Propensity score methods for comparing multiple treatment options

Melissa Garrido, PhD

April 28, 2021

VA | U.S. Department of Veterans Affairs
Veterans Health Administration
Office of Research & Development

PEPReC
Partnered Evidence-based Policy Resource Center
A VA QUERI Center

QUERI
VA Quality Enhancement Research Initiative
EVIDENCE INTO PRACTICE

Acknowledgements

- Jessica Lum
- Yevgeniy Feyman
- Steve Pizer

- HSRD IIR 16-140

- Garrido, Lum, Pizer. Vector-based kernel weighting: A simple estimator for improving precision and bias of average treatment effects in multiple treatment settings. Statistics in Medicine 2021; 40(5): 1204-1223.

Overview

- Illustrate challenges of comparing multiple treatment options in observational studies
- Outline best practices for using propensity scores to compare the effects of multiple treatments
- Introduce vector-based kernel weighting (VBKW)
  - Produces estimates with low bias and high efficiency
  - Straightforward to implement

Poll Question

- How familiar are you with propensity score analyses?
  - Not at all familiar
  - Somewhat familiar
  - Very familiar

Propensity Scores Can Address Selection Bias by “Pre-Processing” Datasets

- Goal: Make treatment and comparison group as similar as possible on observed confounders before proceeding with analysis

- Pre-processing methods include exact matching, coarsened exact matching, propensity scores, and entropy balancing

Ho et al. 2007. Political Analysis 15: 199-236
Stuart 2010. Statistical Science 25: 1-21.

Propensity Scores: Brief Overview

- Create a single composite score of all observed, measured potential confounders of the association between treatment and outcome

- Propensity score is the conditional probability of treatment given the observed covariates X

$$E(X) = P(D=1 \mid X)$$

- Match or weight on this one-dimensional score alone

- Do this without knowledge of the outcome variable

Propensity Score Assumption: Strongly Ignorable Treatment Assignment

- Given a set of covariates:
  - Treatment assignment and outcome are independent
  - Everyone has a nonzero chance of receiving the treatment

Rosenbaum & Rubin 1983. Biometrika 70: 41-45

What happens when you have multiple treatment options?

- Not all treatment decisions are binary (treated vs untreated)
- Can be continuous or categorical

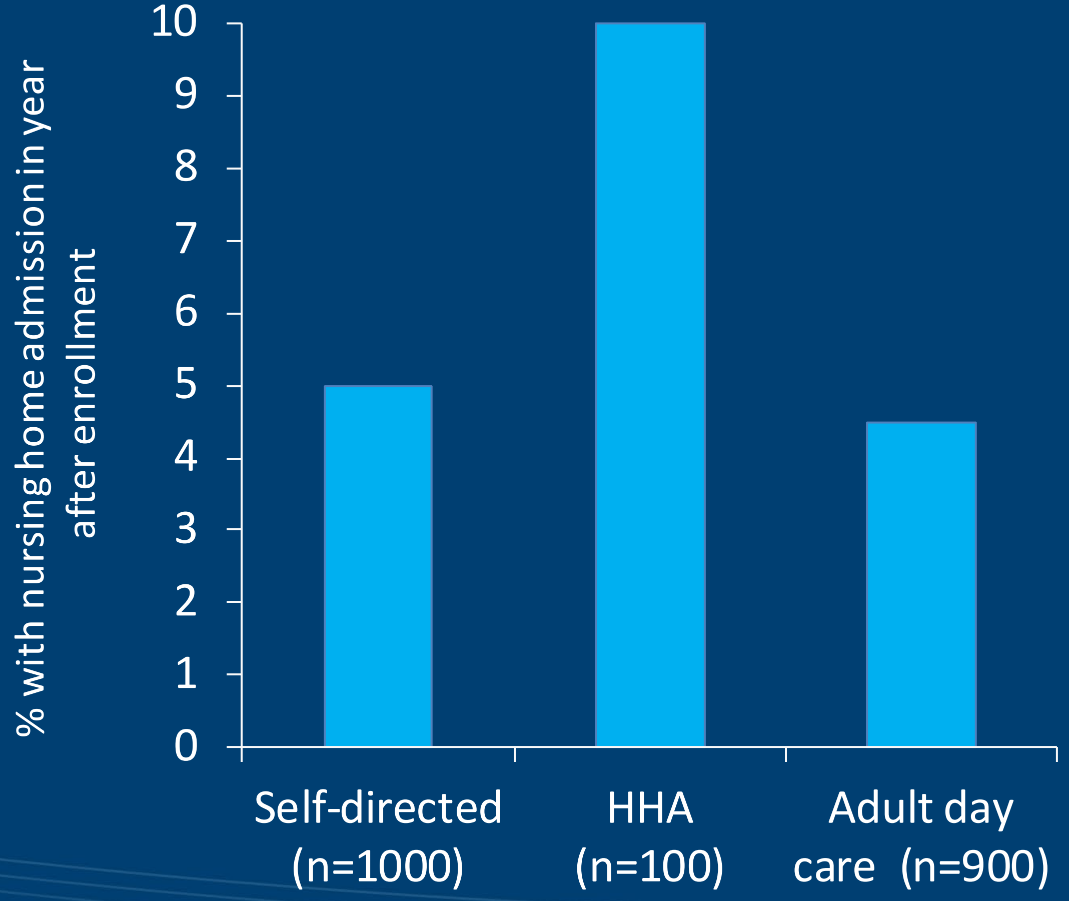
- Examples of categorical treatments:
  - Multiple vaccines for COVID-19
  - Inpatient hospice vs outpatient hospice vs no hospice
  - Self-directed home and community based services vs home health aide services vs adult day care services

