UC DAVIS HEALTH

Propensity Scores: Why, When, and How to Use Them

CLINICAL AND TRANSLATIONAL SCIENCE CENTER

Shuai Chen, Ph.D.

Assistant Professor of Biostatistics

UC Davis School of Medicine

The UC Davis CTSC receives support from the NIH National Center for Advancing Translational Sciences

Learning Objectives

- Understand the role of propensity scores in non-randomized studies.
- How to specify and estimate the propensity score model.
- How to use propensity scores to adjust for confounders when estimating the intervention effect.

Causation and Potential Outcomes

- Treatment X: The “intervention” that could apply or withhold

- Potential outcomes:
  - Potential outcome under treatment: outcome that would be observed if get treatment, Y(X = 1) = Y(1)
  - Potential outcome under control: outcome that would be observed if get control, Y(X = 0) = Y (0)
  - e.g., your headache pain in two hours if you take an aspirin: Y(1)

    your headache pain in two hours if you don’t take the aspirin: Y(0)

- Causal effects are comparisons of these potential outcomes Y(1) - Y(0)

“True” data (Only God Observes)

- e.g., effect of heavy adolescent drug use (X) on earnings at age 40 (Y)

Units	Y_i (1)	Y_i (0)
1	$15,000	$18,000
2	$9,000	$10,000
3	$10,000	$8,000
…		

- Causal effect for *i*th unit (individual) = Y_i (1) − Y_i (0)
- Average treatment effect (ATE) = $\frac{1}{N}\sum_{i=1}^{N}\{Y_i (1) - Yi (0)\}$
  - Average of Y_i (1) − Y_i (0) across individuals

Observed data

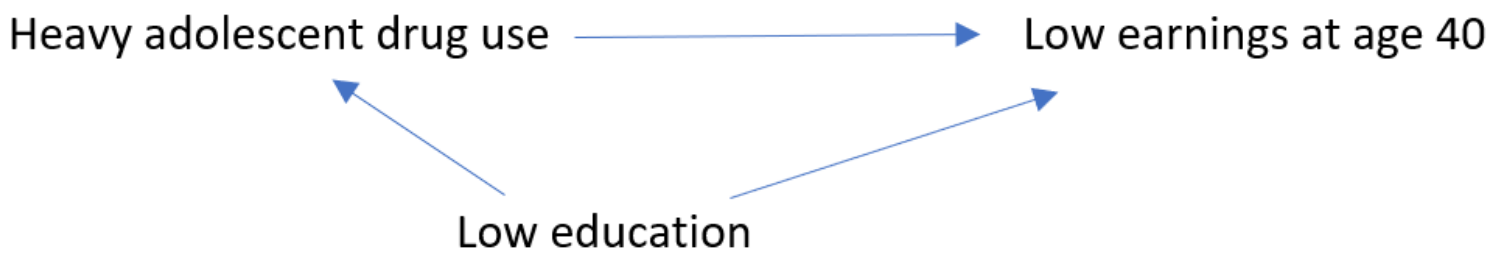
- e.g., effect of heavy adolescent drug use (X) on earnings at age 40 (Y)

Units	Y_i (1)	Y_i (0)
1	$15,000	?
2	?	$10,000
3	?	$8,000
…		

- Only observe Y_i (1) or Y_i (0) for each i
- **Association**: compares observed outcomes (average difference of observed outcomes)
- **Causation**: compares potential outcomes

Causation vs. Association

- Can Association lead to Causation? Potential selection bias!
  - E.g., education may be a confounder (for example only, may not be true):



- **Randomized experiments** are the “gold standard" of causal inference:
  - Remove any possible selection bias
    E.g., Education in both interventions (heavy drug use vs. not) is similar
  - Only difference between two groups is whether or not they receive the intervention
  - Not always feasible: Can’t randomize teenagers to become heavy drug users

Non-randomized Studies (Observational Studies)

- Main problem: People in “treatment” and “control” groups are likely different

- Traditional methods for non-randomized studies:
  - Stratification/matching
    - Put people into groups with same values of covariates (e.g., education)
    - Hard to adjust for many covariates this way
  - Regression analysis
    - e.g., linear regression of outcome given treatment and covariates
    - Predict earnings given covariates and heavy drug use:
      look at coefficient on heavy drug use

Non-randomized Studies (Observational Studies)

- Main problem: People in “treatment” and “control” groups are likely different

- Newer methods: Use propensity score methods to facilitate comparing similar individuals
  - Look for structure in the data that “mimics" a randomized experiment (pseudo-randomization)
  - Reduce selection bias by balancing confounders in “treatment” and “control” groups
- Propensity score $\pi(Z) = P(X = 1|Z)$
  - Probability of being assigned to “treatment” given covariates Z

Needed Assumptions for Propensity Score Methods

- **Stable Unit Treatment Value Assumption (SUTVA)**:
  - Consistency (If X=x, then Y=$Y(x)$), and
  - Non-interference (treating one individual does not affect any others)

- **Positivity** assumption: for any Z, $0 < \pi(Z) < 1$
  - If all low educated participants are heavy drug users, how can we reliably compare heavy drug users vs. non-heavy drug users among the low educated people?

