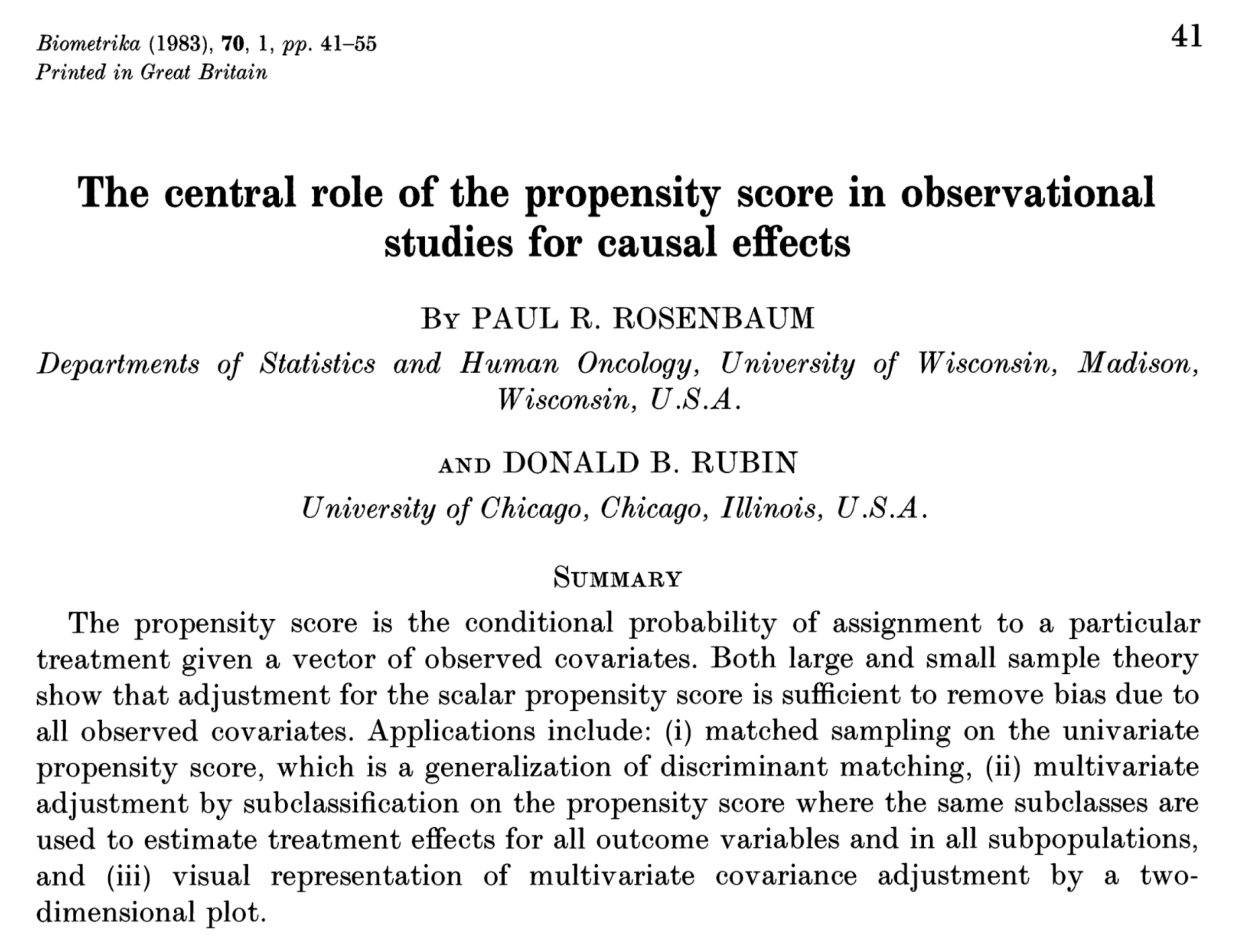
Comparing Medical Treatments Using Randomized Controlled Trials (RCT)

- Examples:
  - ✓ intraoperative transfusion vs. no transfusion on post-operative morbidity
  - ✓ Mastectomy vs. breast conserving therapy (BCT) vs. breast conserving surgery (BCS) on survival
  - ✓ Rank hospitals on quality and performance metrics
- The gold standard for causal inference of treatment effect
- Comparing apples to apples

Comparing Medical Treatments Using Observational Data

- However, randomized controlled trials have limitations:
  - ✓ Infeasible (coaching; drinking wine; smoking)
  - ✓ Costly and time-consuming
  - ✓ Limited external validity and patient heterogeneity
- Majority of published evidence on the effect of medical treatments relies on nonrandomized studies (observational data)
- Observational studies become more common with technology development (Electronic Medical Records)
- More difficult to draw causal inference
- Comparing apples to oranges
- How to turn an “apples to oranges” comparison into an “apples to apples” comparison?

Propensity Score Analysis



Biometrika (1983), **70**, 1, *pp.* 41–55
Printed in Great Britain

The central role of the propensity score in observational studies for causal effects

By PAUL R. ROSENBAUM

Departments of Statistics and Human Oncology, University of Wisconsin, Madison, Wisconsin, U.S.A.

and DONALD B. RUBIN

University of Chicago, Chicago, Illinois, U.S.A.

Summary

The propensity score is the conditional probability of assignment to a particular treatment given a vector of observed covariates. Both large and small sample theory show that adjustment for the scalar propensity score is sufficient to remove bias due to all observed covariates. Applications include: (i) matched sampling on the univariate propensity score, which is a generalization of discriminant matching, (ii) multivariate adjustment by subclassification on the propensity score where the same subclasses are used to estimate treatment effects for all outcome variables and in all subpopulations, and (iii) visual representation of multivariate covariance adjustment by a two-dimensional plot.

The Propensity Score

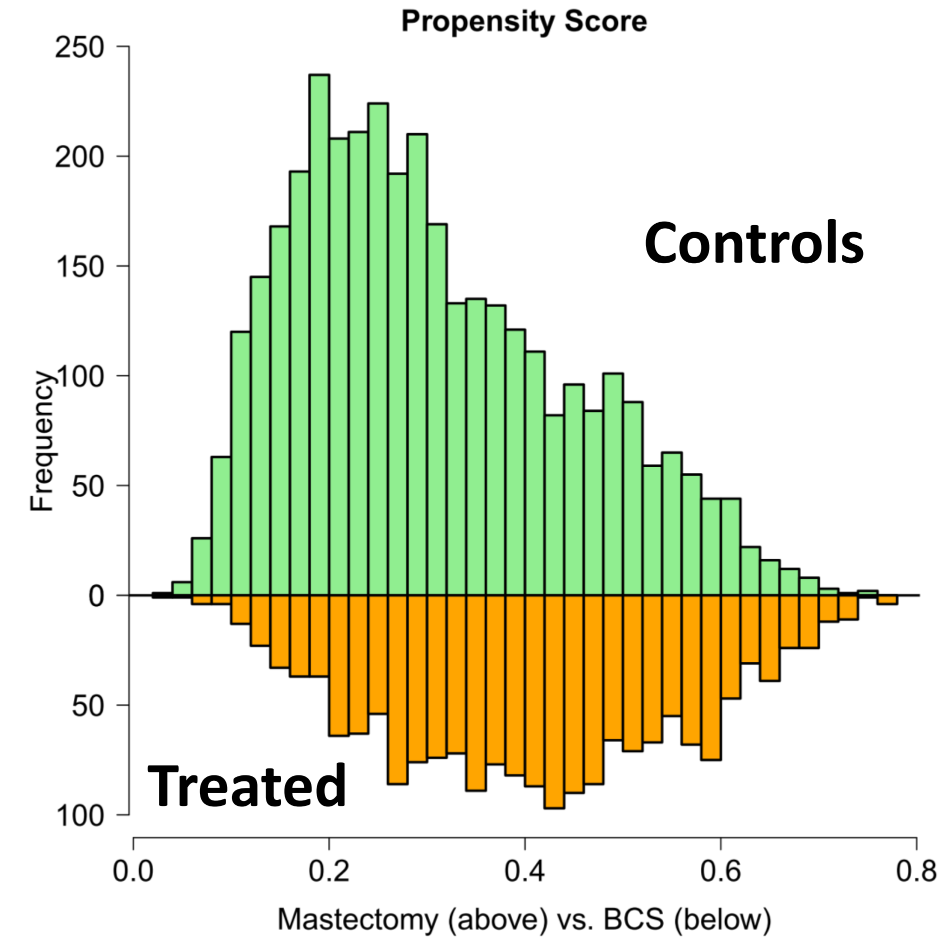
- It is impossible to matching a treated patient to a control on all covariates; not even in randomized controlled trials
- We only need to ensure similarity at group level
- The propensity score is the probability of being assigned to the treated group (vs. control) given the covariates; it can be calculated conveniently from a logistic regression (a working model)

Probability(Treated) ~ Covariate 1 + Covariate 2 + … …

- The covariates must include all that we want to balance; bias may arise if there is an unmeasured variable that may be unbalanced and affects the outcome (a confounder)
- If there is no unmeasured confounder, the propensity score has the balancing property

