PROPENSITY SCORE METHODS: THEORY AND CONCEPTS

Part 1: Reducing model dependence

Brian Lee (bklee@drexel.edu)
Assistant Professor
Dept. of Epidemiology and Biostatistics
Drexel University School of Public Health

SPER workshop June 17, 2013

Model dependence occurs…

- …when estimates depend on the particular model used
- If the model is a poor representation of nature, the conclusions may be wrong

Model dependence in prediction

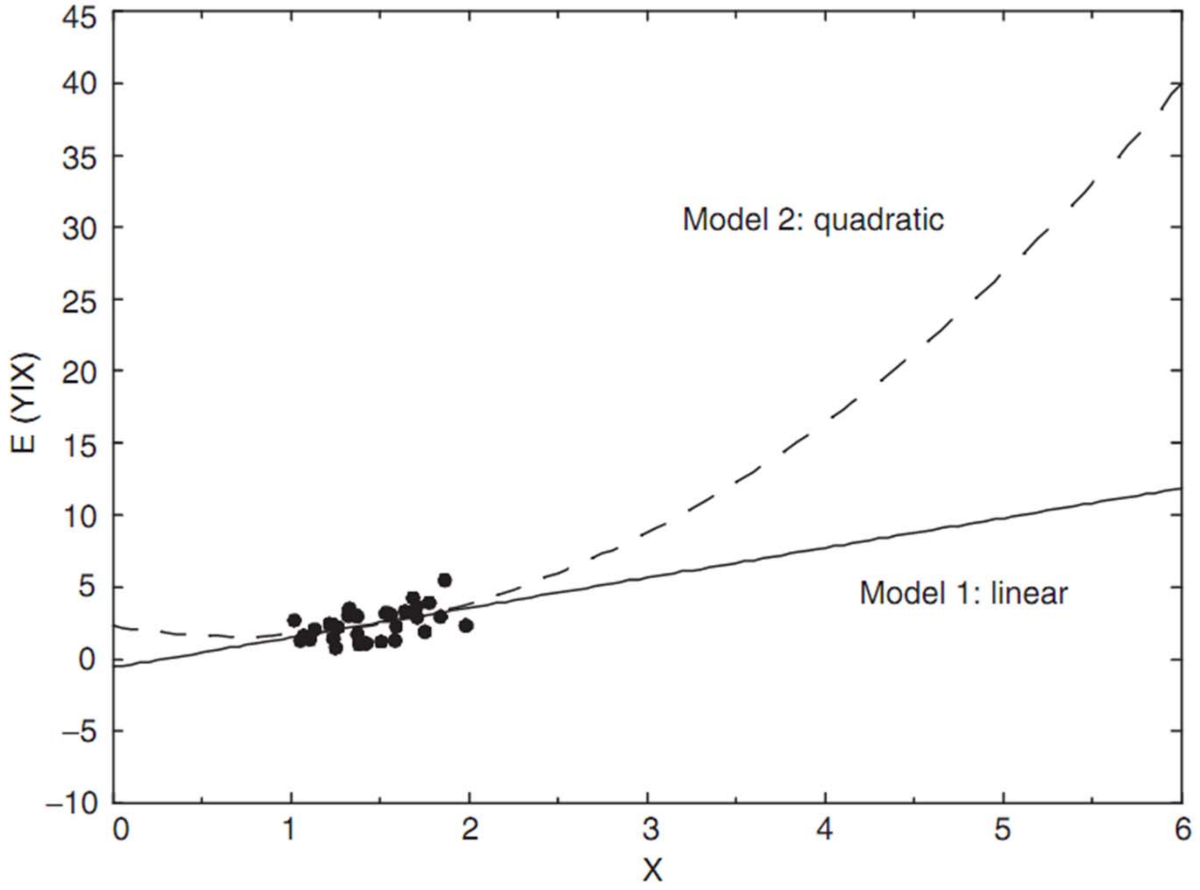


FIG. 1. Linear and Quadratic Models With Equal Fit to Simulated Data But Massively Different Out-of-Sample Implications

King and Zeng, 2007

Model dependence in causal inference

- Example from real study of whether change in political leadership affects drug approval time (more details later)
- 18 covariates to possibly include as linear predictors
- Every combination of covariates (no non-linearities and interactions)!

N choose R	Combinations
(18, 1)	18
…	…
(18, 4)	3,060
…	…
(18, 9)	48,620
…	…
(18, 18)	1
TOTAL NUMBER OF COMBINATIONS	**262,143**