Needed Assumptions for Propensity Score Methods

- **Conditional independence (Unconfoundness)** assumption:
  Assume that the treatment X is being randomly chosen, conditional on covariates Z,
  mathematically: $Y(x) \perp\!\!\!\perp X|\mathrm{Z}$
  - i.e., Z includes all confounders (no unobserved confounders)
  - Can help make this assumption more realistic if think about it during data collection

  - **This conditional independence also holds if conditional on propensity score only**,
    mathematically: $Y(x) \perp\!\!\!\perp X|\mathrm{Z} \rightarrow Y(x) \perp\!\!\!\perp X|\pi(Z)$
    (Propensity score is a summary score combining information from multiple Zs)

Propensity Scores as Summary of All Covariates

Propensity score can be viewed as a balancing score:

- At a given value of propensity score, distributions of observed covariates (that went into propensity score) are similar in treated and control groups
  - Conditional on propensity score, people in two groups are similar in covariates (like randomization)
  - If two people had the same probability of receiving treatment, and one did and one didn't receive treatment in reality (treated and comparison control), they should looks only randomly different on the observed covariates

Potential Dangers of Regression Adjustment on Full Samples

When treated and control groups have very different distributions of confounders (problem of small “common support”):

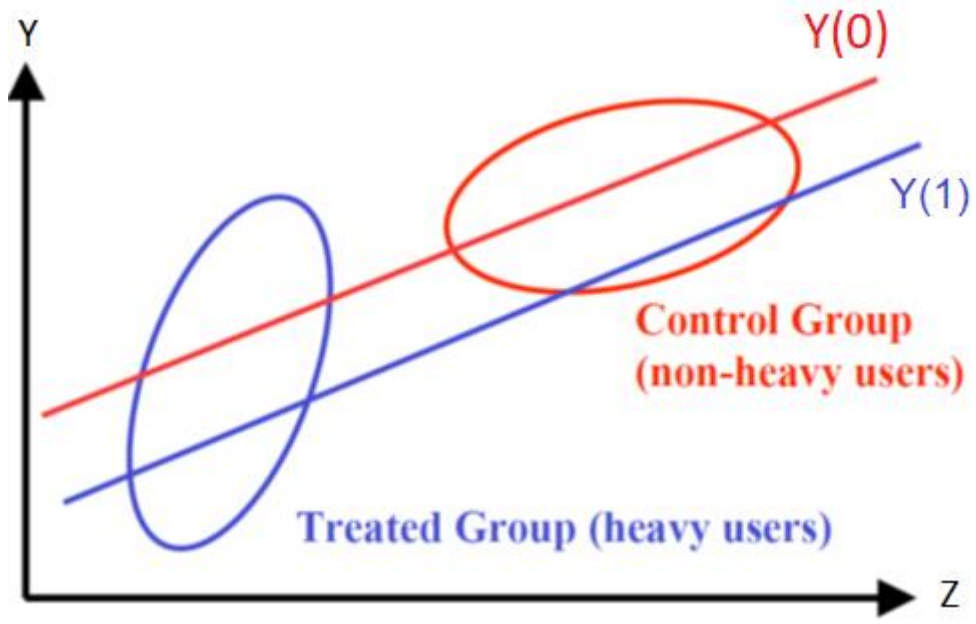
- May lead to bias if model mis-specified
  - Is the world really linear?
  - Several studies provided evidence that effect estimates are more sensitive to outcome regression model than to propensity score model

- Dangerous if simply fitting regression without checking whether the distributions of the confounders overlap.

Potential Dangers of Regression Adjustment on Full Samples

Example:

- Simple linear regression model used to estimate causal effects:
  $Y_i = \alpha + \tau X_i + \beta Z_i + e_i$, $e_i \sim N(0, \sigma^2)$
  - $\hat{\tau}$ is estimate of treatment effect
  - Treatment X=1 or 0
  - Z is a confounder adjusted in regression
- This assumes parallel linear regression lines about potential outcomes:
  - $Y_i(0) = \alpha + \beta Z_i + e_i$
  - $Y_i(1) = \alpha + \tau + \beta Z_i + e_i$
  - $Y_i(1) - Y_i(0) = \tau$



Simply Adjust Propensity Score in Regression Directly?

- What about simply including propensity score in outcome regression model?
  - Propensity scores are sometimes used as a predictor in regression (simply replacing all of the individual covariates)
  - **Not recommended**: If samples unbalanced on covariates, will also be unbalanced on propensity score

Better Methods Using Propensity Scores (Details Later)

- **Matching**
  - For each treated individual, select k controls with similar propensity scores (often k=1)
  - Easier to match just on propensity score, rather than all covariates individually

- **Stratification/ Subclassification**
  - Group individuals into groups with similar propensity score values
  - Often 5 subclasses used (quintiles of propensity scores)
  - Outcome analyses often use subclass as strata (e.g., evaluate intervention effects within strata and then combine)

- **Weighting adjustments**
  - Inverse probability of treatment/exposure weights (IPTW)

How to estimate Propensity Score?

- Propensity Score = probability of treatment receipt given covariates = $P(X = 1|Z)$
  - Most common: logistic regression
    - Dependent variable = treatment indicator X (0 or 1)
    - Independent variables = covariates
  - Non-parametric options using machine learning techniques: classification and regression trees (CART), random forest, GBM, etc.