The Balancing Property

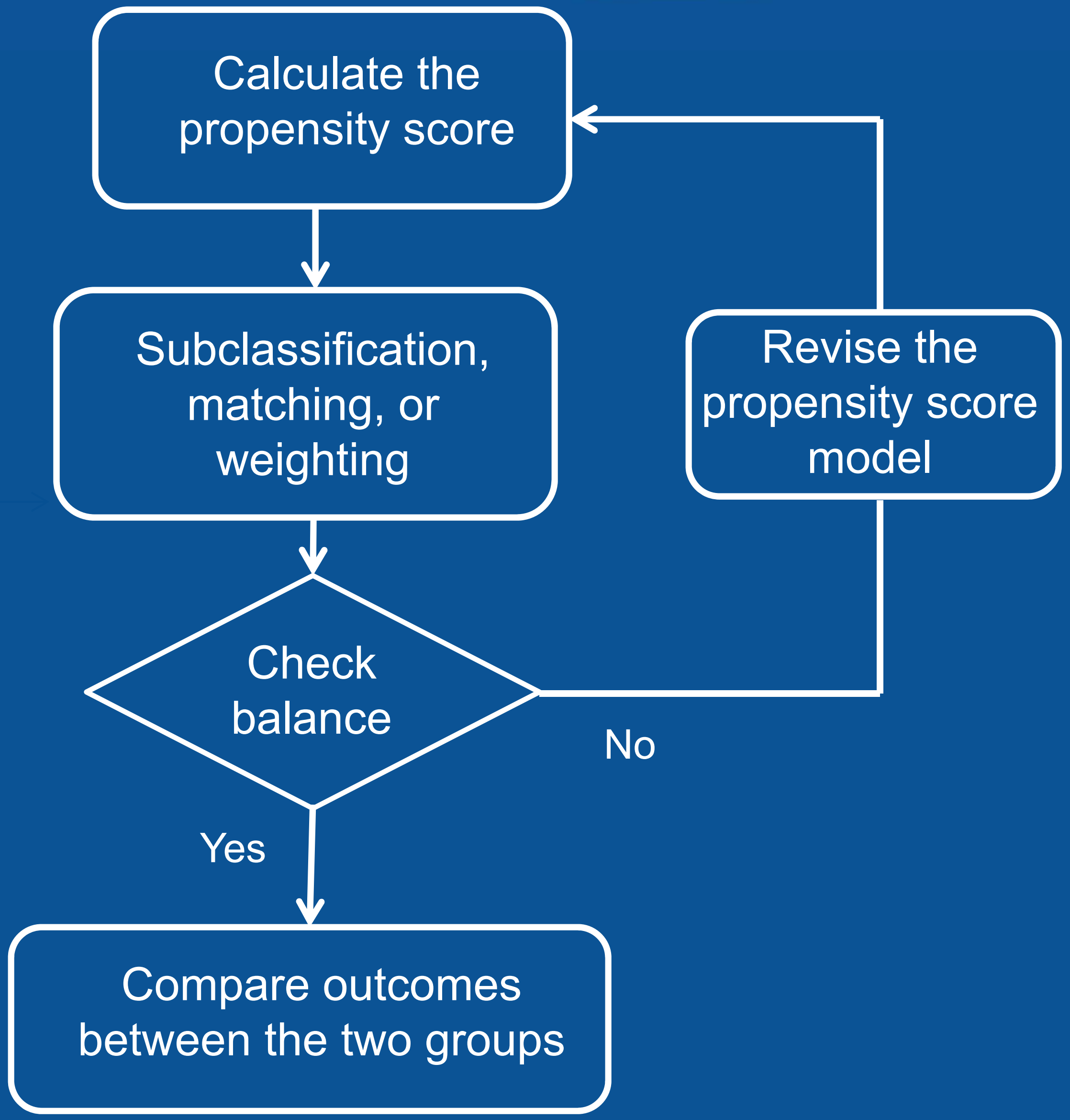
- Given the propensity score, the treatment assignment is independent of the covariates
- If several subjects have the same propensity score, i.e., the probability of treatment assignment, they form a small randomized controlled trial



What does the mirror histogram of a RCT look like?

A Quick Overview of the
Propensity Score Analysis Procedure

Propensity Score Analysis



Calculate the propensity score

Subclassification, matching, or weighting

Check balance

Yes

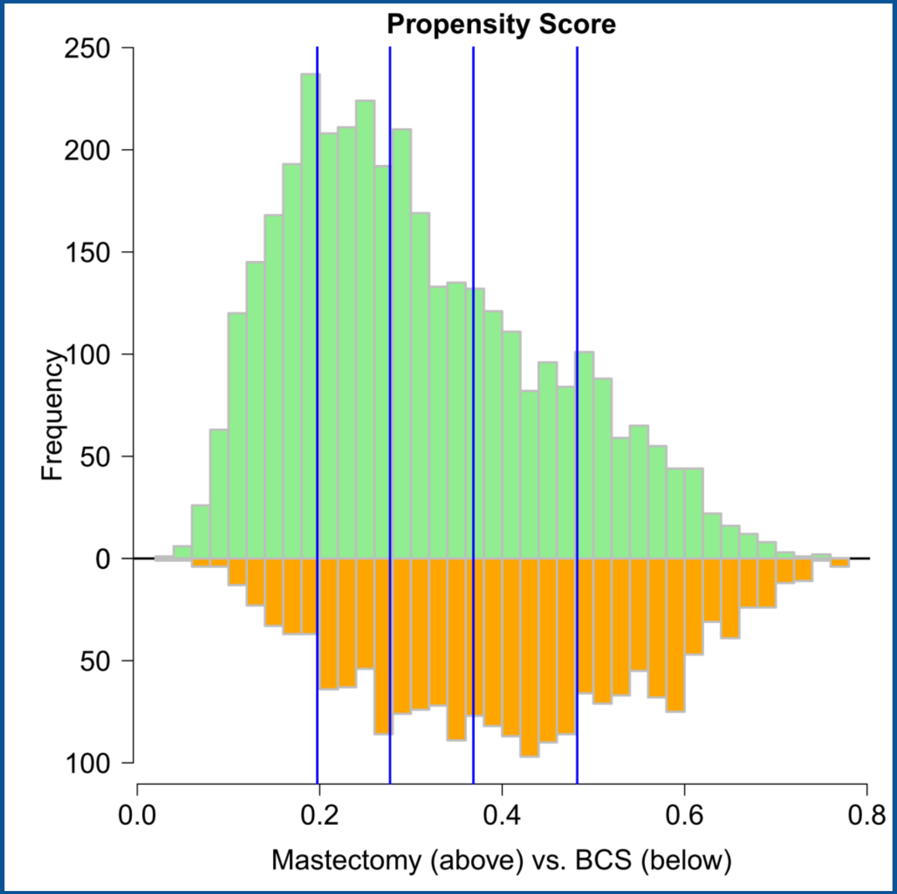
No

Revise the propensity score model

Compare outcomes between the two groups

Subclassification

- Aggregate within each stratum, and then average across all strata, similar to the analysis of stratified RCT (multicenter RCT)
- 5 strata removes 90% of bias (rule of thumb)



Mastectomy	937	827	731	591	501
BCS	150	259	355	493	588

Subclassification

Covariates	Mastectomy	BCS
N	3,587	1,845
Age	62	66
Tumor grade = 1	0.28	0.36
Tumor grade = 2	0.45	0.43
Tumor grade = 3	0.28	0.21
Hormone receptor	0.83	0.86
Hormone therapy	0.53	0.33
Chemo therapy	0.27	0.14
Charlson Deyo Comorb. index > 0	0.16	0.12

strata	1		2		3		4		5	
	Mast.	BCS	Mast.	BCS	Mast.	BCS	Mast.	BCS	Mast.	BCS
Age	54	55	61	61	64	66	67	66	71	71
Grade = 1	0.14	0.11	0.25	0.26	0.3	0.3	0.31	0.33	0.49	0.53
Grade = 2	0.49	0.5	0.46	0.42	0.42	0.43	0.44	0.43	0.4	0.41
Grade = 3	0.37	0.39	0.29	0.32	0.28	0.28	0.25	0.24	0.11	0.05
H receptor	0.85	0.76	0.79	0.76	0.78	0.79	0.82	0.86	0.93	0.96
H therapy	0.86	0.8	0.71	0.63	0.54	0.51	0.2	0.26	0.02	0.03
CHEMO	0.57	0.58	0.29	0.32	0.19	0.17	0.07	0.06	0.01	0.01
CDCI > 0	0.23	0.31	0.17	0.18	0.15	0.15	0.12	0.10	0.07	0.04