Ho, 2007

Estimates vary according to model choice

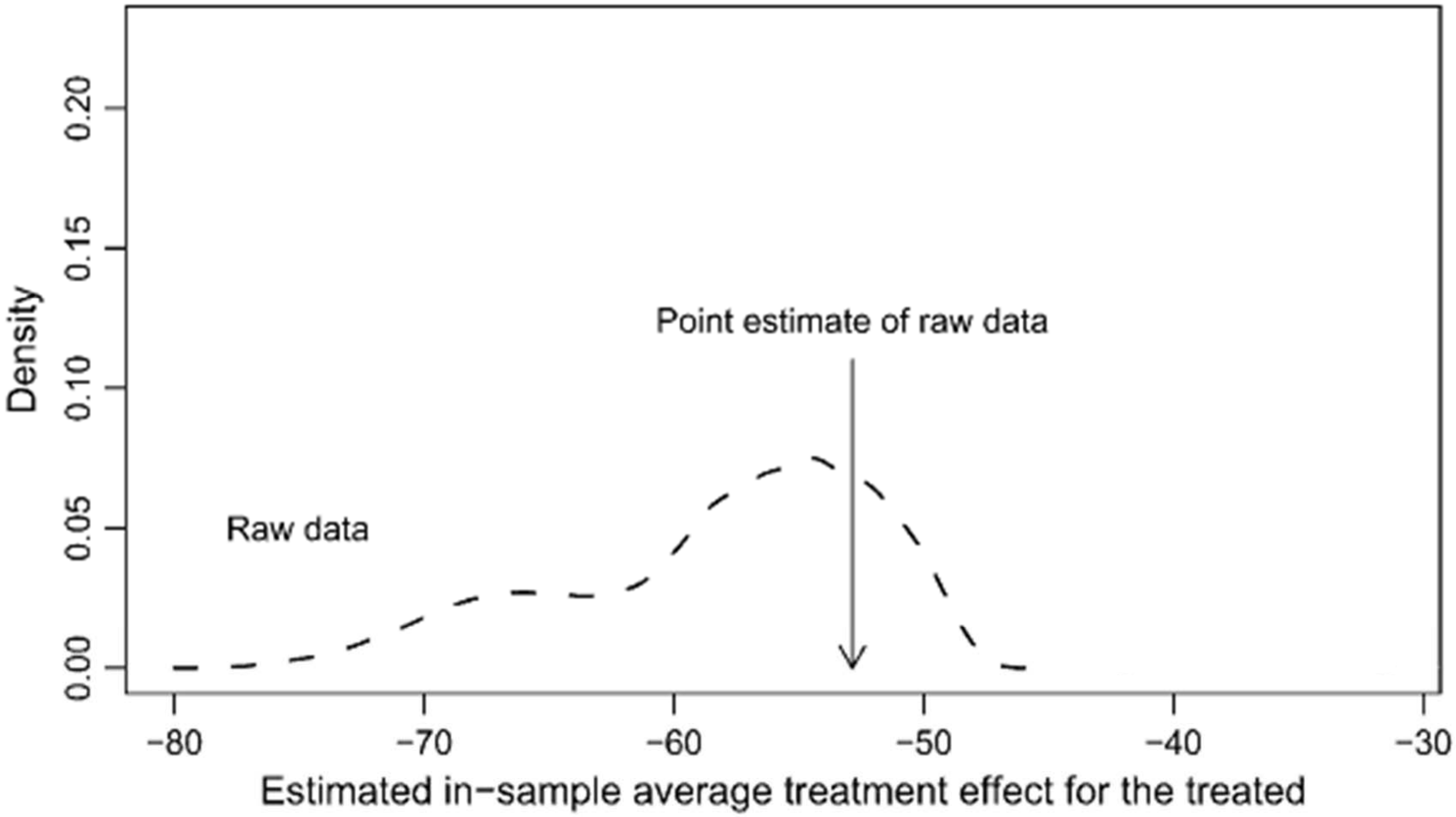


Fig. 2 Kernel density plot (a smoothed histogram) of point estimates of the in-sample ATT of the Democratic Senate majority on FDA drug approval time across 262,143 specifications. The solid line

Ho, 2007

Why do we need models?

The "treated"

The "controls"