- Propensity scores are predicted probability for each person from these models, but
  - **Mainly care about whether it results in balanced samples**
    - Can easily check this!
  - Don't care much about predictive ability of model
  - Don't care about interpretation of covariates: only need predicted probabilities
  - If using **propensity scores weighting**, need to care somewhat about accuracy of the predicted propensity scores themselves

Select Variables into Propensity Score Model

- Main idea: Select variables related to treatment receipt and the outcomes

- Trade-offs involved:
  - Excluding confounders may violate uncounfoundedness assumption
  - Including too many unnecessary covariates Z's can exacerbate problem of "common support" and increase variance

- Suggestion by Stuart (2011):
  - For large samples, be generous in what you include and err on including more rather than less
  - For small samples (e.g., 100), focus on variables believed to be **strongly related to outcome**

Matching

Traditional Matching:

- Find treated and control individuals with similar covariate values
  - Might be hard to get matches on all covariates separately

Propensity Score Matching:

- Find treated and control individuals with similar propensity scores
- What if a treated individual just doesn't have any controls with similar propensity scores?
  - Can impose a "caliper": limits matches to be within some range of propensity score values (often 0.25 or 0.5 propensity score standard deviation)

Diagnostics for Propensity Score Matching

Main idea: Compare the covariate distributions between matched treated and controls

- Histograms of propensity scores and covariates
- Descriptive statistics
  - e.g., means of covariates, variances of covariates
- Standardized difference
  - Difference in means (or probabilities for binary variable) between two groups, divided by pooled standard deviation

$$D = \frac{\bar{Z}_{treat} - \bar{Z}_{control}}{\sqrt{\frac{\sigma^2_{treat} + \sigma^2_{control}}{2}}} \text{ (continuous Z), or}$$

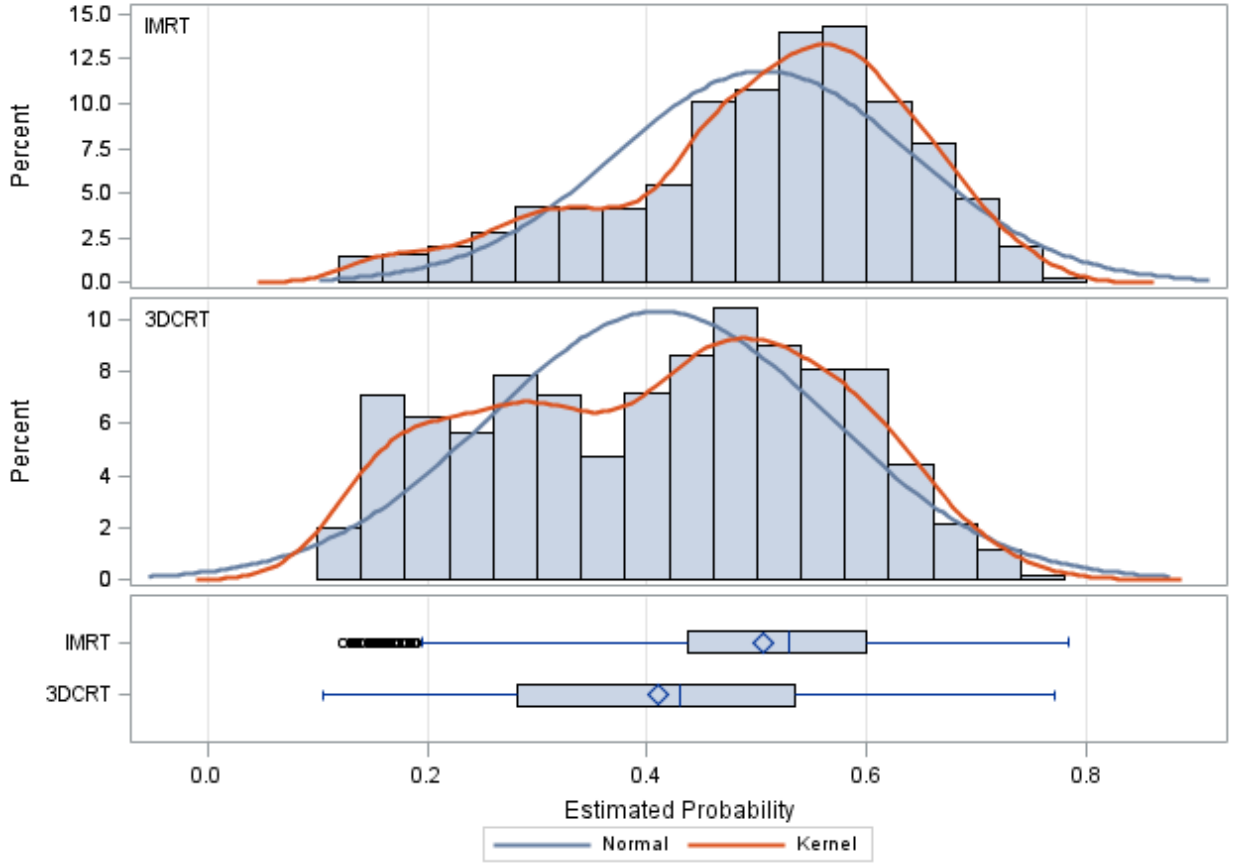
$$D = \frac{\hat{p}_{treat} - \hat{p}_{control}}{\sqrt{\frac{\hat{p}_{treat}(1-\hat{p}_{treat}) + \hat{p}_{control}(1-\hat{p}_{control})}{2}}} \text{ (binary Z)}$$

  - See how much smaller it is after matching
  - Standardized difference < 0.2 (or even smaller) is often considered as small