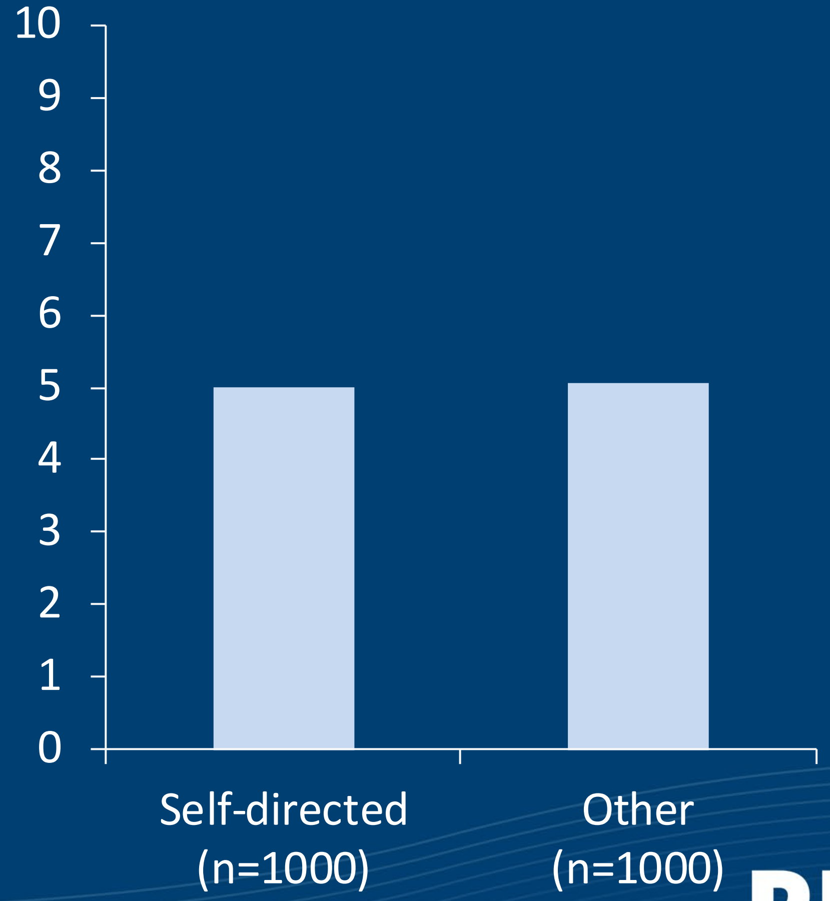
Motivating Example: Home and Community Based Long-Term Services and Supports

- Hypothetical study comparing self-directed care, home health aide (HHA) services, and adult day care

- Options with binary propensity scores:
  - Pairwise comparisons of one treatment vs the other two
    - Self-directed care vs other options, HHA vs others, adult day care vs others
  - Pairwise comparisons among specific treatments
    - Self-directed care vs HHA, adult day care vs HHA

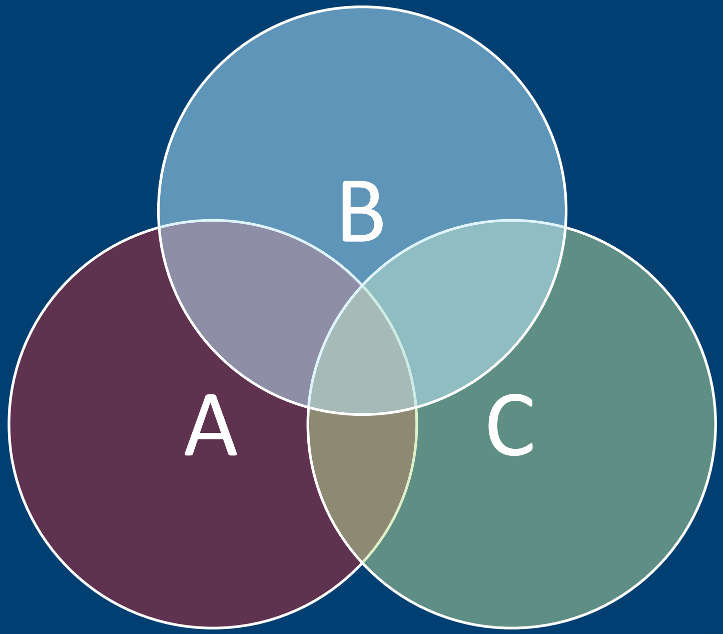
Restricting treatments to binary indicators can obscure between-group differences