Subclassification

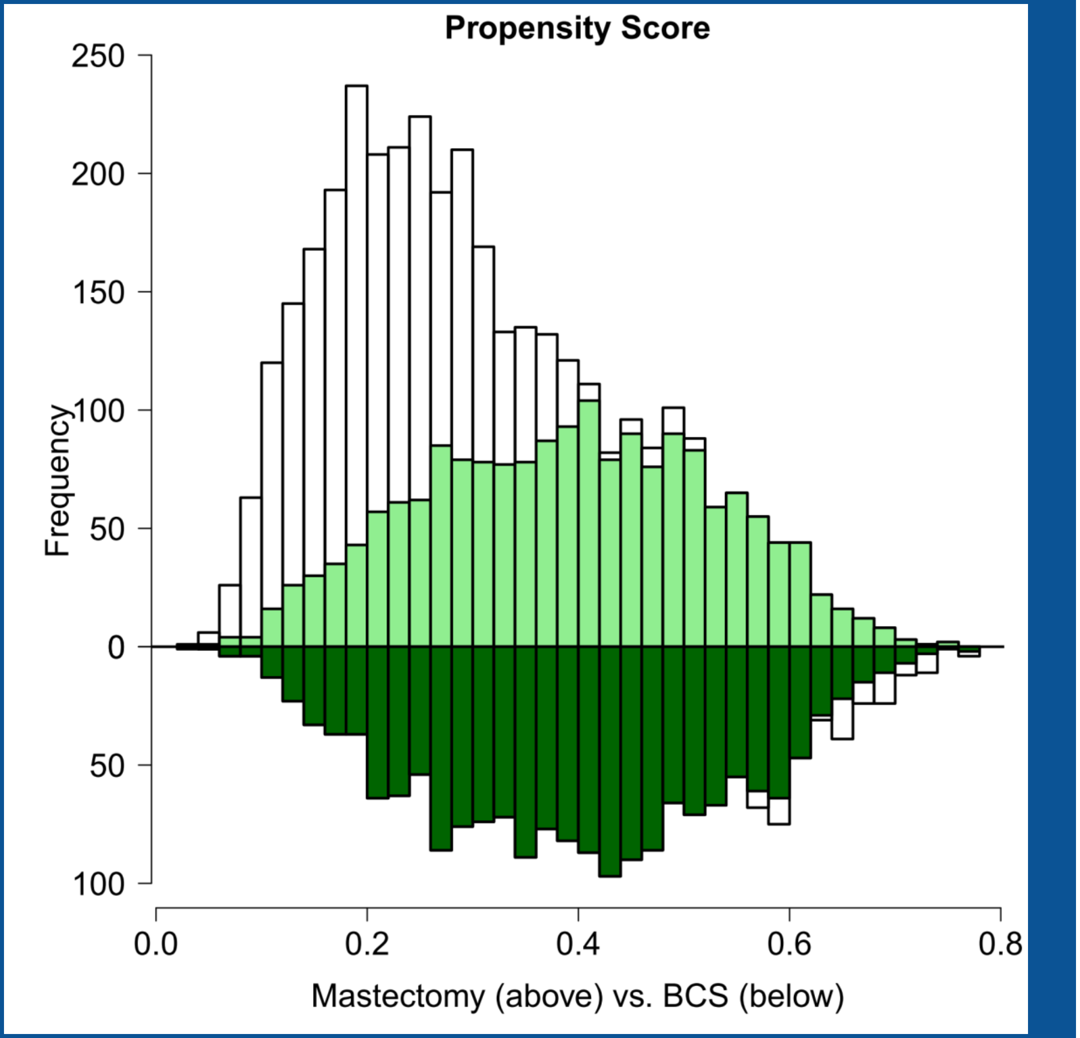
- Recapitulation:
  - ✓ Stratification is a crude way of partitioning the data from a nonrandomized study into several small data sets that look like randomized controlled trials
  - ✓ How to choose the number of strata?
  - ✓ Developed in the 1980s; simple to implement, but the within-strata balance may not always be good
  - ✓ Refinement available: nonparametric (kernel) regression

    Outcome ~ g(propensity score) + β × treatment indicator

    - ➢ Averaging over overlapping, continuously moving strata
    - ➢ Produce better results than simple stratification but need fine-tuning: the kernel, balance checking, and p-values

Pair Matching

- Pair matching (1:1 matching without replacement) is the most widely used in medicine
- It resembles the randomized controlled trials with 1:1 allocation



Results from Pair Matching

Covariates	Before matching		After matching	
	Mastectomy	BCS	Mastectomy	BCS
N	3,587	1,845	1,769	1,769
Age	62	66	65	66
Tumor grade = 1	0.28	0.36	0.34	0.34
Tumor grade = 2	0.45	0.43	0.43	0.44
Tumor grade = 3	0.28	0.21	0.23	0.22
Hormone receptor	0.83	0.86	0.84	0.85
Hormone therapy	0.53	0.33	0.35	0.34
Chemo therapy	0.27	0.14	0.14	0.15
Charlson Deyo Comorb. index > 0	0.16	0.12	0.12	0.12

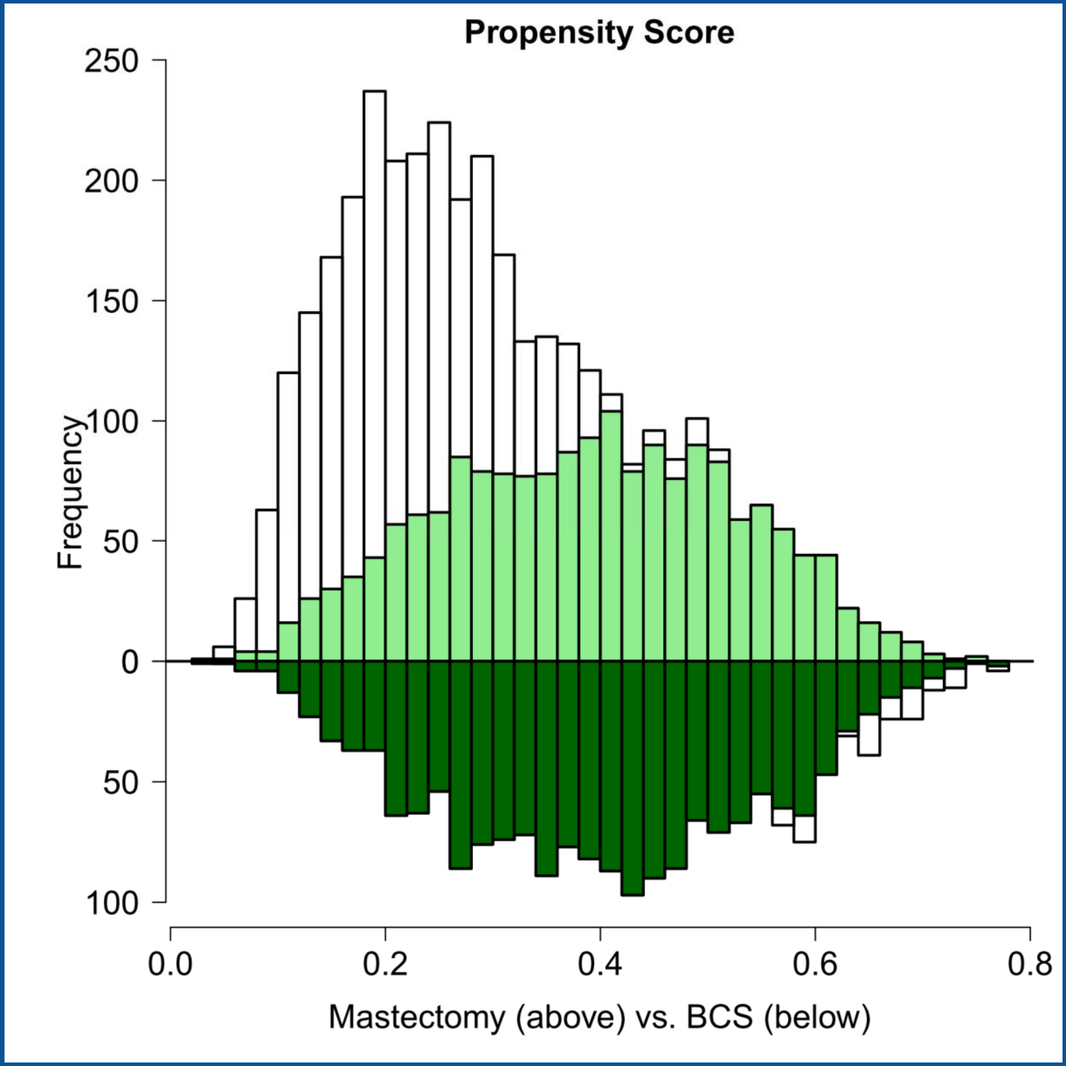
Weighting

	White	Hispanic	African American	Asian	Other
TX	43.9	26.5	11.8	3.8	13.9
CA	35.9	21.7	6.2	13.0	23.2
OH	80.2	2.5	12.2	1.7	3.5