Propensity Scores (Pre-Match) from NCDB Data

Example: Lung cancer patients from US registry data (National Cancer Database)
Treatment = IMRT vs 3D-CRT, Propensity score $\pi(Z)=P(\text{X="IMRT"}|Z)$

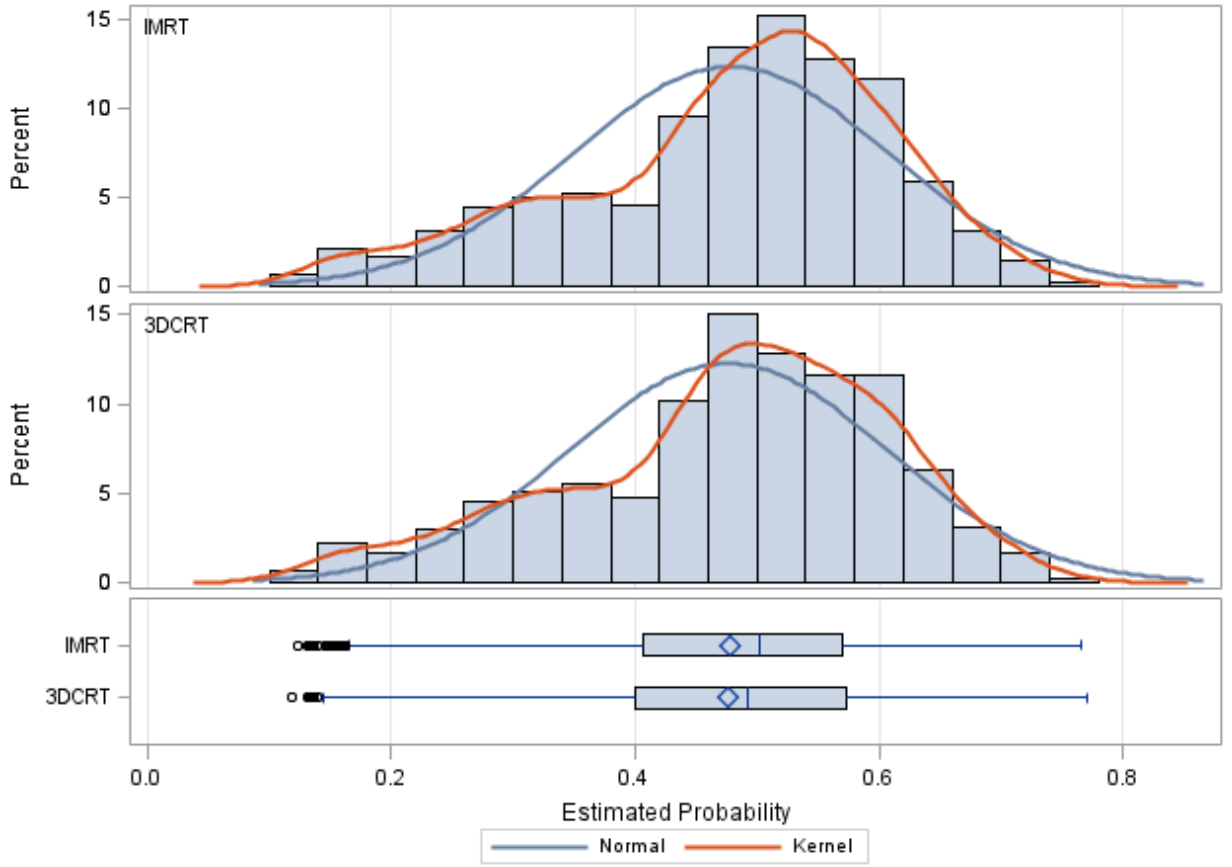
Distribution of Estimated Propensity Scores (Pre-Match)



Propensity Scores (Post-Match) from NCDB Data

Example: Lung cancer patients from US registry data (National Cancer Database)
Treatment = IMRT vs 3D-CRT, Propensity score $\pi(Z)=P(X=\text{“IMRT”}|Z)$

Distribution of Estimated Propensity Scores (Post-Match)



Improved Covariate Balance After Matching

	Pre-Matching		Post-Matching	
	3D-CRT (N=1,888)	IMRT (N=1,556)	3D-CRT (N=1,238)	IMRT (N=1,238)
Age, median (range)	66 (28-90)	67 (34-90)	67 (40-90)	67 (40-90)
Sex				
Male	1,035 (54.8%)	831 (53.4%)	689 (55.5%)	676 (54.4%)
Female	853 (45.8%)	725 (46.6%)	553 (44.5%)	566 (45.6%)
Race				
White	1,600 (85.8%)	1,312 (84.6%)	1,058 (85.2%)	1,060 (85.3%)
Non-white	264 (14.2%)	239 (15.4%)	184 (14.8%)	182 (14.7%)
Charlson/Deyo comorbidity score				
<2	1,692 (89.6%)	1,390 (89.3%)	1106 (89.1%)	1107 (89.1%)
≥2	196 (10.4%)	166 (10.7%)	136 (10.9%)	135 (10.9%)
Insurance type				
Private	616 (33.1%)	490 (32.0%)	408 (32.1%)	386 (31.1%)
Other	1243 (66.9%)	1043 (68.0%)	834 (67.2%)	856 (68.9%)
Income				
<$48,000	795 (42.9%)	683 (44.7%)	552 (44.4%)	548 (44.1%)
≥$48,000	1,060 (57.1%)	844 (55.3%)	690 (55.6%)	694 (55.9%)
Location				
Urban	1263 (68.0%)	1047 (68.5%)	830 (66.8%)	834 (67.1%)
Non-urban	594 (32.0%)	482 (31.5%)	412 (33.2%)	408 (32.9%)
Facility				
Academic	548 (29.1%)	643 (41.4%)	417 (33.6%)	427 (34.4%)
Non-Academic	1,332 (70.9%)	909 (58.6%)	825 (66.4%)	815 (65.6%)
Histology				
Adenocarcinoma	626 (33.2%)	648 (41.6%)	466 (37.5%)	465 (37.4%)
Other	1262 (66.8%)	908 (58.4%)	776 (62.5%)	777 (62.6%)