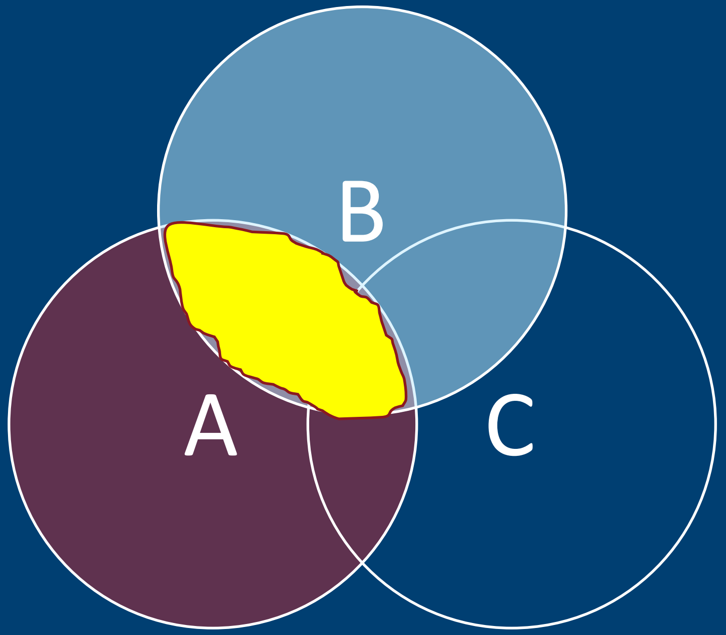
Pairwise comparisons among single treatments exclude sample receiving alternates

- If only interested in that specific comparison, this isn't an issue
- But, it complicates inferences about differences across the *entire set of treatments*



Pairwise comparisons among single treatments exclude sample receiving alternates

- If only interested in that specific comparison, this isn't an issue
- But, it complicates inferences about differences across the *entire set of treatments*

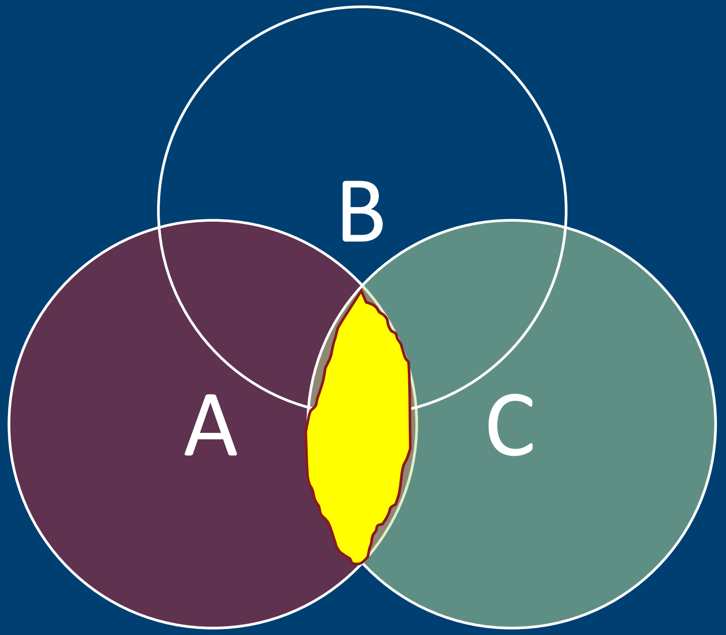


A vs B

PEPReC
Partnered Evidence-based Policy Resource Center
A VA QUERI Center

Pairwise comparisons among single treatments exclude sample receiving alternates

- If only interested in that specific comparison, this isn't an issue
- But, it complicates inferences about differences across the *entire set of treatments*