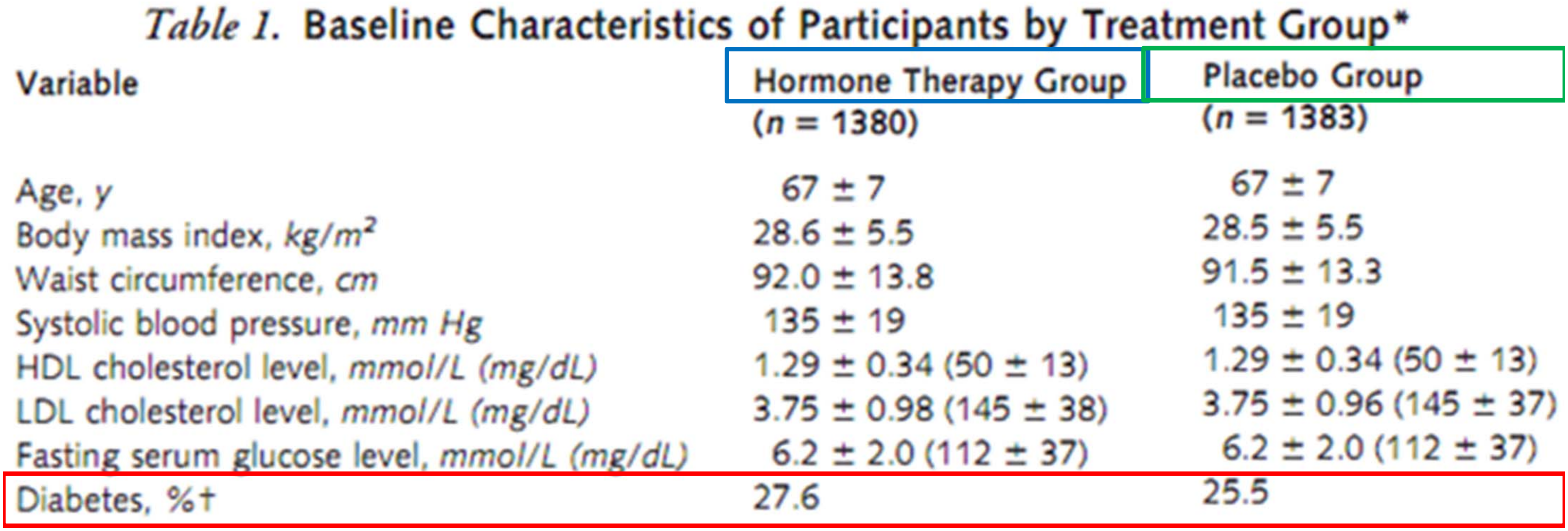


Table 1. Baseline Characteristics of Participants by Treatment Group*

Variable	Hormone Therapy Group (n = 1380)	Placebo Group (n = 1383)
Age, *y*	67 ± 7	67 ± 7
Body mass index, *kg/m²*	28.6 ± 5.5	28.5 ± 5.5
Waist circumference, *cm*	92.0 ± 13.8	91.5 ± 13.3
Systolic blood pressure, *mm Hg*	135 ± 19	135 ± 19
HDL cholesterol level, *mmol/L (mg/dL)*	1.29 ± 0.34 (50 ± 13)	1.29 ± 0.34 (50 ± 13)
LDL cholesterol level, *mmol/L (mg/dL)*	3.75 ± 0.98 (145 ± 38)	3.75 ± 0.96 (145 ± 37)
Fasting serum glucose level, *mmol/L (mg/dL)*	6.2 ± 2.0 (112 ± 37)	6.2 ± 2.0 (112 ± 37)
Diabetes, %†	27.6	25.5

But imbalance can arise even in randomized studies, due to finite samples, and this imbalance could result in confounding

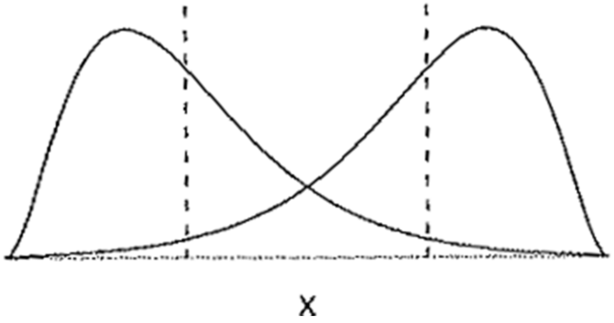
Kanaya, 2003

Covariate balance

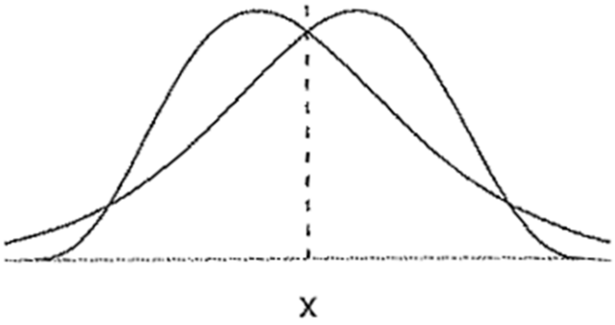
- When a covariate *X* does not differ on average between treatment groups, *X* is said to be “balanced”
    - i.e., distribution of *X* is identical between groups

- If *X* is balanced, this removes the possibility that *X* could confound

How model-dependent are our inferences?

