Vector-Based Kernel Weighting:

A Simple Estimator for Improving Precision and Bias of Average Treatment Effects in Multiple Treatment Settings

Jessica Lum, MA[1]
Steven Pizer, PhD[1, 2]
Melissa Garrido, PhD[1, 2]

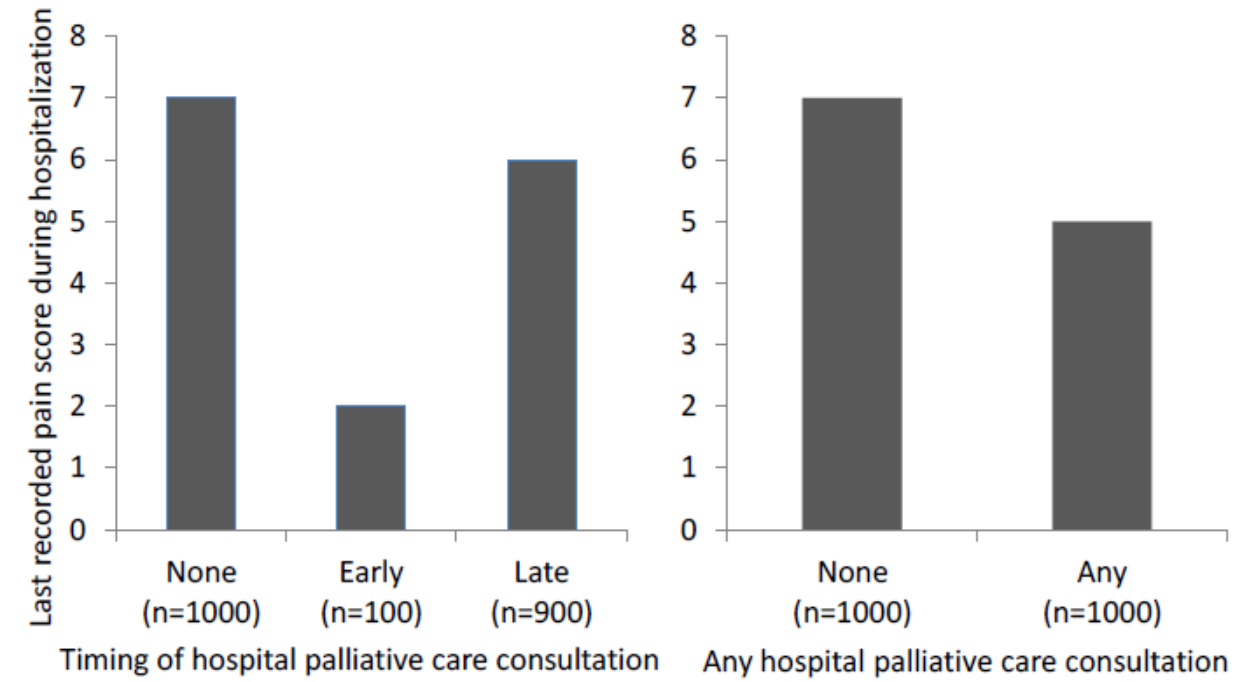
1. Department of Veterans Affairs
2. Boston University School of Public Health

Stata Conference
Columbus, OH
July 20th, 2018

Overview

1. Importance of using full propensity score vector
2. Common support in multiple treatment setting
3. Transitive treatment effects
4. Weighting/Matching strategies
   - Introduce new treatment effect estimator
5. Monte Carlo (MC) simulation design
6. Demonstrate bias and efficiency of estimators via MC simulations