Pre/Post-PS Matching Standardized Differences

	Pre-Matching (N_{IMRT}=1556 $N_{3D\text{-}CRT}$=1888)	Post-Matching (N_{IMRT}=1238 $N_{3D\text{-}CRT}$=1238)
Age (continuous)	0.010	0.008
Gender (Male vs Female)	0.028	0.010
Race (White vs Non-white)	0.035	0.021
Charlson Score (≥2 vs <2)	0.009	0.023
Insurance (Private vs Other)	0.025	0.007
Income (≥$48,000 vs <$48,000)	0.038	0.016
Location (Urban vs Non-urban)	0.010	0.045
Facility (Academic vs Non-academic)	0.259	0.003
Histology (Adenocarcinoma vs Other)	0.176	0.005

(Some patients cannot find similar match and were discarded)

Outcome Analysis after Matching

Matches generally pooled together into just “treated” and “control” groups:

- Can run the same outcome analyses for matched data (e.g., t-test)
  - just like randomized studies

- Don’t need to account for match pairs
  - Suggested by Schafer and Kang, but there are some argument that standard error needs to be corrected to account for match pairs

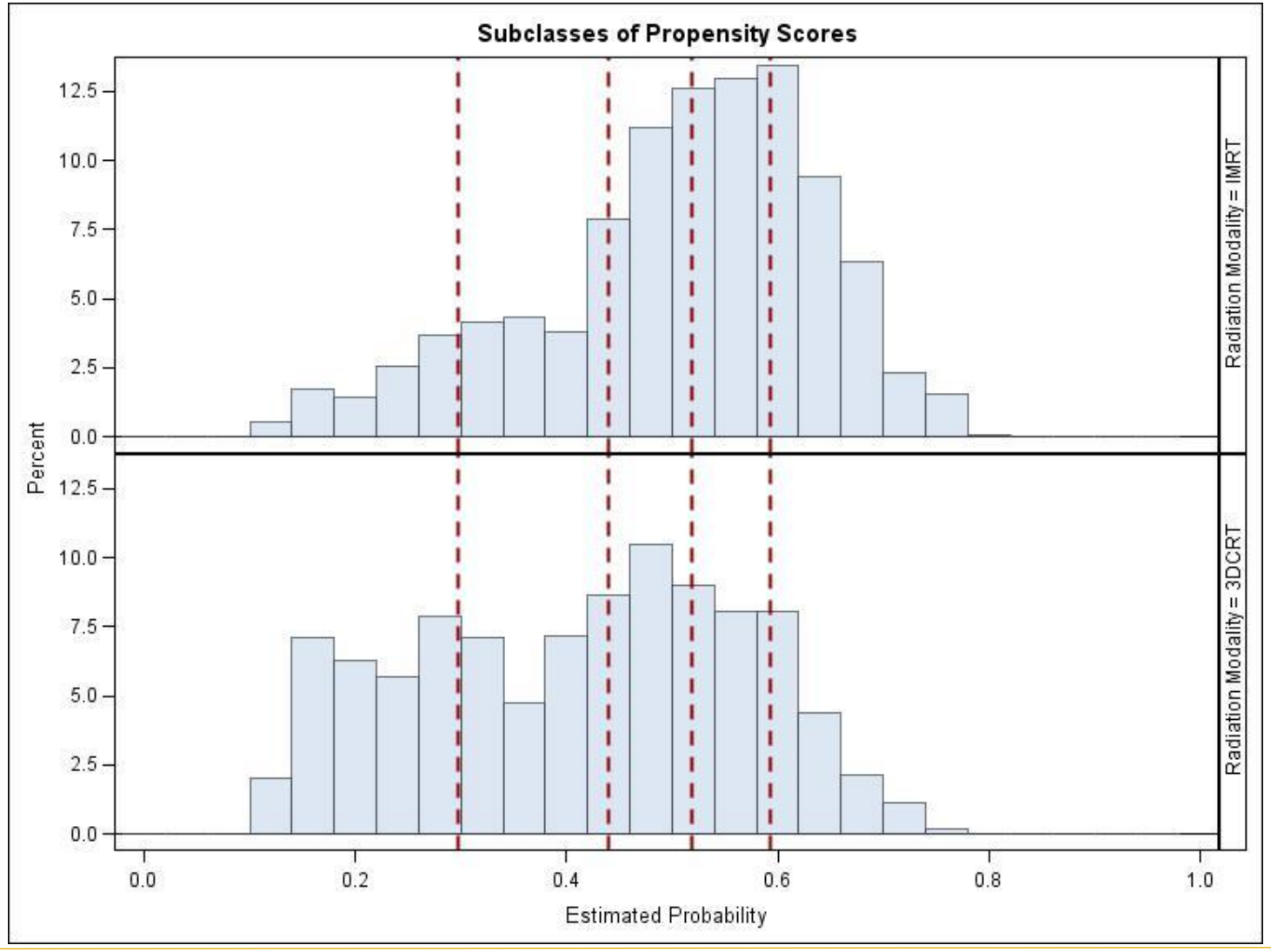
Subclassification/Stratification

- Creates subclasses of individuals with similar propensity score values
  - Often 5 subclasses (quintiles of propensity scores)
  - With large sample sizes, can use more than 5 subclasses
  - Ensuring enough treated and control in each subclass