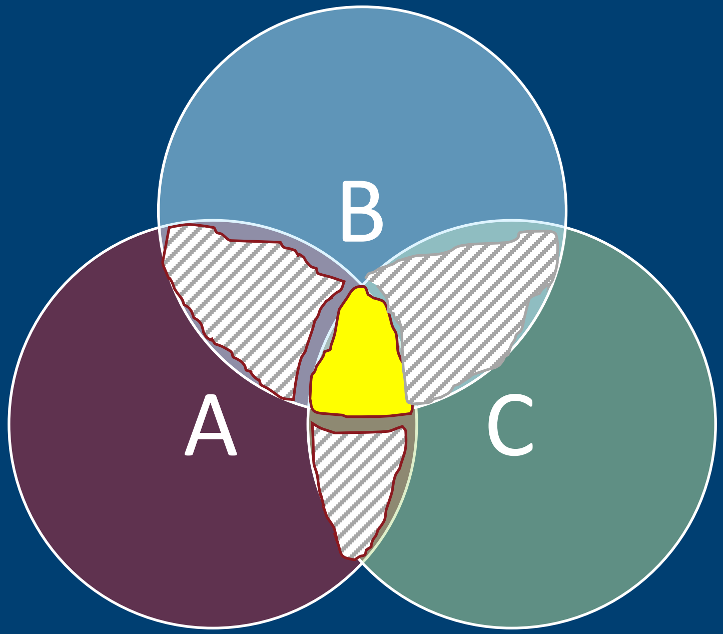


A vs C

PEPReC
Partnered Evidence-based Policy Resource Center
A VA QUERI Center

Pairwise comparisons among single treatments exclude sample receiving alternates

- If only interested in that specific comparison, this isn't an issue
- But, it complicates inferences about differences across the *entire set of treatments*



A vs B vs C

PEPReC
Partnered Evidence-based Policy Resource Center
A VA QUERI Center

Pairwise comparisons among single treatments exclude sample receiving alternates

- Example:
  - Estimate binary propensity score for self-directed care vs HHA
    - Estimated among those with non-zero probability of receiving either care option; could include individuals with 0 probability of receiving adult day care
  - Estimate binary propensity score for self-directed care vs adult day care
    - Estimated among those with non-zero probability of receiving either care option; could include individuals with 0 probability of receiving HHA
  - Cannot directly compare these two estimates – derived from different subsets of the sample

Choice of Strategy Depends on Goal of Analysis

- Series of pairwise comparisons
  - Apply standard propensity score methods

- Simultaneous comparison of multiple options
  - Requires additional restrictions on observations used to estimate treatment effect
    - Non-zero chance of receiving *any* of the treatment options
  - Uses a generalized propensity score approach
  - Requires additional thought about vectors of propensity scores

Generalized propensity score

- Probability of receiving one treatment level/option, conditional on observed covariates
- Each level/option has its own propensity score, but all propensity scores are estimated from a single multinomial model

- Can be estimated with:
  - Maximum likelihood estimation (multinomial logit or probit)
  - Covariate-balancing propensity score method (uses generalized method of moments)
  - Nonparametric machine learning models

Degree of Similarity of Vector of Propensity Scores

- Vector of propensity scores = a collection of an observation’s estimated probabilities of receiving each treatment option

- Binary treatment (yes/no)
  - Vector contains p(treatment) and p(no treatment)

- Multiple treatment options (A, B, C)
  - Vector contains p(A), p(B), p(C)

Degree of Similarity of Vector of Propensity Scores

- Vector of propensity scores = a collection of an observation's estimated probabilities of receiving each treatment option
- Binary treatment (yes/no)
  - Vector contains p(treatment) and p(no treatment)
  - By matching on p(treatment), implicitly matching on p(no treatment)
- Multiple treatment options (A, B, C)
  - Vector contains p(A), p(B), p(C)
  - Strategies vary in whether they require matching on probability of each treatment

Degree of Similarity of Vector of Propensity Scores – Multiple Treatment Options

- Non vector-based methods
  - Require matching on probabilities of two treatment levels; require non-zero probabilities of other treatments

- Vector-based methods
  - Require similarity on probabilities of all treatment levels; require non-zero probabilities of all treatments
    - Enhances ability to make direct comparisons of pairwise treatment effects derived from the same sample

Example – Simultaneous Comparison of Multiple Options

- Compare two pairwise average treatment effects (ATEs):
  - Differences in nursing home admission between self-directed care and HHA (ATE 1)
  - Differences in nursing home admission between self-directed care and adult day care (ATE 2)

Example – Simultaneous Comparison of Multiple Options

- Compare two pairwise average treatment effects (ATEs):
  - Differences in nursing home admission between self-directed care and HHA (ATE 1)
  - Differences in nursing home admission between self-directed care and adult day care (ATE 2)

- Non-vector based methods:
  - ATE1 estimated in individuals with nonzero, but widely varying p(adult day care)
  - ATE2 estimated in individuals with nonzero, but widely varying p(HHA)
  - Direct comparison of ATE1 and ATE2 is challenging

Example – Simultaneous Comparison of Multiple Options

- Compare two pairwise average treatment effects (ATEs):
  - Differences in nursing home admission between self-directed care and HHA (ATE 1)
  - Differences in nursing home admission between self-directed care and adult day care (ATE 2)