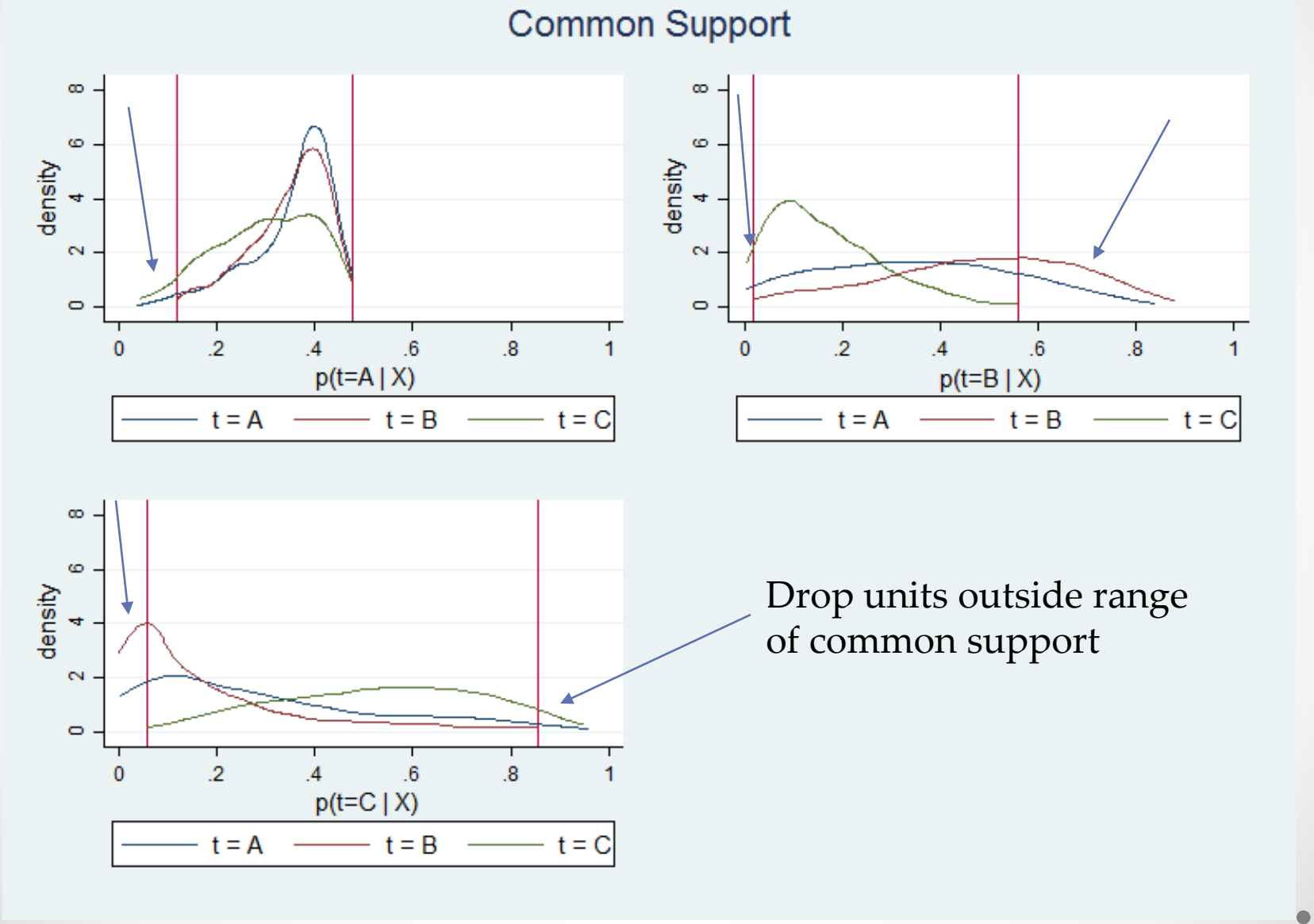
Multiple Treatment Groups



- Accounting for all values of a treatment variable in a single equation helps ensure propensity scores from a multinomial model leads to treatment effect estimation among patients with non-zero probabilities of receiving any of the other treatments (common support).

- Multinomial choice model: Predicts several generalized propensity scores, each one representing probability of receiving one of the treatments. Predicted probabilities are represented by a ***propensity score vector*** of values for each observation.

Common Support



Drop units outside range of common support

Transitive Treatment Effects*

- Treatment effect** estimation involves constructing counterfactual outcomes from a comparison group determined to be most "similar" to the reference group based on propensity scores.

- Pairwise treatment effects are transitive iff conditioning on a sample eligible to receive the same treatment groups.

$$E[Y(A) - Y(C) \mid T = A] - E[Y(A) - Y(B) \mid T = A]$$
$$= E[Y(B) - Y(C) \mid T = A]$$

*Lopez and Gutman (2017).

** All estimates are obtained as weighted mean differences of outcomes, with weights normalized to sum to 1 in each treatment group.

Goals

- The degree to which different weighting or matching strategies lead to robust inferences in messy empirical scenarios with multiple treatment groups is unknown. We seek to understand the scenarios in which all methods perform similarly, as well as scenarios that produce divergent inferences.

- To identify when estimators produce unbiased and efficient estimators in a variety of settings, we compare 4 estimators which each utilize propensity scores differently in treatment effect estimation:

1. Inverse Probability of Treatment Weighting (IPTW) (weighting)
2. Kernel Weighting (KW) (weighting + matching)
3. Vector Matching (VM) (matching)
4. Vector-Based Kernel Weighting (VBKW) (weighting + matching)

Inverse Probability of Treatment Weights

- In estimating E[Y(A) – Y(B)],

$$W = \begin{cases} \frac{1}{p(t=A \mid x)}, & if\ t = A \\ \frac{1}{p(t=B \mid x)}, & if\ t = B \end{cases}$$

- In estimating E[Y(A) – Y(B) | T = A],

$$W = \begin{cases} 1, & if\ t = A \\ \frac{p(t=A \mid x)}{p(t=B \mid x)}, & if\ t = B \end{cases}$$

- Incorrectly estimated IPTWs may have extreme values, increasing variance of treatment effect estimate, and potentially leading to biased estimates.
- In pairwise comparisons, the IPTW estimator does not utilize the full propensity score vector.

Kernel Weights

- In estimating E[Y(A) – Y(B) | T = A],

$$W = \begin{cases} 1, & if\ t = A \\ K_j(D_{iA}) / \sum_j^{N_B} K_j(D_{iA}) & if\ t = B \end{cases}$$

$$K_j(D_{iA}) = \begin{cases} \frac{3}{4}\left(1 - \left(\frac{D_{iA}}{0.06}\right)^2\right), & if\ D_{iA} < 0.06 \\ 0, & otherwise \end{cases}$$

$$D_{iA} = | p_j(A \mid X) - p_i(A \mid X) |$$

where *i* and *j* index T = A and T = B units, respectively, and N_B is the total T = B units.