Weights					
CA	1	1	0.64	4.2	2
OH	1.8	0.09	1	0.42	0.25

Weighting

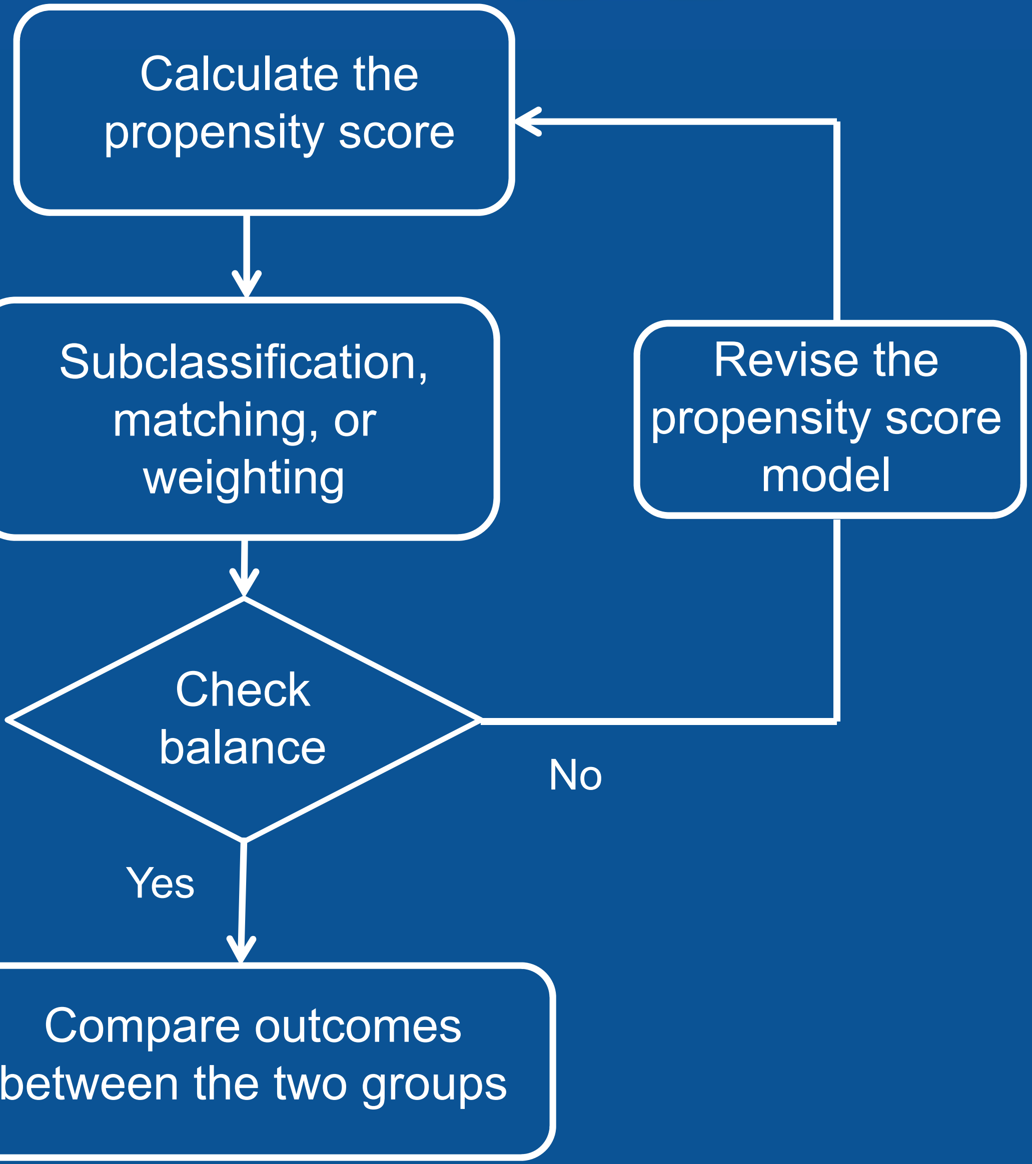
- Weight the control subjects to make them resemble the treated (Average treatment effect for the treated)
- Weight the treated subjects to make them resemble the controls (Average treatment effect for the controls)
- Or, weight both the treated and control subjects to make them similar but deviate from either sample



Results from Weighting

Covariates	Before weighting		After weighting	
	Mastectomy	BCS	Mastectomy	BCS
N	3,587	1,845		
Age	62	66	65	65
Tumor grade = 1	0.28	0.36	0.34	0.34
Tumor grade = 2	0.45	0.43	0.43	0.43
Tumor grade = 3	0.28	0.21	0.23	0.23
Hormone receptor	0.83	0.86	0.84	0.85
Hormone therapy	0.53	0.33	0.35	0.36
Chemo therapy	0.27	0.14	0.15	0.15
Charlson Deyo Comorb. index > 0	0.16	0.12	0.13	0.13

Propensity Score Analysis



Calculate the propensity score

Subclassification, matching, or weighting

Check balance

Yes

Compare outcomes between the two groups

No

Revise the propensity score model

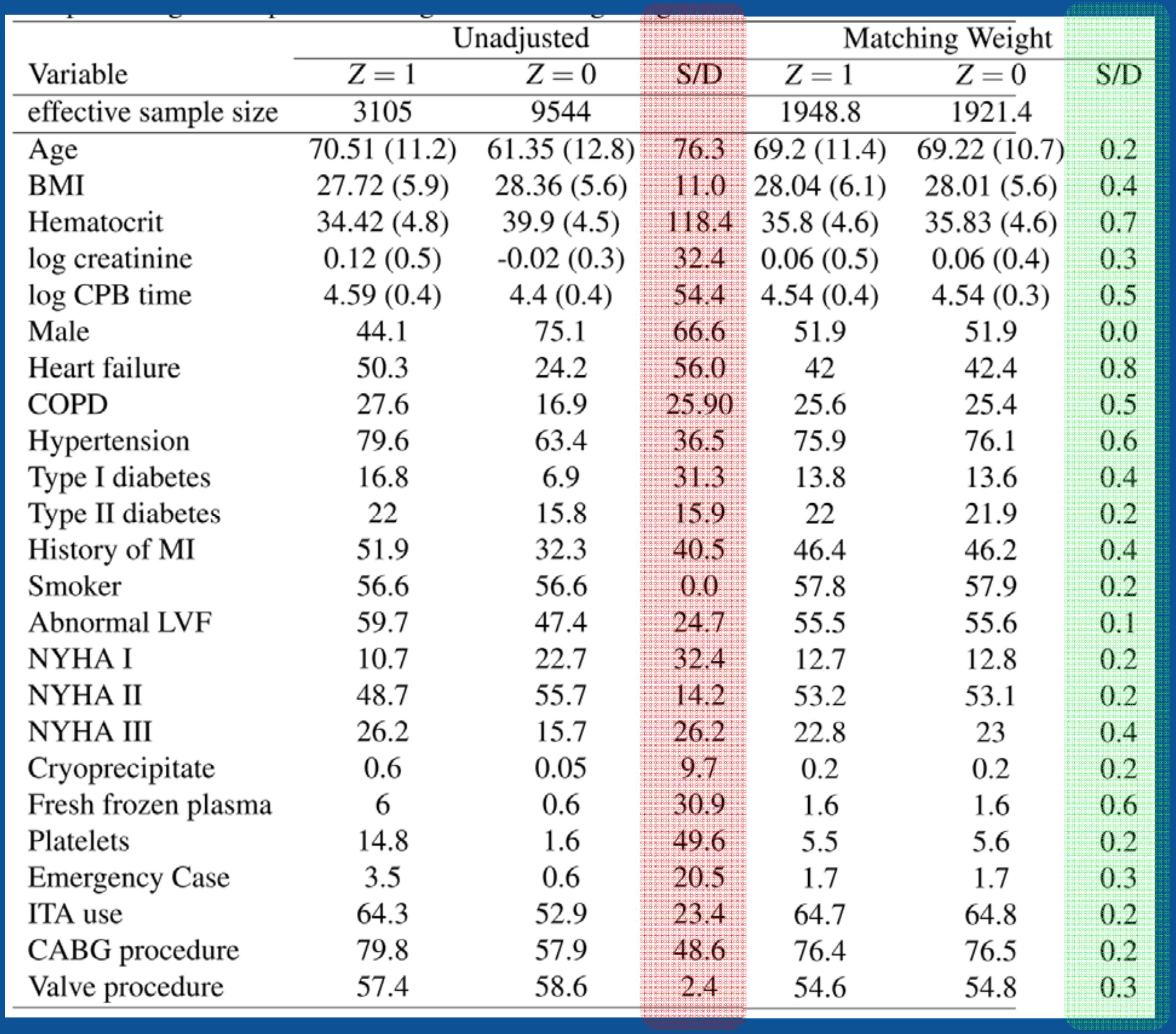
Checking Balance

	Before matching		After matching		After matching	
Covariates	Mastectomy	BCS	Mastectomy	BCS	Mastectomy	BCS
N	3,587	1,845	1,769	1,769	1,769	1,769
Age	62	66	65	66	65	66
Tumor grade = 1	0.28	0.36	0.34	0.34	0.34	0.34
Tumor grade = 2	0.45	0.43	0.43	0.44	0.40	0.44
Tumor grade = 3	0.28	0.21	0.23	0.22	0.23	0.22
Hormone receptor	0.83	0.86	0.84	0.85	0.84	0.85
Hormone therapy	0.53	0.33	0.35	0.34	0.38	0.34
Chemo therapy	0.27	0.14	0.14	0.15	0.14	0.15

- The balance can be checked in a similar way as randomized controlled trials
  - ✓ Do not use P-values; standardized differences or empirical distribution OK
  - ✓ Including interaction terms, nonlinear terms, or matching on both propensity score and some important covariates may help (just keep trying … …)
  - ✓ How small the differences need to be for adequate balance? (zero differences are impossible)