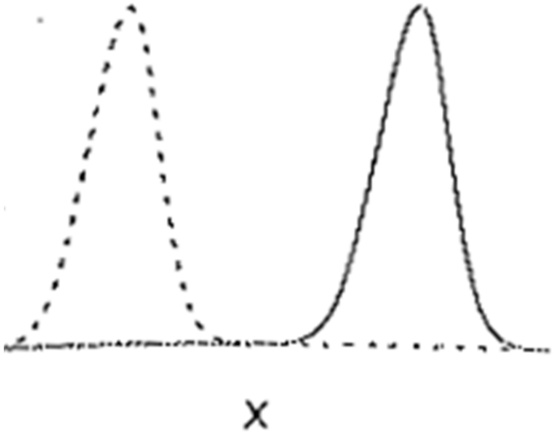


Severe imbalance, good overlap

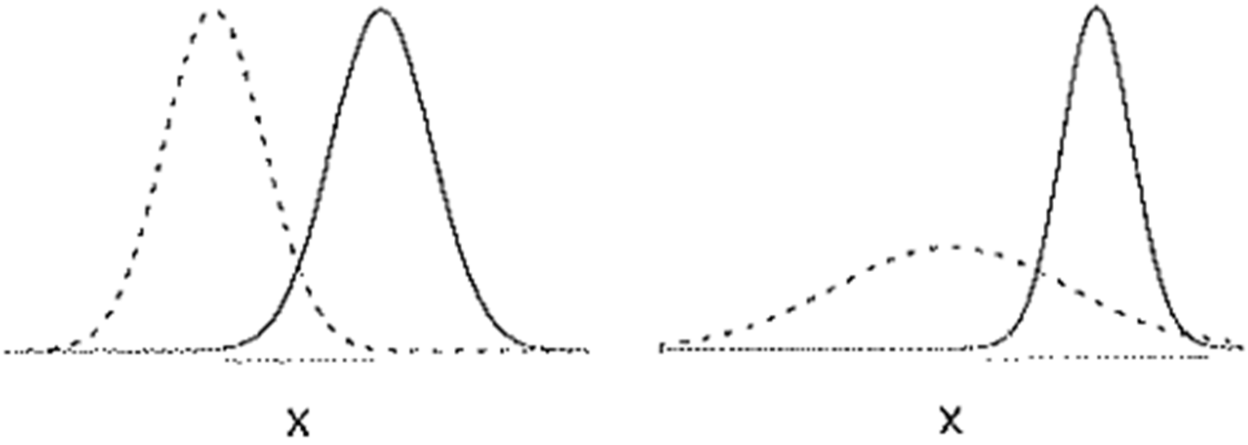


Slight imbalance, good overlap

Severe imbalance, no overlap



Moderate imbalance, partial overlap



From Gelman and Hill, 2003

Imbalance and lack of overlap

- Both forces us to rely more on the correctness of our model than we would have
- Interpolation, extrapolation, and residual confounding from observed covariates are all possible

PROPENSITY SCORE METHODS: CONCEPTS

Part 2: An introduction to matching

Brian Lee (bklee@drexel.edu)

Matching to reduce model dependence

- In case-control context
  - Subsetting cases and controls together on important covariates
- In a randomized experimental context
  - Subsetting treated units and control units together based on identical distributions of background characteristics
- In a more general context
  - Restricting the sample so that contrasting groups (either by treatment or outcome status) are more comparable to each other…in other words so that groups are *balanced*

Matching

- Matching attempts to replicate two features of randomized experiments
  - Create groups that look only randomly different from one another (at least on observed variables)

- Find treated/control units with similar covariate values

- Depends on the idea of sample restriction
  - not everybody in the sample is fit for analysis, so you restrict your analysis to those who can contribute meaningful data
  - clear parallel with the design of studies (e.g., who should I include in my study cohort and who should I exclude?)

Steps in implementing matching methods

1. Calculate the distance measure
   - Distance: the measure of how similar a treated is with a control unit

2. Match units using a method

3. Assess quality of matches
   - Iterate between steps 1 and 2 until have good matches

4. Estimate the treatment effect

Steps in implementing matching methods

1. Calculate the distance measure
   - Distance: the measure of how similar a treated is with a control unit

2. Match units using a method

3. Assess quality of matches
   - Iterate between steps 1 and 2 until have good matches

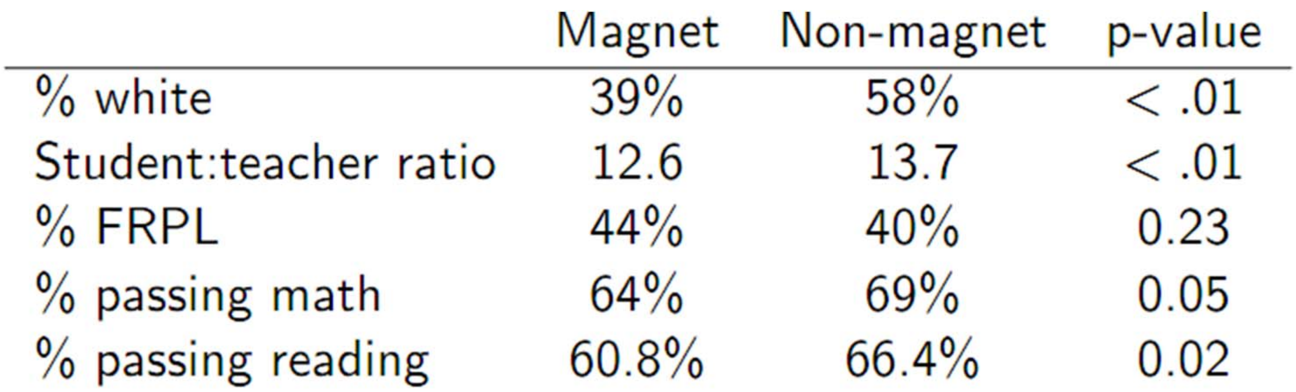
4. Estimate the treatment effect

Distance

- We want treated and control units to be as similar as possible
- Ideally, treated and control units match on the exact values of covariates *k*
  - E.g., race, sex, age..
  - In an infinite sample, is the ideal
  - But is impossible with continuous variables
  - “Coarsened exact matching” – match on ranges of variables
    - E.g., using income categories instead of a continuous measure

The curse of dimensionality

- National school-level dataset
- 55 elementary magnet schools; 384 non-magnet



	Magnet	Non-magnet	p-value
% white	39%	58%	$< .01$
Student:teacher ratio	12.6	13.7	$< .01$
% FRPL	44%	40%	0.23
% passing math	64%	69%	0.05
% passing reading	60.8%	66.4%	0.02

- Define variables based on quartiles
  - But even with just these 5 demographic variables with 4 levels each, only 35 schools have an exact match

- So what to do?