- Within each subclass,
  - individuals have a similar probability of receiving treatment
  - should look similar on covariates between treated and control

- Use all individuals in data (not discarding lots data like matching)
- Diagnostics can use stratified analysis to examine covariate balance between treated and control (e.g., Cochran–Mantel–Haenszel test)

Subclassification/Stratification



Subclassification/Stratification

- Individuals across subclasses may look different
  - Patients in academic facility with adenocarcinoma are more likely to choose IMRT

	Subclass 1		Subclass 3		Subclass 5	
	3D-CRT N=515	IMRT N=143	3D-CRT N=350	IMRT N=309	3D-CRT N=234	IMRT N=425
Age, Median (range)	67 (41-90)	66 (42-86)	67 (40-90)	68 (42-89)	65 (41-89)	65 (40-88)
Sex						
Female	228 (44.3%)	58 (40.6%)	158 (45.1%)	140 (45.3%)	114 (48.7%)	231 (54.4%)
Male	287 (55.7%)	85 (59.4%)	192 (54.9%)	169 (54.7%)	120 (51.3%)	194 (45.6%)
Race						
White	453 (88.0%)	127 (88.8%)	300 (85.7%)	267 (86.4%)	193 (82.5%)	342 (80.5%)
Non-white	62 (12.0%)	16 (11.2%)	50 (14.3%)	42 (13.6%)	41 (17.5%)	83 (19.5%)
Location						
Urban	365 (70.9%)	99 (69.2%)	235 (67.1%)	202 (65.4%)	164 (70.1%)	319 (75.1%)
Non-urban	150 (29.1%)	44 (30.8%)	115 (32.9%)	107 (34.6%)	70 (29.9%)	106 (24.9%)
Facility						
Academic	73 (14.2%)	14 (9.8%)	59 (16.9%)	43 (13.9%)	174 (74.4%)	331 (77.9%)
Non-Academic	442 (85.8%)	129 (90.2%)	291 (83.1%)	266 (86.1%)	60 (25.6%)	94 (22.1%)
Histology						
Adenocarcinoma	100 (19.4%)	29 (20.3%)	110 (31.4%)	107 (34.6%)	138 (59.0%)	270 (63.5%)
Other	415 (80.6%)	114 (79.7%)	240 (68.6%)	202 (65.4%)	96 (41.0%)	155 (36.5%)

Outcome Analysis After Subclassification

- Main idea:
  1. Calculate effect within each subclass
  2. Then average effects across subclasses

- Possible methods, e.g.,
  - Simple t-test (or other analysis you would do) within each subclass, and combine
  - Regression adjustment within each subclass, and combine
    - Adjust for small differences within subclasses
  - Regression adjustment using all individuals together, including subclass and treatment×subclass interactions

Subclassification/Stratification analysis from NCDB data

Results for Survival Rate in Each Propensity Score Subclass

Subclass	Treatment	3-year Survival Rate	Difference (3D-CRT - IMRT)	SE of Difference
1	3D-CRT IMRT	23.9% 24.4%	-0.5%	4.1%
2	3D-CRT IMRT	29.3% 30.8%	-1.6%	3.9%
3	3D-CRT IMRT	31.5% 29.1%	2.4%	4.2%
4	3D-CRT IMRT	26.9% 33.9%	-7.0%	4.4%
5	3D-CRT IMRT	33.3% 34.9%	-1.6%	4.8%

Calculating Overall Effects

- Overall average treatment effect is weighted average of subclass-specific effects
  - Weighted by sample sizes of subclasses
- Variance of the estimate is calculated as weighted summation of subclass-specific variances

- e.g., if subclass is determined by quintiles of propensity scores
  - Take average of 5 subclass-specific effects to get an overall average treatment effect

$$\text{Effect}_{\text{overall}} = \frac{1}{5}\sum_{i=1}^{5} \text{Effect}_i$$

    - For NCDB data, overall difference in 3-year survival rate

$$= \frac{1}{5}(-0.5\% - 1.6\% + 2.4\% - 7.0\% - 1.6\%) = -1.7\%$$

  - Variance of this estimated effect is

$$\text{Variance}_{\text{overall}} = \frac{1}{5^2}\sum_{i=1}^{5} \text{Variance}_i$$

    - For NCDB data, overall standard error of the difference

$$= \frac{1}{5}\sqrt{(4.1\%)^2+(3.9\%)^2+(4.2\%)^2+(4.4\%)^2+(4.8\%)^2} = 1.9\%$$

Inverse Probability of Treatment/Exposure Weights (IPTW)