Goodness of Fit Test for Balance

- "How small the standardized difference is small enough?" 10% ?
- Test of balance is often undesirable in the matching framework because (Austin 2008; Imai, King, Stuart 2008)
  - ✓ Balance is a property of the sample, but not the population
  - ✓ Sample size is often reduced after matching
- Cast the problem as a test of misspecification of the PS model (Li and Greene 2013; currently only available for weighting methods)
  - ✓ Null hypothesis: the propensity score model is correctly specified
  - ✓ Alternative hypothesis: the propensity score model is misspecified
  - ✓ It is a chi-square test and the test statistic is a measure of balance

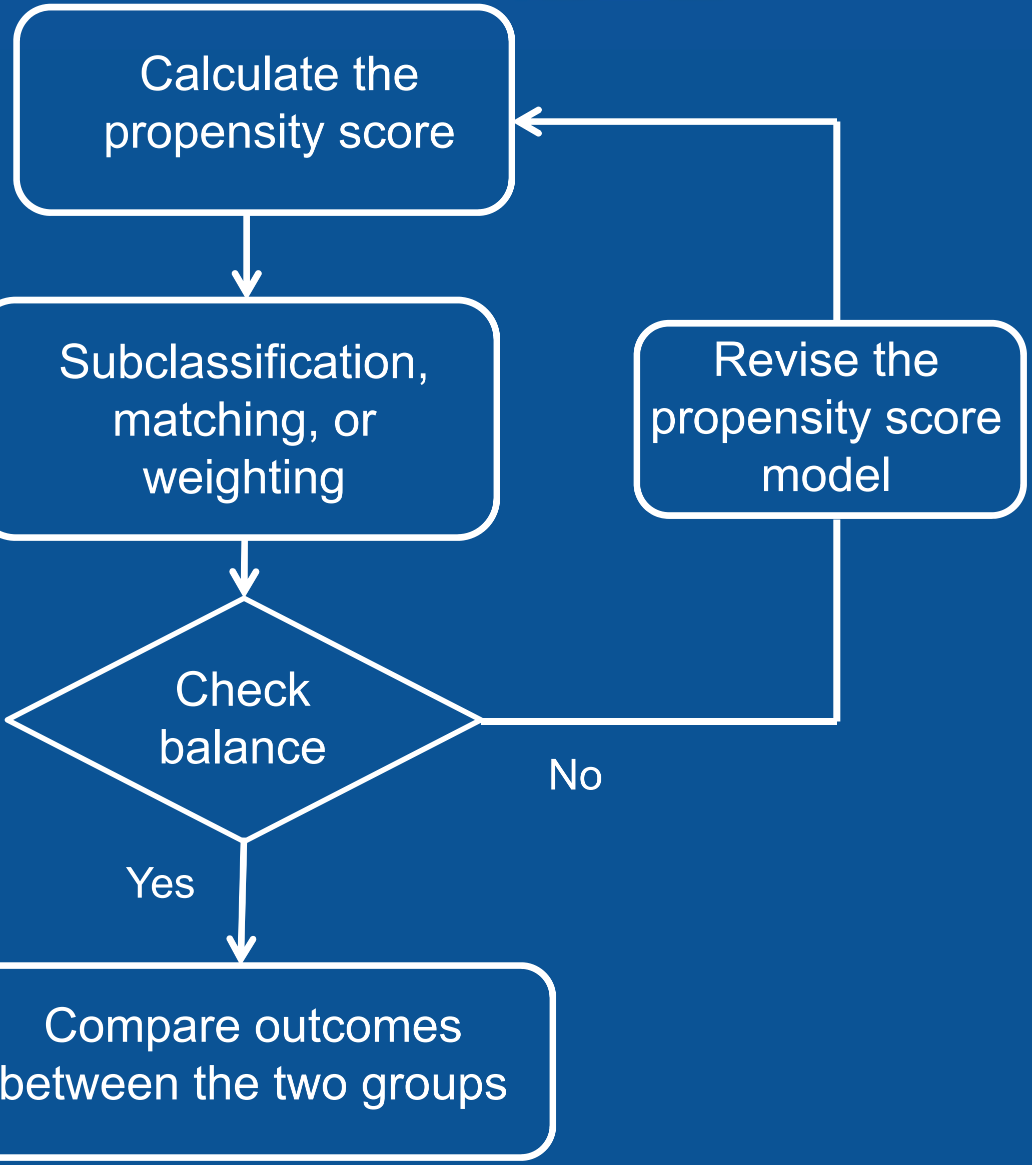


Variable	Unadjusted $Z=1$	 $Z=0$	 S/D	Matching Weight $Z=1$	 $Z=0$	 S/D
effective sample size	3105	9544		1948.8	1921.4	
Age	70.51 (11.2)	61.35 (12.8)	76.3	69.2 (11.4)	69.22 (10.7)	0.2
BMI	27.72 (5.9)	28.36 (5.6)	11.0	28.04 (6.1)	28.01 (5.6)	0.4
Hematocrit	34.42 (4.8)	39.9 (4.5)	118.4	35.8 (4.6)	35.83 (4.6)	0.7
log creatinine	0.12 (0.5)	-0.02 (0.3)	32.4	0.06 (0.5)	0.06 (0.4)	0.3
log CPB time	4.59 (0.4)	4.4 (0.4)	54.4	4.54 (0.4)	4.54 (0.3)	0.5
Male	44.1	75.1	66.6	51.9	51.9	0.0
Heart failure	50.3	24.2	56.0	42	42.4	0.8
COPD	27.6	16.9	25.90	25.6	25.4	0.5
Hypertension	79.6	63.4	36.5	75.9	76.1	0.6
Type I diabetes	16.8	6.9	31.3	13.8	13.6	0.4
Type II diabetes	22	15.8	15.9	22	21.9	0.2
History of MI	51.9	32.3	40.5	46.4	46.2	0.4
Smoker	56.6	56.6	0.0	57.8	57.9	0.2
Abnormal LVF	59.7	47.4	24.7	55.5	55.6	0.1
NYHA I	10.7	22.7	32.4	12.7	12.8	0.2
NYHA II	48.7	55.7	14.2	53.2	53.1	0.2
NYHA III	26.2	15.7	26.2	22.8	23	0.4
Cryoprecipitate	0.6	0.05	9.7	0.2	0.2	0.2
Fresh frozen plasma	6	0.6	30.9	1.6	1.6	0.6
Platelets	14.8	1.6	49.6	5.5	5.6	0.2
Emergency Case	3.5	0.6	20.5	1.7	1.7	0.3
ITA use	64.3	52.9	23.4	64.7	64.8	0.2
CABG procedure	79.8	57.9	48.6	76.4	76.5	0.2
Valve procedure	57.4	58.6	2.4	54.6	54.8	0.3

The Analysis of the Outcome

- The analysis of the outcome variable is straightforward in propensity score analysis: a two sample comparison, similar to a randomized controlled trial
- This simplicity is very attractive compared with regression methods
- Be careful about estimated variances and, hence, p-values, particularly when the p-values are on the borderline:
  - ✓ The p-values from matching tend to be larger than they should be
  - ✓ The p-values from weighting methods can be very accurate
  - ✓ Currently developing software for weighting methods in general

Propensity Score Analysis


Calculate the propensity score
Subclassification, matching, or weighting
Check balance
No
Revise the propensity score model
Yes
Compare outcomes between the two groups

Additional Methodological and Practical Issues

Propensity Score vs. Regression

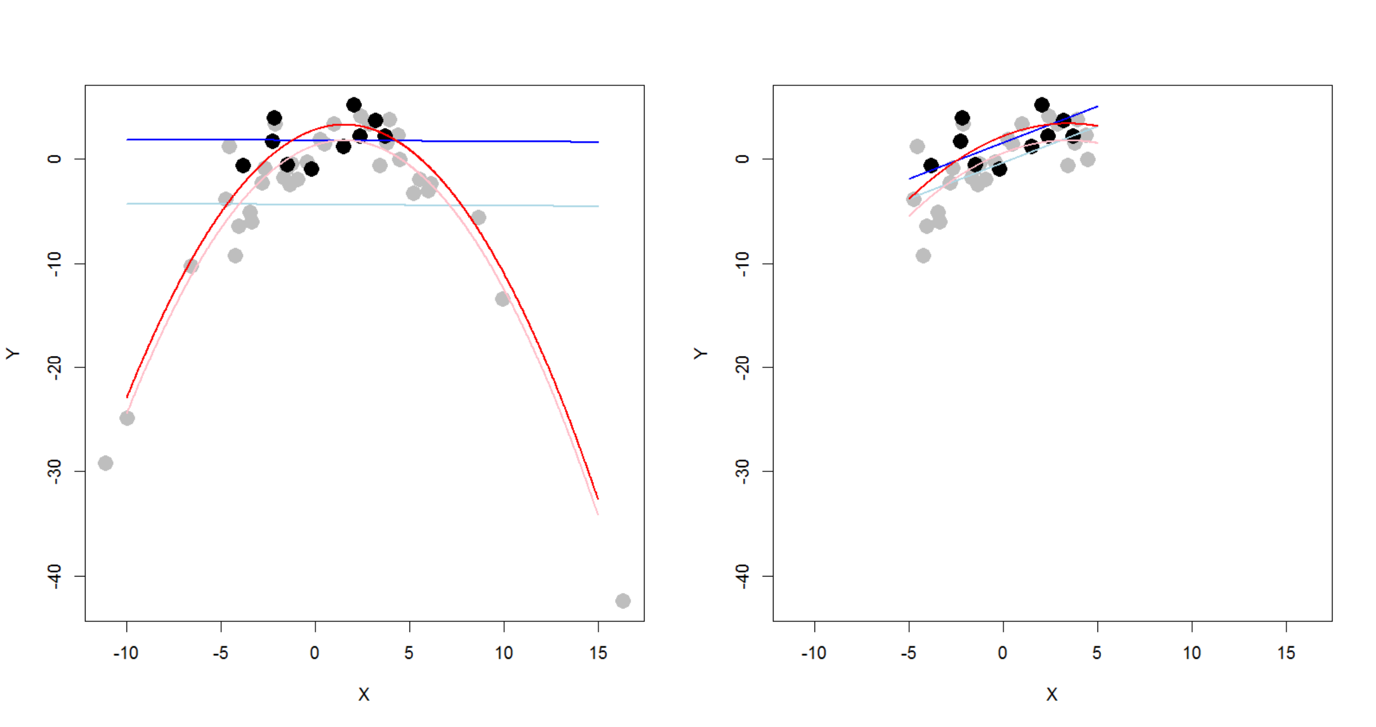
- How about studying the treatment effect using regression?