Stuart, 2007

- Instead of trying to match on multiple covariates at once, match on a single distance measure
- One distance measure: *the probability of treatment*
- Remember: in a randomized trial, treatment and control units both have equal probabilities of treatment

The propensity score

- Propensity score = Pr(*T*=1 | *X*)
    - The propensity score is the probability of receiving the treatment *T* conditional on the covariate(s) *X*
- Ranges from 0 to 1

Steps in implementing matching methods

1. Calculate the distance measure
   - Distance: the measure of how similar a treated is with a control unit

2. Match units using a method

3. Assess quality of matches
   - Iterate between steps 1 and 2 until have good matches

4. Estimate the treatment effect

Lots of possible matching algorithms

- Exact matching

- K:1 Nearest neighbor
  - With replacement
  - Without replacement
  - Greedy
  - Optimal
  - Caliper
  - Radius

  - …and more

Exact matching

- Match exactly on X covariates
    - Great with binary variables, e.g. sex

- Infeasible for more than several covariates

- So, use in combination with another matching algorithm

Nearest neighbor matching

- K:1 NN matching

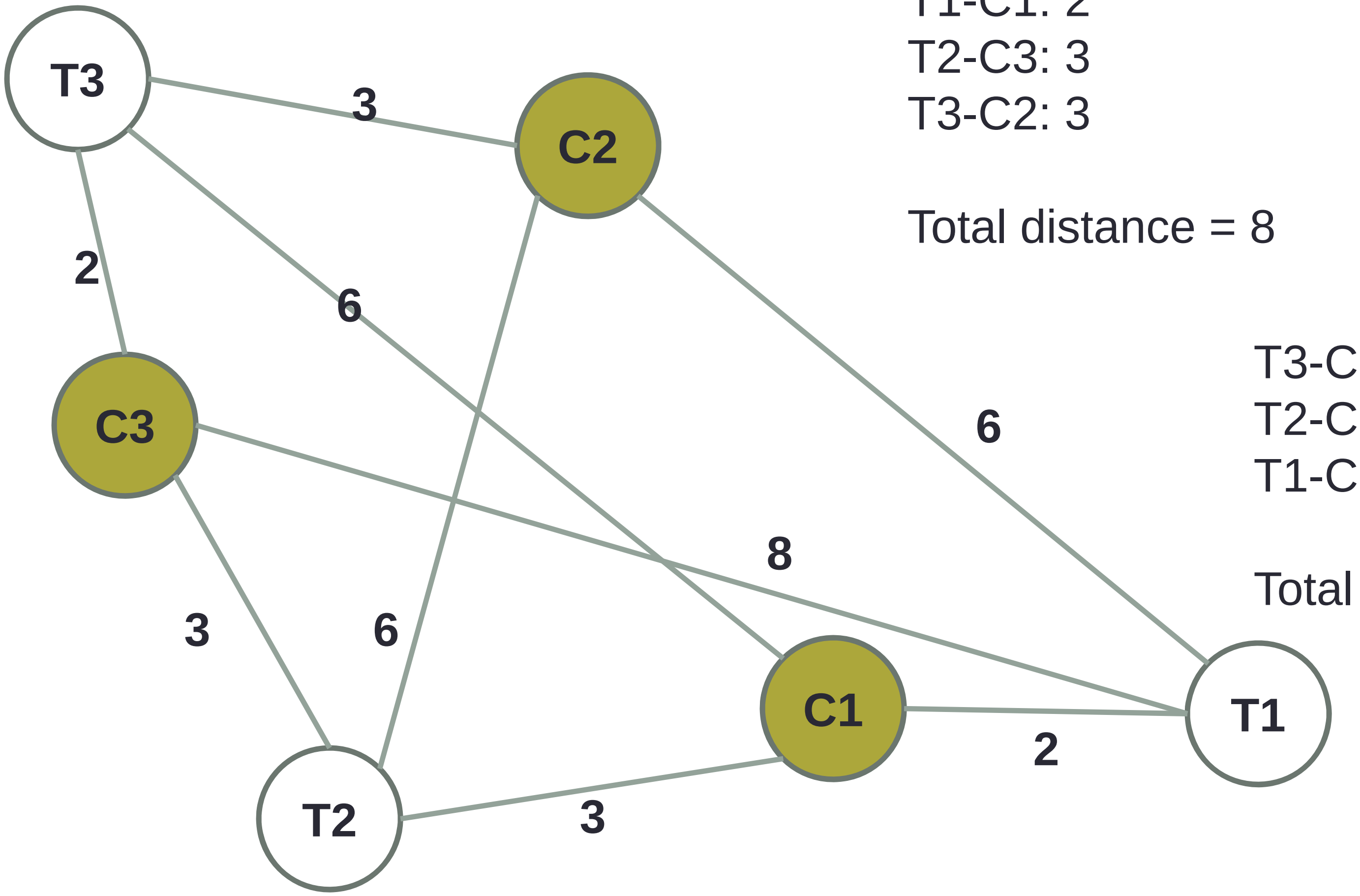
- Simplest form: 1:1 NN matching selects for each treated unit *i* the control unit with the smallest distance from *i*