- Use propensity score $\pi(Z)$ to weight treated and control groups back to the whole population
- Like survey sampling weights
  - Treated group: weight = $\frac{1}{\pi(Z)}$
  - Control group: weight = $\frac{1}{1-\pi(Z)}$
- e.g., a treated individual with $\pi(Z)$=0.2
  - 20% probability to be assigned to treatment
  - get weight 1/0.2 = 5, representing 5 potential people in whole population
- e.g., a control with $\pi(Z)$=0.666
  - 1-0.666 = 33% probability to be assigned to control
  - get weight 1/0.333 = 3, representing 3 potential people in whole population

Extreme Weights

- Weights can be extreme and lead to unstable results
  - E.g., a treated individual with estimated propensity score 0.01
    $\rightarrow$ weight$=1/0.01=100$

- Most widely used solution is Weight Trimming:
  - Set a maximum value for weights (e.g., 10). If a weight>10, trims it to 10

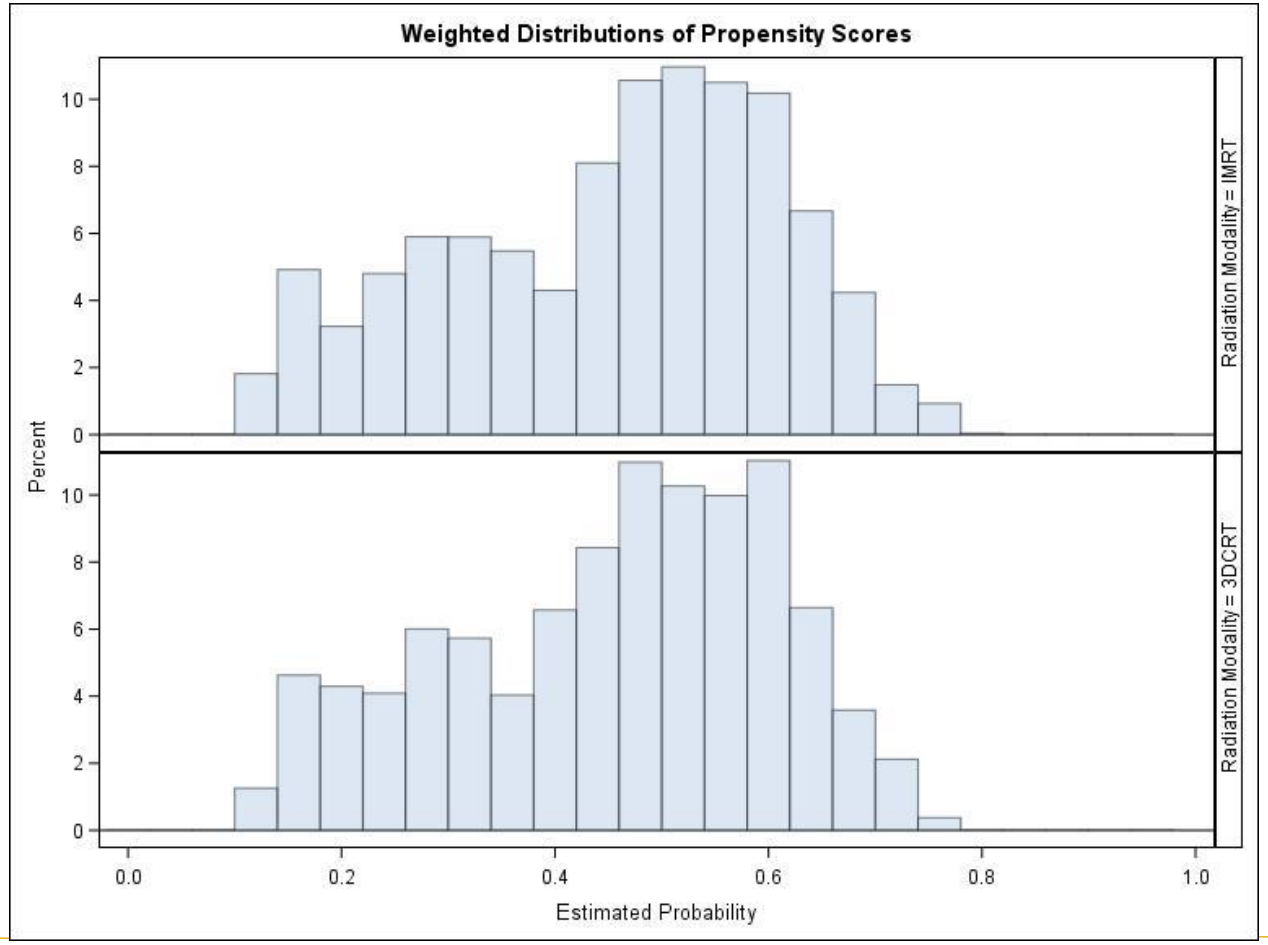
Outcome Analysis After Weighting

- Treat the weights like sampling weights:
  - Perform weighted outcome analysis
  - E.g., weighted t-tests to compare outcomes in treated vs. control
  - May correct standard error to account for uncertainty in estimated propensity score (e.g., formula by Williamson et al. 2013), but more complicated

- Diagnostics to examine covariate balance between treated and control:
  - Can compare the weight-adjusted covariate distributions
  - e.g., weighted t-test to compare weighted means of covariates