Outcome ~ Treatment/control + covariates

Propensity Score	Regression
Only study the effect of the treatment	Study the effect of many covariates
Weaker assumptions due to model checking (logistic regression is a means to an end)	Stronger modeling assumptions
Separate design from analysis, like an RCT	Not separate
Propensity score matching or weighting can be used in combination with an outcome regression	

- Conclusion: Propensity score is superior to regression when only the treatment effect is of interest

Propensity Score vs. Regression



- Before: 6.16 in linear model, 1.51 in quadratic model
- After: 1.91 in linear model, 1.62 in quadratic model
- Propensity score matching as a preprocessing step of regression modeling (Ho, Imai, King, Stuart, 2007): robust and may be even doubly robust

Which Covariates to Use?

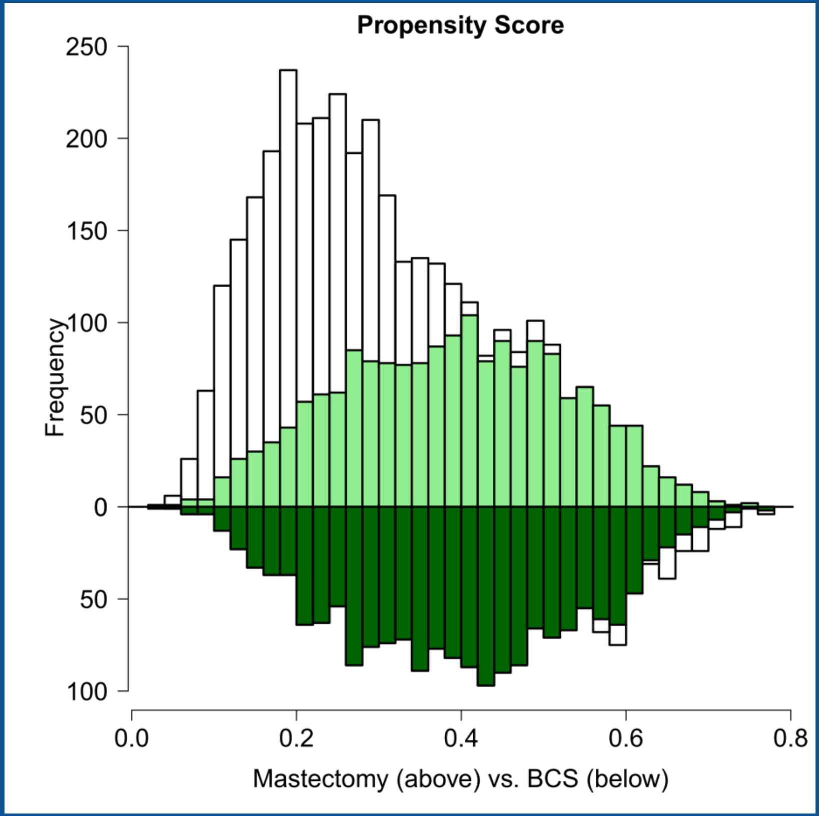
- Rule 1: If you believe that the imbalance of a covariate may cause bias in estimating the treatment effect, that covariate should be included
- Rule 2: If you are not sure, include it (be conservative and avoid bias).
- The form of the logistic regression and the interpret of the coefficients is not important; all we need is a logistic regression model that produce satisfactory balance (a means to an end)
- Therefore, despite the use a parametric logistic regression model, the propensity score analysis is not viewed as a parametric method

Choosing Matching or Weighting Methods

Matching	Weighting
Pair matching	Matching weight
Matching with replacement	Inverse probability weight for the treated
Matching without replacement	Inverse probability weight for the control
Greedy matching	Inverse probability weight for both the treated and control
Optimal matching	Overlapping weight
Full matching	Kernel weight
… …	… …

Matching vs. Weighting

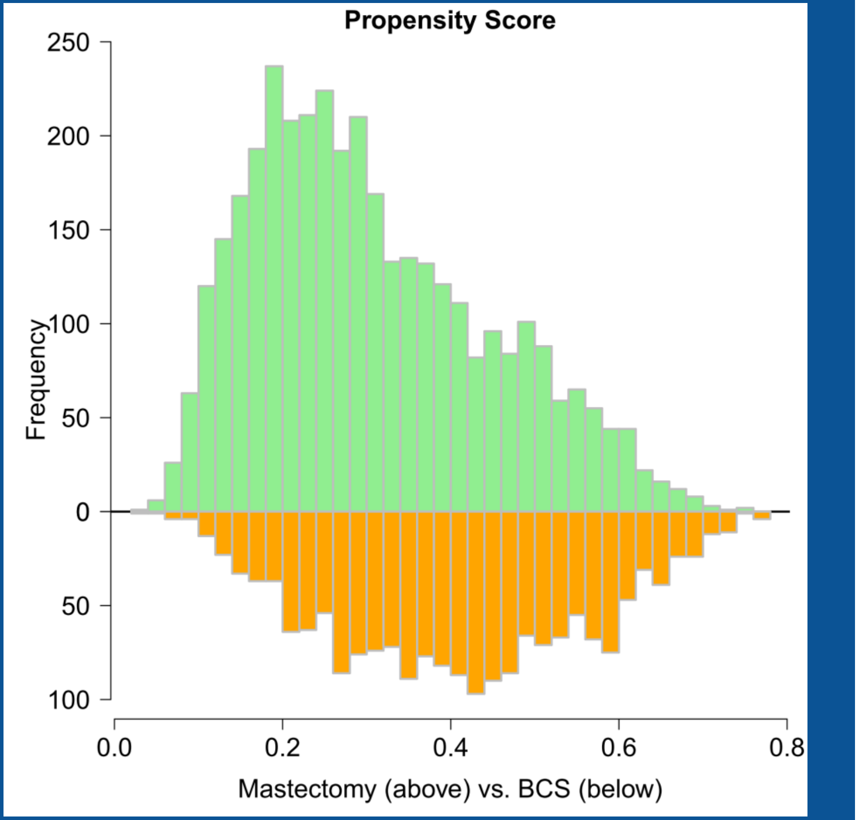
Covariates	Original		Matching Wt	
	Mastectomy	BCS	Mastectomy	BCS
N	3,587	1,845		
Age	62	66	65	65
grade 1	0.28	0.36	0.34	0.34
grade 2	0.45	0.43	0.43	0.43
grade 3	0.28	0.21	0.23	0.23
Hormone R	0.83	0.86	0.84	0.85
Hormone T	0.53	0.33	0.35	0.36
Chemo T	0.27	0.14	0.15	0.15
Charlson > 0	0.16	0.12	0.13	0.13



- Matching Weight: balance is good after weighting, the weighted sample is different from Mastectomy, BCS, or the original sample, it resembles the 1:1 matched sample

Matching vs. Weighting

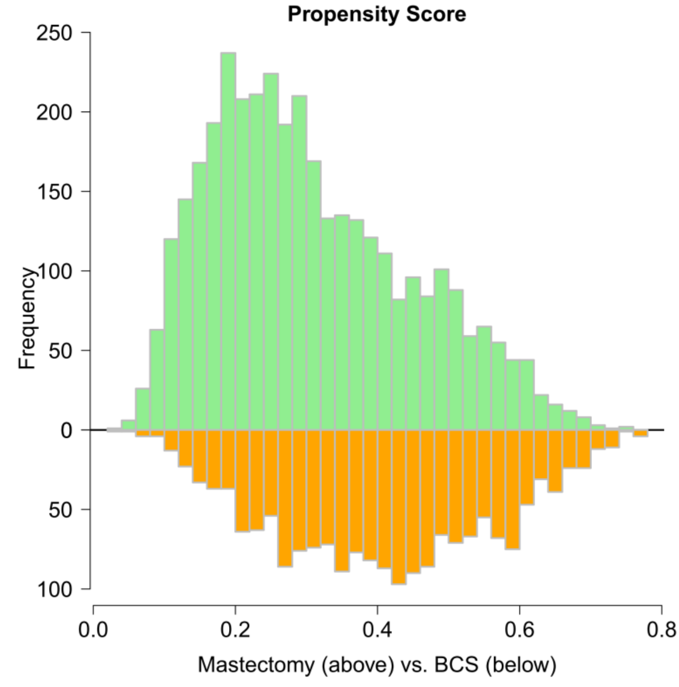
Covariates	Original		IPW T	
	Mastectomy	BCS	Mastectomy	BCS
N	3,587	1,845		
Age	62	66	66	66
grade 1	0.28	0.36	0.36	0.36
grade 2	0.45	0.43	0.42	0.43
grade 3	0.28	0.21	0.22	0.21
Hormone R	0.83	0.86	0.85	0.86
Hormone T	0.53	0.33	0.32	0.33
Chemo T	0.27	0.14	0.14	0.14
Charlson > 0	0.16	0.12	0.12	0.12