- Can discard treated units as well
  - Especially if no reasonable controls exist

Nearest neighbor: greedy vs. optimal

- Simplest form – greedy matching
  - Once a match is made, it’s fixed
  - But the order that treated units are matched may affect quality of matches
- Greedy matching performs poorly when lots of competition for controls
- Optimal matching
  - Takes into account the overall set of matches when choosing individual matches, by minimizing the global distance measure

White: treated; Filled: controls



T1-C1: 2
T2-C3: 3
T3-C2: 3

Total distance = 8

T3-C3: 2
T2-C1: 3
T1-C2: 6

Total distance = 11

Nearest neighbor: with/without replacement

- Generally, we match without replacement (once a control is matched to a treated unit, it can't be selected again)

- If there are few control units that are comparable, may have to use a control as a match for multiple treated units
  - But this makes things more complicated
  - Need to account for weights
  - Intuitively: if a control is matched to 2 different treated units, the control is now counted twice and must receive a mathematical weight of 2 to signify this

Nearest neighbor: caliper and radius

- Nearest neighbor matching may yield some bad matches if there is not a good match nearby
  - Often happens at tails of the PS distribution, lower possibility of overlap with the other treatment group
- Can impose a caliper
  - Matches have to occur within a pre-defined distance
  - Rubin suggests 0.25 of SD of PS
- Or radius
  - One to many: take all matches within a pre-defined distance
- This can potentially discard some treated units, since there may not be controls that fall within the caliper/radius

Steps in implementing matching methods

1. Calculate the distance measure
   - Distance: the measure of how similar a treated is with a control unit

2. Match units using a method

3. Assess quality of matches
   - Iterate between steps 1 and 2 until have good matches

4. Estimate the treatment effect

Matching diagnostics

- Goal is to have similar covariate distributions in the matched treated/control groups
  - Therefore: assess quality of matching through checking covariate balance on the individual covariates
- One useful balance measure:

  ASAM – <u>*average standardized absolute mean distance*</u>

- If imbalance is found on particular variables, re-work the estimation of the distance measure or choose a different matching algorithm to improve balance in subsequent matched samples

UNMATCHED

	Mean (treated)	Mean (control)	SD (treated)	Standardized Mean Difference
Age	68.5	45.2	18.4	1.27
Male	0.49	0.44	0.50	0.10
Education	2.66	2.94	0.90	-0.31

ASAM: $\frac{1.27 + 0.10 + |-0.31|}{3}$ = 0.56

MATCHED

	Mean (treated)	Mean (control)	SD (treated)	Standardized Mean Difference
Age	68.5	68.6	18.4	-0.01
Male	0.49	0.48	0.50	0.02
Education	2.66	2.72	0.90	-0.07