- Vector based methods:
  - ATE1 estimated in individuals with similar probabilities of receiving any of the treatment options
  - ATE2 estimated in individuals with similar probabilities of receiving any of the treatment options
  - Direct comparison of ATE1 and ATE2 is possible

Treatment Effects of Interest

- ATEs and Average Treatment Effects on the Treated (ATT)
- For 3 treatment groups, have three ATEs:
  - A vs B
  - B vs C
  - A vs C
- And 9 ATTs:
  - Each ATE among observations that received a single treatment
  - Includes transitive treatment effects (e.g., A vs B among those that received C)

Choice: Incorporating Generalized Propensity Score

- Weighting
- Matching
- Subclassification
- Regression adjustment

Choice: Incorporating Generalized Propensity Score

- Weighting

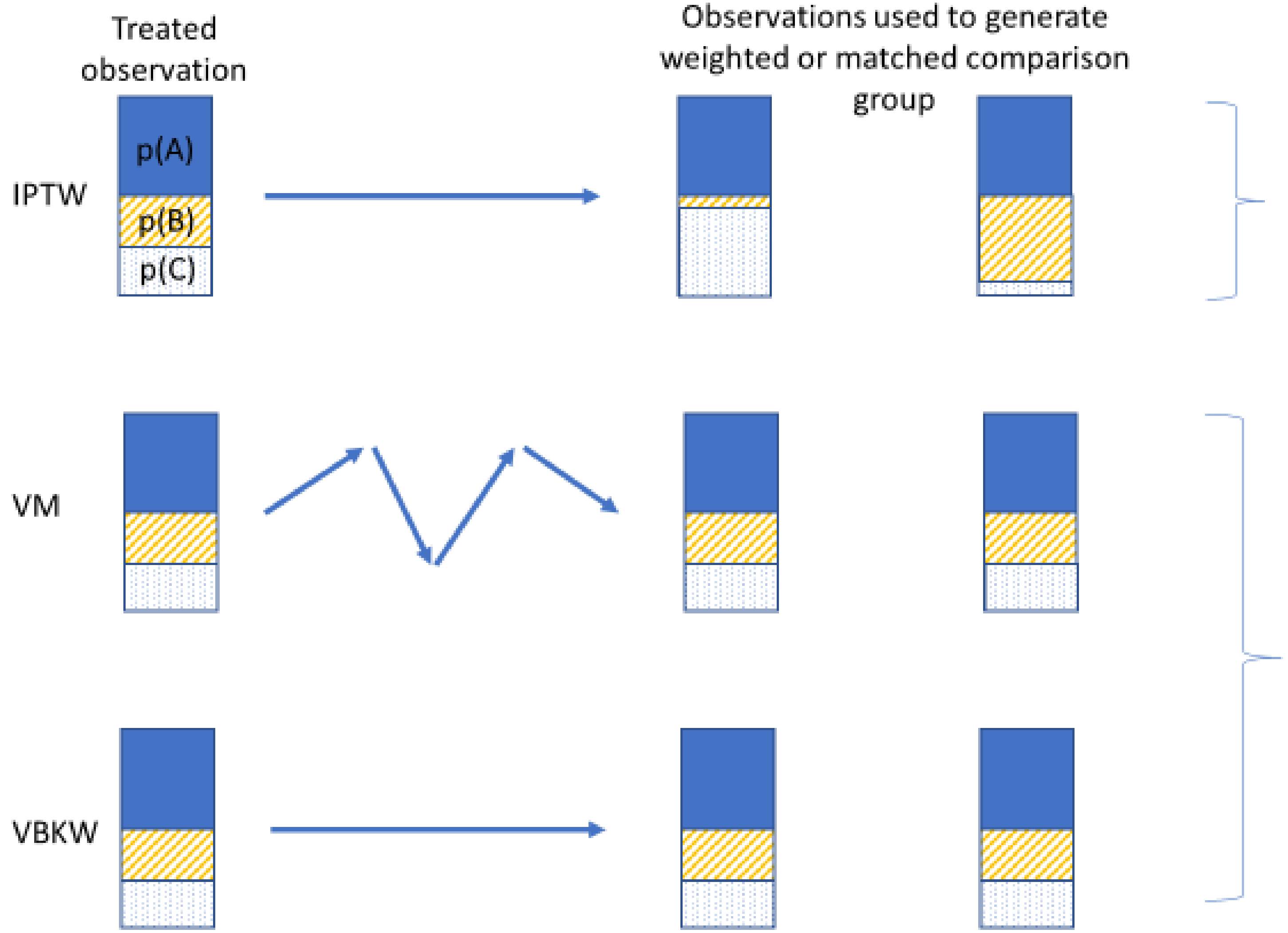
- Matching

- Subclassification
  - Optimal number of strata required to reduce selection bias varies with sample size
  - With traditional number of strata, less effective at reducing bias than weighting

- Regression adjustment
  - Produces inferior covariate balance relative to weighting or matching
  - Can introduce greater bias into treatment effect estimates

Which Weighting or Matching Method is Best?