- Inverse Probability Weight for the Treated: balance is good after weighting, the weighted sample resembles the BCS group

Matching vs. Weighting

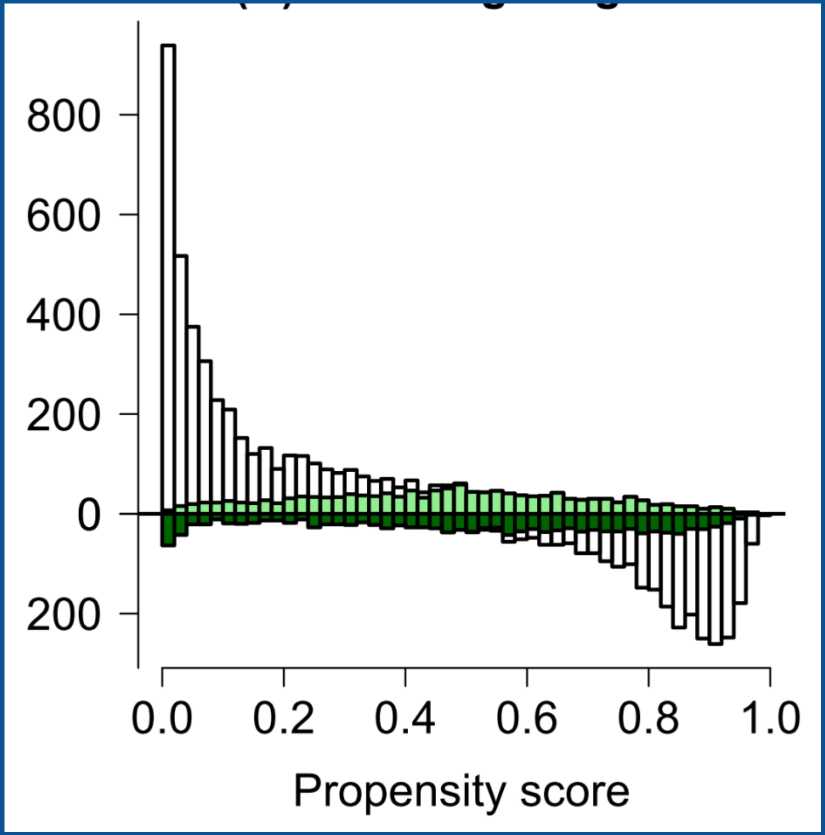
Covariates	Original			IPW	
	Mastectomy	BCS	combo	Mastectomy	BCS
N	3,587	1,845	5,432		
Age	62	66	63	63	63
grade 1	0.28	0.36	0.30	0.30	0.30
grade 2	0.45	0.43	0.44	0.44	0.44
grade 3	0.28	0.21	0.25	0.26	0.26
Hormone R	0.83	0.86	0.84	0.83	0.83
Hormone T	0.53	0.33	0.46	0.46	0.46
Chemo T	0.27	0.14	0.22	0.23	0.24
Charlson > 0	0.16	0.12	0.15	0.15	0.16



- Inverse Probability Weight: balance is good after weighting, the weighted sample resembles the original data (mastectomy + BCS)

Matching vs. Weighting

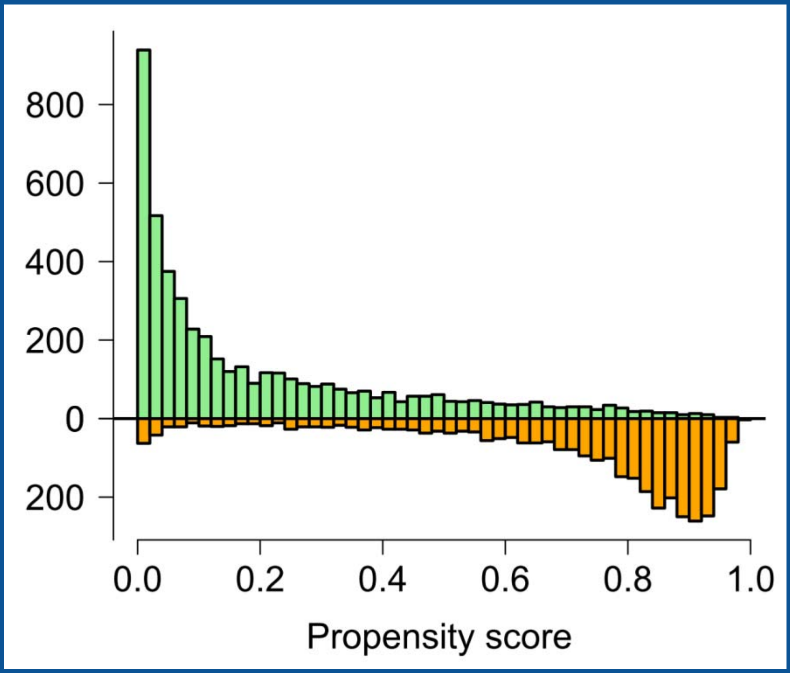
	Original		Matching Wt	
Covariates	None	COPD	None	COPD
N	4,842	3,422		
Age	45	57	55	54
BMI	24	24	25	25
Log(creatinine)	-0.20	-0.19	-0.15	-0.15
FEV1	43	25	35	33
Female	0.44	0.51	0.47	0.46
Diabetes	0.18	0.05	0.08	0.09
Hypertension	0.21	0.23	0.24	0.23
Double lung Tx	0.68	0.41	0.52	0.53



- Matching Weight: balance is good after weighting, the weighted sample is different from either group, or the original sample, it resembles the 1:1 matched sample

Matching vs. Weighting

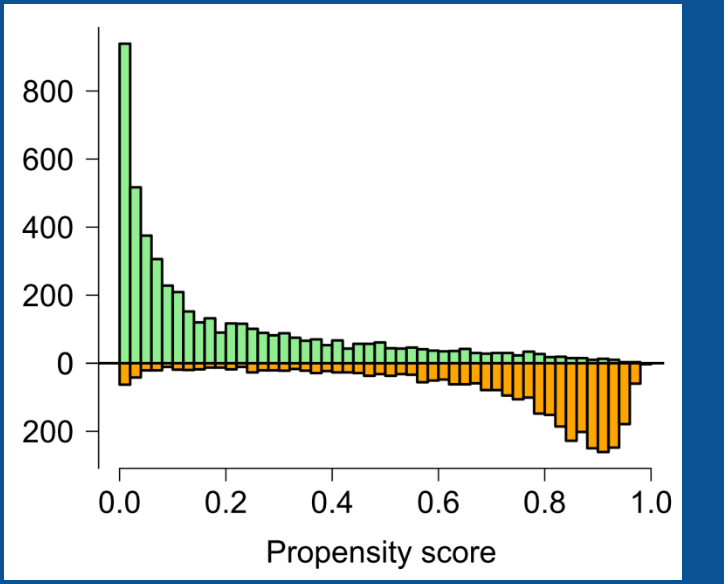
Covariates	Original		IPW T	
	None	COPD	None	COPD
N	4,842	3,422		
Age	45	57	58	57
BMI	24	24	24	24
Log(creatinine)	-0.20	-0.19	-0.15	-0.19
FEV1	43	25	30	25
Female	0.44	0.51	0.49	0.51
Diabetes	0.18	0.05	0.05	0.05
Hypertension	0.21	0.23	0.23	0.23
Double lung Tx	0.68	0.41	0.45	0.41



- Inverse Probability Weight for the Treated: balance is not good after weighting, the weighted sample resembles the COPD group