ASAM: 0.03

Steps in implementing matching methods

1. Calculate the distance measure
   - Distance: the measure of how similar a treated is with a control unit

2. Match units using a method

3. Assess quality of matches
   - Iterate between steps 1 and 2 until have good matches

4. Estimate the treatment effect

Estimate the treatment effect

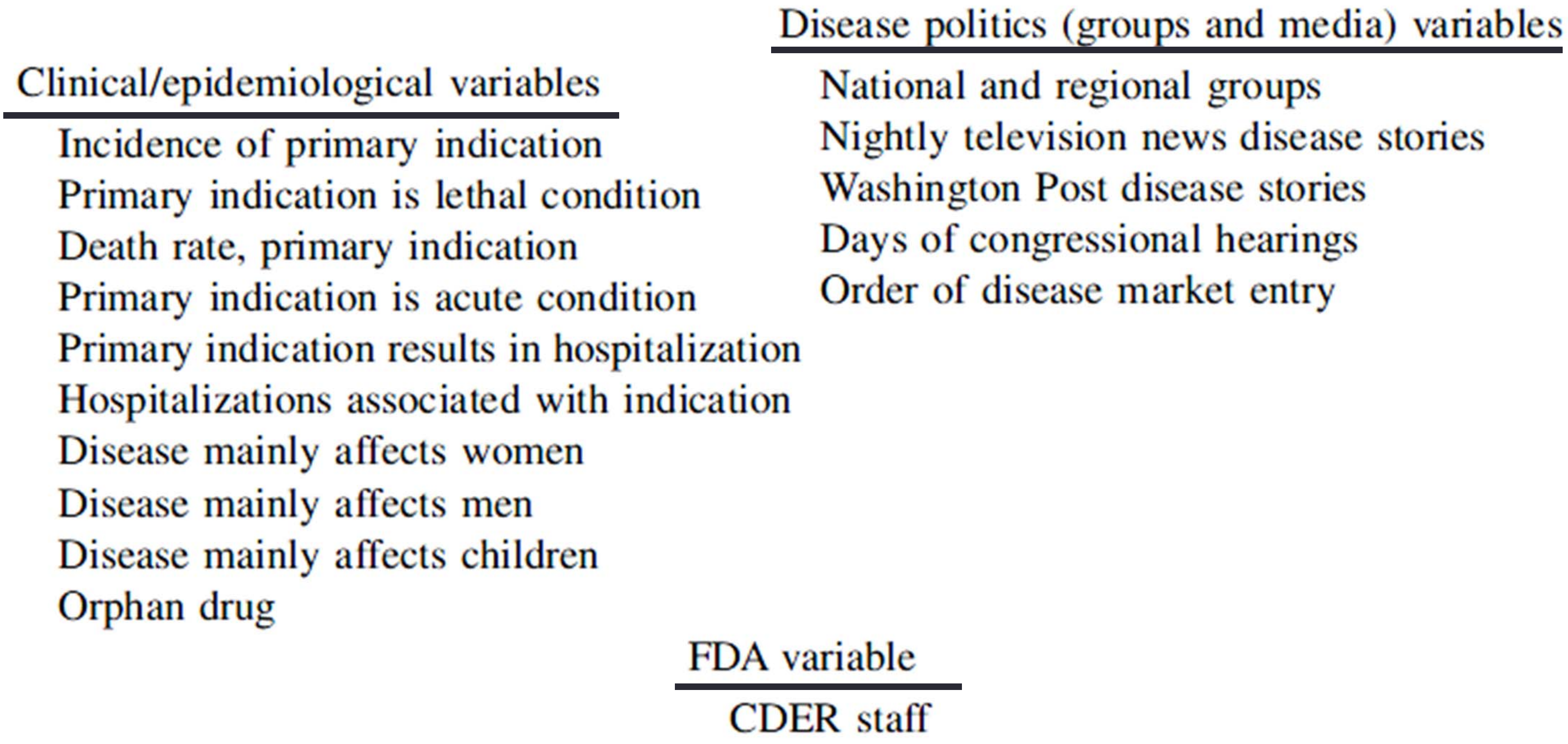
- After matching, use parametric model to adjust for residual imbalances
- This is considered to be “doubly robust” – two chances to remove confounding, once in the matching phase and again with the regression model
- After matching, effect estimates should depend less on the particular model used

Example: Democratic Senate majority and FDA Drug Approval Time

Does a Democratic majority (the treatment), compared with a Republican majority (the control), change the length of time it takes the FDA to approve a new drug?

Ho, 2007

The covariates



Clinical/epidemiological variables

- Incidence of primary indication
- Primary indication is lethal condition
- Death rate, primary indication
- Primary indication is acute condition
- Primary indication results in hospitalization
- Hospitalizations associated with indication
- Disease mainly affects women
- Disease mainly affects men
- Disease mainly affects children
- Orphan drug

Disease politics (groups and media) variables

- National and regional groups
- Nightly television news disease stories
- Washington Post disease stories
- Days of congressional hearings
- Order of disease market entry

FDA variable

- CDER staff

Examining how model dependence changes with matching

- 18 covariates to possibly include as linear predictors
- Ho et al. considered every possible combination of covariates, ignoring non-linearities and interactions!

N choose R	Combinations
(18, 1)	18
…	…
(18, 4)	3,060
…	…
(18, 9)	48,620
…	…
(18, 18)	1
TOTAL NUMBER OF COMBINATIONS	**262,143**

- Ho et al. examined model results from all 262,143 models in 2 different contexts

Raw data	Matched data
Take data-as-is	Take data-as-is
	Estimate propensity score from all 18 covariates
	Discard 15 control units and 2 treated units outside of common support of PS
	Match on PS
Run 262,143 models on data-as-is	Run 262,143 models on matched data