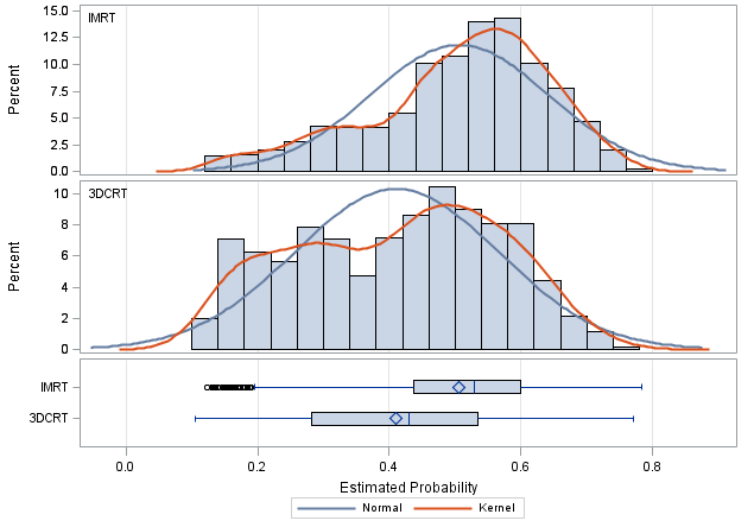
Example: Examining Balance of Propensity Scores

- Propensity scores are more balanced after weighting
- Can check the balance of covariates similarly

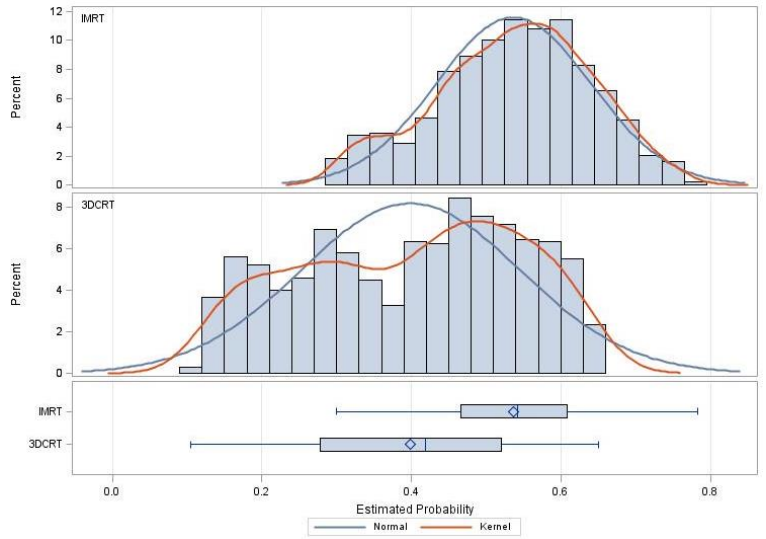


Restricting Analyses to Common Support of Propensity Scores

- For all propensity score methods, ensure individuals are comparable
  - Remember to check histograms of propensity scores to make sure sufficient overlap between treated and control
- Sometimes it may make sense to restrict analyses to only those individuals with propensity scores that overlap with the other group, e.g.,
  - Drop controls with propensity score < min(propensity scores in treatment)
  - Drop treated individuals with propensity score > max(propensity scores in control)



versus



Software for Propensity Score Methods

- Many propensity score tasks don't require special software, e.g.,
  - estimating propensity scores using logistic regression
  - doing propensity score-weighted outcome analysis

- Matching methods require specialized software/package. Some examples:
  - R: MatchIt package (http://gking.harvard.edu/matchit), etc.
  - Stata: psmatch2, etc.
  - SAS: PSMATCH (introduced in SAS/STAT v14.2), and some user-written macros
  - These functions/package also include other propensity score methods such as subclassification and weighting.

Discussion

- Important to control for confounding in non-randomized studies
- Benefits of using propensity scores
  - Force you to see the amount of overlap ("balance") in the data
    - standard regression diagnostics don't show this
  - Clear diagnostics of the use of propensity scores on balance
- Whenever estimating causal effects using non-randomized data, good to estimate propensity scores
  - Even if don't end up using them in analysis, good to use them to do some diagnostics for covariate balance
- If you do use them, ensures comparison of similar individuals (reduced confounding)
- Could combine one of the three propensity scores approaches with regression adjustment

References

- Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. Biometrika 70, 41-55.
- Lunceford, J. K. and Davidian, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. Statistics in Medicine 23, 2937-2960.
- Schafer, J. L., & Kang, J. (2008). Average causal effects from nonrandomized studies: a practical guide and simulated example. Psychological methods, 13(4), 279.
- Lee, B. K., Lessler, J., & Stuart, E. A. (2010). Improving propensity score weighting using machine learning. Statistics in medicine, 29(3), 337–346.
- Ho D, Imai K, King G, Stuart E (2007). Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference. Political Analysis, 15(3), 199–236.
- Williamson Elizabeth J, Forbes A, Ian RW. (2013). Variance reduction in randomised trials by inverse probability weighting using the propensity score. Statistics in Medicine 33(5),721–37.

Help is available

- CTSC and Cancer Center Biostatistics Office Hours
  - Every Tuesday from 12 – 2:00 currently via WebEx
  - 1st & 3rd Monday from 1:00 – 2:00 currently via WebEx
  - Sign-up through the CTSC Biostatistics Website
- EHS Biostatistics Office Hours
  - Upon request
- Request Biostatistics Consultations
  - CTSC
  - MIND IDDRC
  - Cancer Center Shared Resource
  - EHS Center

Questions?