Matching vs. Weighting

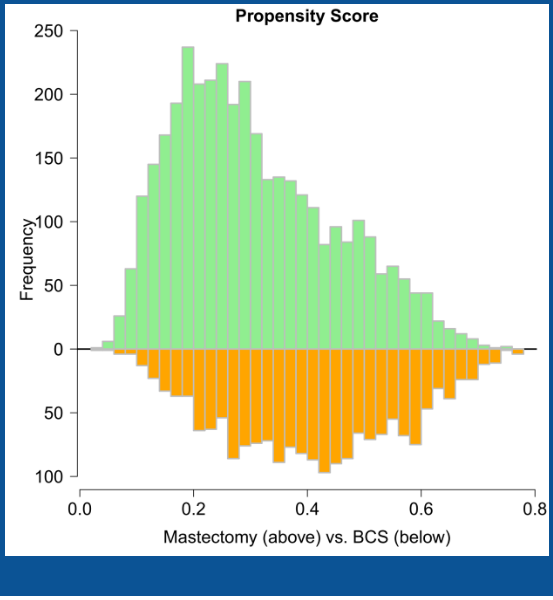
Covariates	Original			IPW	
	None	COPD	combo	None	COPD
N	4,842	3,422	8,264		
Age	45	57	50	50	53
BMI	24	24	24	24	24
Log(creatinine)	-0.20	-0.19	-0.19	-0.18	-0.06
FEV1	43	25	36	38	92
Female	0.44	0.51	0.47	0.46	0.51
Diabetes	0.18	0.05	0.12	0.13	0.09
Hypertension	0.21	0.23	0.22	0.21	0.30
Double lung Tx	0.68	0.41	0.57	0.60	0.87

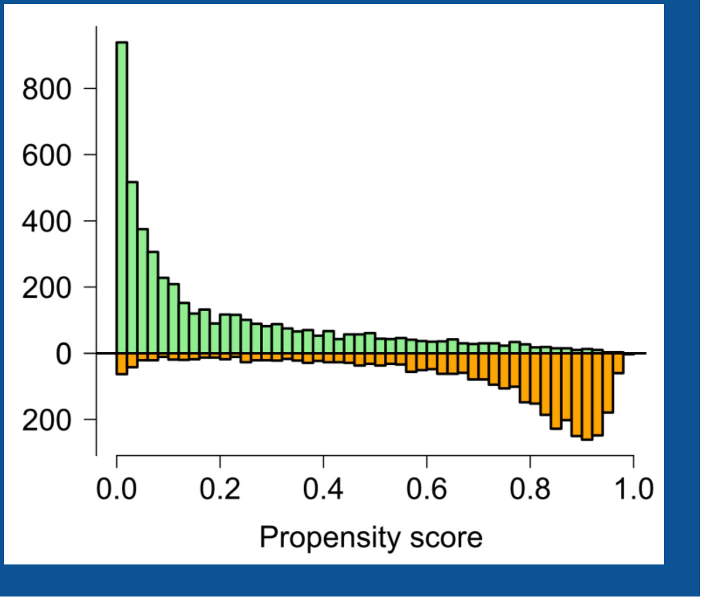


- Inverse Probability Weight: balance is not good after weighting, the weighted sample resembles the original data

Matching vs. Weighting

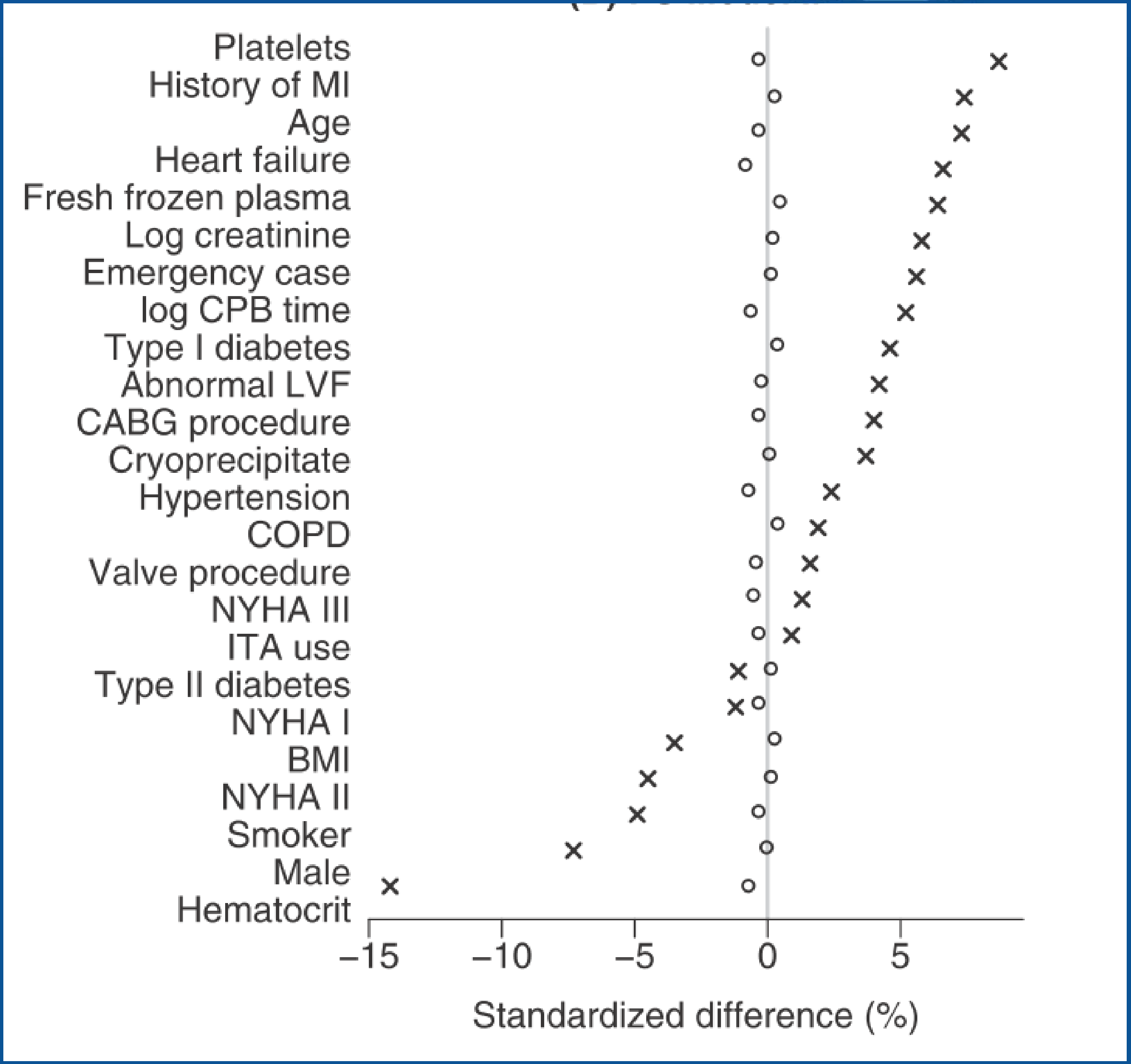
- Conclusion:
  - ✓ The results using various matching methods are similar
  - ✓ Different matching or weighting methods estimate different quantities and have different interpretation. Need to be clear about what we are trying to estimate in the paper or grant proposal.
  - ✓ Must justify the choice of the method from both a medical and statistical perspective.





Matching vs. Weighting

- Choosing between matching and weighting:
  - ✓ Weighting usually produces better balance than matching (next page)
  - ✓ It is also more accurate (higher efficiency)
  - ✓ CAUTION: The p-values from matching is usually larger than it should be (conservative), but the p-values from weighting can be made accurately (Abadie and Imbens 2009; Stuart 2010; Li, Greene, Bauer, in press)
  - ✓ Be careful when the weights are excessively large:
    - ➢ Use matching weight or overlapping weight instead
    - ➢ Use Covariate Balancing Propensity Score (Imai 2014)


Platelets
History of MI
Age
Heart failure
Fresh frozen plasma
Log creatinine
Emergency case
log CPB time
Type I diabetes
Abnormal LVF
CABG procedure
Cryoprecipitate
Hypertension
COPD
Valve procedure
NYHA III
ITA use
Type II diabetes
NYHA I
BMI
NYHA II
Smoker
Male
Hematocrit
–15
–10
–5
0
5
Standardized difference (%)

Generalized Propensity Score

- Compare more than two treatment groups; compare continuous exposure variables (e.g., antibiotic timing in the operating room)
- Instead of using logistic regression, use nested logistic regression (proportional odds/cumulative logistic model undesirable)
- Instead of a single propensity score for each subject, there are several propensity scores for each subject
- Instead of matching on a single score, match on several propensity scores
- Lack of overlap becomes a bigger problem
- More research is needed

Generalized Propensity Score

Covariates	Before weighting			After weighting		
	BCS	BCT	Mastectomy	BCS	BCT	Mastoctomy
Age	66	59	62	66	66	66
grade 1	0.36	0.32	0.28	0.36	0.36	0.36
grade 2	0.43	0.43	0.45	0.43	0.43	0.42
grade 3	0.21	0.25	0.28	0.21	0.21	0.22
Hormone R	0.86	0.85	0.83	0.86	0.86	0.85
Hormone T	0.33	0.71	0.53	0.33	0.32	0.32
Chemo T	0.14	0.31	0.27	0.14	0.14	0.14

Summary

- Fundamental to causal inference; widely used, and under active research
- Superior to multivariate regression model in estimating the treatment effect (but not estimating the effect of many risk factors); more objective because design and analysis are separated
- The two can be used together
- Recent research suggests that weighting generally have better performance than stratification and matching, but they must be dealt with care, particularly when the overlap is not good. We must consider from both scientific and technical perspectives when choosing methods.
- Sensitivity analysis on unmeasured covariates (Guo and Fraser, 2009)
- Software for matching (Stuart 2010); CBPS (Imai 2014); our weighting package in R

Thank You!