- Non-vector based methods
  - Inverse probability of treatment weighting (IPTW)
    - Commonly used, but often leads to highly biased, inefficient estimates
  - Generalized propensity score matching
    - Performs better than IPTW but still likely to lead to biased, inefficient estimates
- Vector-based methods
  - Vector-based matching
  - Vector-based kernel weighting (VBKW)
    - New
    - Builds off of principles of vector matching, but easier to implement
    - Reliably produces unbiased, efficient treatment effect estimates


Treated observation
Observations used to generate weighted or matched comparison group
IPTW
p(A)
p(B)
p(C)
VM
VBKW

Non-vector based matching and weighting methods focus on the probability of a single treatment level, leading to comparison groups with disparate probabilities of receiving other treatment levels

Vector-based methods lead to the construction of a comparison group with similar probabilities of each treatment level

VBKW does this in fewer steps than VM and produces the least biased and most efficient estimates

Garrido et al. Statistics in Medicine 2021; 40(5): 1204-1223.

Inverse Probability of Treatment Weighting (IPTW)

- Default option: Calculate ATE
  - Observations receive weights of 1/p(observed treatment)
    - Inverse of the propensity score for the treatment option received

- Can be modified to calculate ATT
  - Treated observations receive weight of 1
  - Comparison observations receive weight of p(observed treatment)/p(comparison treatment)

- Requires each observation have a non-zero probability of any of the treatment levels
- Does not require similarity across entire vector of propensity scores

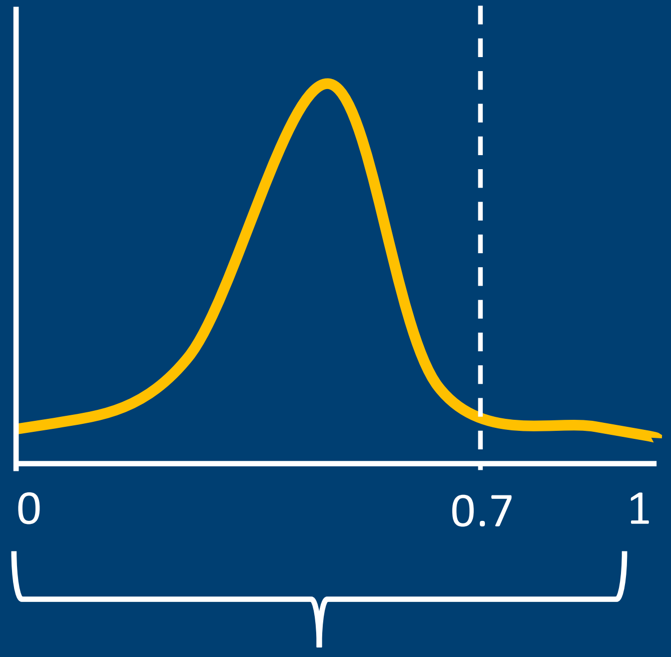
Generalized Propensity Score Matching (GPSM)

- Each matching step is only based on propensity of receiving a single treatment
- Example: Estimate ATE of A vs B
  - Take average difference of observed outcomes among a sample matched on p(A) and a sample matched on p(B)
  - Matches can be completed in two ways
    - Can complete matches for p(A) from sample receiving B or C, if only interested in ATEs
    - Can complete matches for p(A) from sample receiving B, if interested in ATEs and ATTS

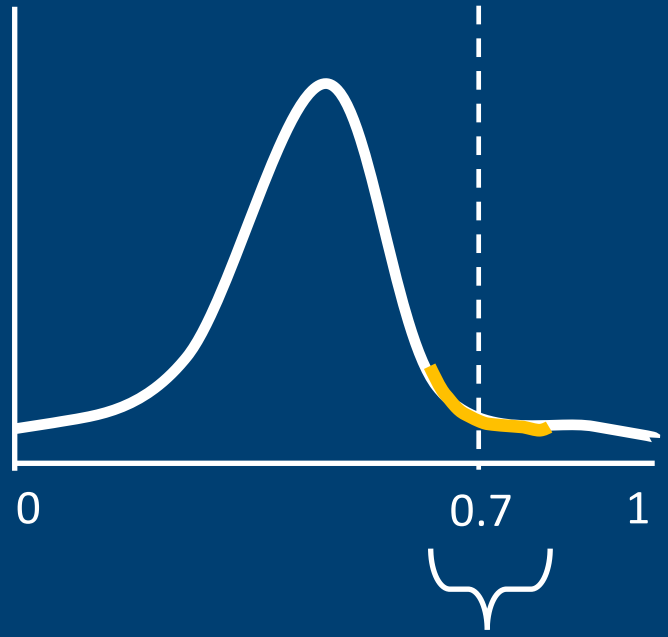
- Requires each observation have a non-zero probability of any of the treatment levels
- Does not require similarity across entire vector of propensity scores