Matching reduces model dependence

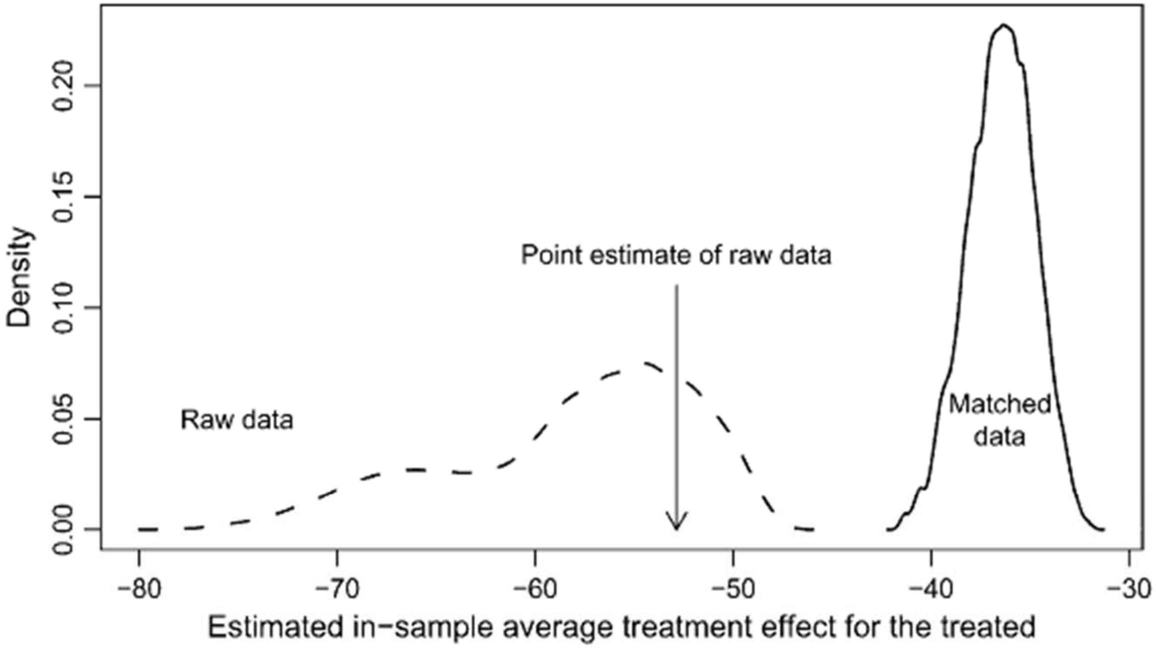


Fig. 2 Kernel density plot (a smoothed histogram) of point estimates of the in-sample ATT of the Democratic Senate majority on FDA drug approval time across 262,143 specifications. The solid line presents a density plot of the MLEs of ATT using the matched data set, whereas the dashed line is based on the raw data. The vertical arrow shows the point estimate from Carpenter's Model 1 based on the raw data. The estimate does not match Carpenter's estimate exactly because it is on a different scale and also because of the slightly different set of predictors used, as discussed above. The figure shows that ATT estimates are considerably more sensitive to model specification using the raw data as compared with the preprocessed matched data.

FAQs

- I’m uncomfortable with selectively removing observations from my dataset. I (spent so much money collecting the data / am very fond of the study subjects / am worried what reviewers will say)..
  - It is well-accepted to use procedures to test whether model estimates are sensitive to specific observations (e.g., DBETAs to systematically estimate parameters from a leave-one-out sample). Even the eyeball test to delete observations with significant leverage (potential to influence) is well-accepted as standard practice.

FAQs

- Does matching reduce statistical power since it will reduce the number of persons in the sample?
  - Not necessarily
  - Precision driven largely by smaller group size
  - Higher precision when comparing groups that are similar since less variance

FAQs

- How do I choose a matching algorithm?

    - Rely on your substantive knowledge but refer to balance measures
    - My philosophy: if there are specific variables that are extremely important to match on, do exact matching on those covariates and then propensity score match to adjust for other variables

References

- Breiman L. Statistical modeling: the two cultures. Statistical Science. 2001;16:199-215.
- Gelman A, Hill J. Data analysis using regression and multilevel/hierarchical models. New York, NY: Cambridge University Press; 2007.
- Ho DE, Imai K, King G, Stuart EA. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. Political Analysis. 2007;15:199-236.
- Ho DE, Stuart EA, Imai K, King G. MatchIt: MatchIt. R package version 2.3-1. 2007.
- Howe CJ, Cole SR, Westreich DJ, Greenland S, Napravnik S, Eron JJ, Jr. Splines for trend analysis and continuous confounder control. Epidemiology. 2011 Nov;22(6):874-5.
- Kanaya AM, Herrington D, Vittinghoff E, Lin F, Grady D, Bittner V, et al. Glycemic effects of postmenopausal hormone therapy: the Heart and Estrogen/progestin Replacement Study. A randomized, double-blind, placebo-controlled trial. *Ann Intern Med. 2003;138(1):1-9.*
- King G, Zeng L. When Can History Be Our Guide? The Pitfalls of Counterfactual Inference1. International Studies Quarterly. 2007;51(1):183-210.
- Lee BK, Glass TA, James BD, Bandeen-Roche K, Schwartz BS. Neighborhood psychosocial environment, apolipoprotein e genotype, and cognitive function in older adults. *Arch Gen Psychiatry. 2011;68(3):314-321*
- Stuart, E.A. and Rubin, D.B. (2007). Best Practices in Quasi-Experimental Designs: Matching methods for causal inference Chapter 11 (pp. 155-176) in Best Practices in Quantitative Social Science. J. Osborne (Ed.). Thousand Oaks, CA: Sage Publications.
- Westreich D, Cole SR. Invited Commentary: Positivity in Practice. Am J Epidemiol. 2010 Feb 5.