Vector-Based Kernel Weighting

Weights based on a kernel function



Inverse probability of treatment weights



Kernel weights

Vector-Based Kernel Weighting

- Kernel weights are assigned to comparison observations that have similar vectors of propensity scores
- To estimate ATT of A vs B | A, generate the following weights:
  - Assign treated observations (A) a weight of 1
  - Assign comparison observations (B) a kernel weight if p(A) is within bandwidth of treated observation’s p(A), p(B) is within bandwidth of treated observation’s p(B), and p(C) is within bandwidth of treated observation’s p(C)
- To estimate ATEs, generate weights that are the sum of non-transitive ATT weights(ATE of A vs B = ATT of A vs B | A + ATT of A vs B | B)
- Requires each observation have a non-zero probability of any of the treatment levels
- Requires similarity across entire vector of propensity scores

Standard Errors in IPTW, GPSM, VBKW

- IPTW – Bootstrapped SEs
- GPSM – Abadie-Imbens adjustment
- VBKW – Abadie-Imbens adjustment vs bootstrapped?
  - Can use bootstrapped standard errors if bandwidth for kernel weight is large enough (bandwidth = 0.5*sd [logit (pscore)])

Comparing IPTW, GPSM, VBKW: Simulation

- Simulations on analytic scenarios (unique combinations of the following characteristics):
  - Sample size (n = 600, 1200, 3000, 9600)
  - Misspecification of the estimated propensity score
  - Number of treatment groups (3, 5)
  - Sample distribution across treatment groups
  - Treatment effect heterogeneity
  - Coefficient set
- 4,584 scenarios; 1000 replications each

Comparing IPTW, GPSM, VBKW: Outcomes

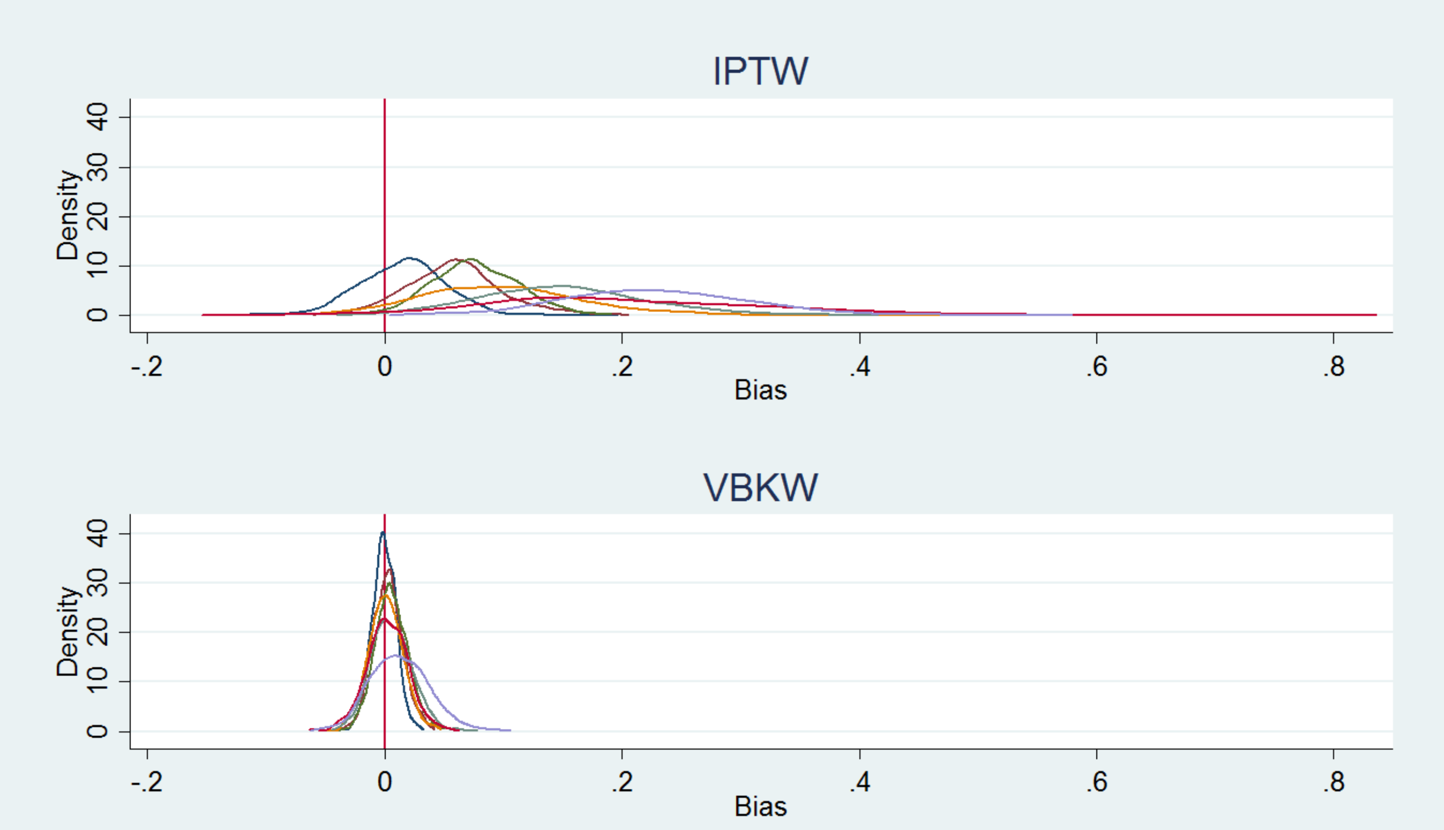
- Bias
  - Absolute bias – distance between estimated and true treatment effect
  - Absolute mean relative bias (AMRB) – bias as % of true treatment effect
- Efficiency
  - Interquartile range (IQR)
  - Root-mean-squared error (RMSE)
  - Median absolute error (MAE)
- Covariate balance
  - Absolute standardized differences in prognostic score values
- Confidence interval coverage

Estimates based on IPTW more likely to be biased and inefficient than estimates based on GPSM or VBKW

Strategy	% of scenarios with < 20% AMRB	Median AMRB	Median absolute bias	Median IQR	Median RMSE	Median MAE
IPTW	37%	40.4	0.050	0.080	0.09	0.06
GPSM	50%	19.9	0.025	0.080	0.09	0.06
VBKW	95%	4.21	0.005	0.055	0.04	0.03

1008 scenarios; n = 1200, 3 treatment groups

IPTW more sensitive to propensity score misspecification than GPSM or VBKW



Dark blue line represents fully saturated propensity score model

All other lines represent misspecified propensity score models

PEPReC

Partnered Evidence-based Policy Resource Center
A VA QUERI Center

VBKW more likely to lead to covariate balance than other methods

Strategy	Median AMRB	Median absolute mean standardized differences in prognostic scores
IPTW	30.96	0.111
GPSM	20.56	0.129
VBKW	3.82	0.032

VBKW produces unbiased, efficient estimates within commonly encountered analytic scenarios

- Robust to
  - Propensity score model misspecification
  - Sample size
  - Distribution of sample across treatment groups
  - Kernel function choice
  - Use of multinomial logit vs multinomial probit
  - Baseline covariate imbalance

Choice: Estimation via maximum likelihood estimation (MLE) or CBPS?

- Covariate Balancing Propensity Score (CBPS)
  - Generalized method of moments
  - Estimates a propensity score model that optimizes covariate balance across treatment groups

- Preliminary results suggest estimated ATEs derived from IPTWs estimated via CBPS are less biased than IPTWs estimated via MLE
- VBKW estimates, whether based on MLE or CBPS, are less biased than IPTW estimates with CPBS
- Similar patterns are observed for efficiency

Choice: VBKW vs Entropy Balancing?

- Entropy balancing
  - Creates treatment and comparison groups with similar moments (mean, variance, skew) of covariate distributions
  - Does not require specification of a propensity score model

- Preliminary results suggest
  - Entropy balancing produces estimates with less bias than VBKW when baseline imbalance in covariates is relatively low
  - VBKW is more robust to baseline imbalance in covariates than entropy balancing

Refining VBKW – Next Steps

- Test in empirically-based (plasmode) simulations (*in progress*)
- Develop Stata command (*in progress*)
- Develop Abadie-Imbens adjustment for standard errors with multinomial propensity score
- Optimal tuning of bandwidth
- Test performance when combined with covariates in doubly robust estimates

Conclusions

- Account for vectors of propensity scores when creating propensity score matches or weights
- Ensuring similarity across vectors of propensity scores will lead to estimates with less bias and greater efficiency
- Failure to account for vectors will limit comparisons of pairwise treatment effects
- VBKW is a relatively straightforward method to account for similarity across vectors of propensity scores

References

- Garrido, Lum, Pizer. Vector-based kernel weighting: A simple estimator for improving precision and bias of average treatment effects in multiple treatment settings. Statistics in Medicine 2021; 40(5): 1204-1223.
- Ho DE et al. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. Political Analysis 2007; 15: 199–236.
- Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika. 1983; 70: 41-45
- Stuart EA. Matching methods for causal inference: A review and look forward. Statistical Science 2010; 25 (1): 1–21.

Questions?

melissa.garrido@va.gov

@GarridoMelissa