- Weight for estimating E[Y(A) – Y(B)] = $W_{E[Y(A) - Y(B) \mid T = A]} + W_{E[Y(A) - Y(B) \mid T = B]}$

- In pairwise comparisons, the KW estimator does not utilize the full propensity score vector.

Vector Matching*

- VM creates matched sets with units that are close on one component of the PS vector, and *roughly* similar on the other components. To estimate E[Y(A) – Y(B) | T = A], E[Y(A) – Y(C) | T = A], or E[Y(B) – Y(C) | T = A]:

1. Refit PS model to obtain new propensity scores, take logit transform of scores.

2. 1:1 greedy match T=A units to T = B units with replacement on logit(p(A|**X**)) within k-means strata of logit(p(C|**X**)), within caliper.

3. 1:1 greedy match T=A units to T = C units with replacement on logit(p(A|**X**)) within k-means strata of logit(p(B|**X**)), within caliper.

- Combination of multiple steps in creating this matched set makes VM relatively complex to implement.

- Weight = The number of times a subject is used to create a matched set.

* Lopez MJ, Gutman R. Estimation of causal effects with multiple treatments: A review and new ideas. Statistical Science 2017; 32(3): 432-454.

Vector-Based Kernel Weighting

- In estimating E[Y(A) – Y(B) | T = A],

$$W = \begin{cases} 1, & if\ t = A \\ K_j(D_{iA}) / \sum_j^{N_B} K_j(D_{iA}) & if\ t = B \end{cases}$$

$$K_j(D_{iA}) = \begin{cases} \frac{3}{4}\left(1 - \left(\frac{D_{iA}}{0.06}^2\right)\right), & if\ D_{iA} < 0.06\ \textbf{and}\ D_{iB} < 0.06\ \textbf{and}\ D_{iC} < 0.06 \\ 0, & otherwise \end{cases}$$

$$D_{iA} = |\, p_j(A \mid X) - p_i(A \mid X) \,|$$
$$D_{iB} = |\, p_j(B \mid X) - p_i(B \mid X) \,|$$
$$D_{iC} = |\, p_j(C \mid X) - p_i(C \mid X) \,|$$

where *i* and *j* index T = A and T = B units, respectively, and N_B is the total T = B units.

- Weight for estimating E[Y(A) – Y(B)] = $W_{E[Y(A) - Y(B) \mid T = A]} + W_{E[Y(A) - Y(B) \mid T = B]}$
- This translates to non-zero weight assignment to controls with a similar propensity score **vector** instead of just being similar on p(A | **X**), as in KW.
- Rather than matching in several steps, as in VM, VBKW takes one step to apply propensity score vector matching.

Vector-Based Kernel Weighting

Features	VM	KW	VBKW
Requires one step to match		x	x
Requires clustering	x		
Weighting		x	x
Matching	x	x	x
Utilizes full PS vector	x		x
Transitivity of estimates	x		x

Expectations

- We wish to identify scenarios in which inferences are most likely to diverge under finite samples.

- We expect estimates from kernel weights (with a low emphasis on extreme weights) to be less biased than IPTW estimates when the data-generating process for the true propensity score is nonlinear and the estimated propensity score model is misspecified.

- We expect differences in inferences to be more likely when the presence of extreme weights is more likely or when identification of matches may be more difficult.

Simulation

- We report results from 3 treatment levels, n=999, as results from other simulation designs are qualitatively similar.

- We look at 12 Estimands: 3 ATEs, 9 ATTs. True ATTs equal to true ATEs when treatment effects were homogeneous.

- When the simulation included 3 treatment groups, the true ATEs, E[Y(A) – Y(B)], E[Y(A) – Y(C)], and E[Y(B) – Y(C)] were set to -0.1, -0.2, -0.1, respectively.

- Model misspecification via estimation with (mlogit) main effects only.

Monte Carlo simulation design

Simulation parameters*

Functional form of the true propensity score model. Increasing model complexity through nonlinearity and/or nonadditivity. Based on Setoguchi et al. (2008).

Number of treatments (k = 3, 4)

Sample size (n = 999, n = 9,999)

Sample distribution across treatment groups:

- Equal distribution of units into treatment groups
- 50% of sample in one group, remaining split equally
- 10% of sample in one group, remaining split equally

Treatment effect distribution:

- Homogenous treatment effect
- Heterogeneous treatment effect (associated with confounder)
- Heterogeneous treatment effect (associated with outcome only variable)

*For a given k, and n, there are 7 model misspecifications x 12 Estimands x 3 sample dist. x 3 effect dist. = 756 unique analytic scenarios to compare estimator performance.

Monte Carlo simulation design

Evaluation metrics

- Bias*
- Bias as % of SD of effect estimate*
- Interquartile Range (IQR)
- Root-mean-squared-error (RMSE)
- Median absolute error (MAE)*
- Number of analytic scenarios with < 40% Bias %

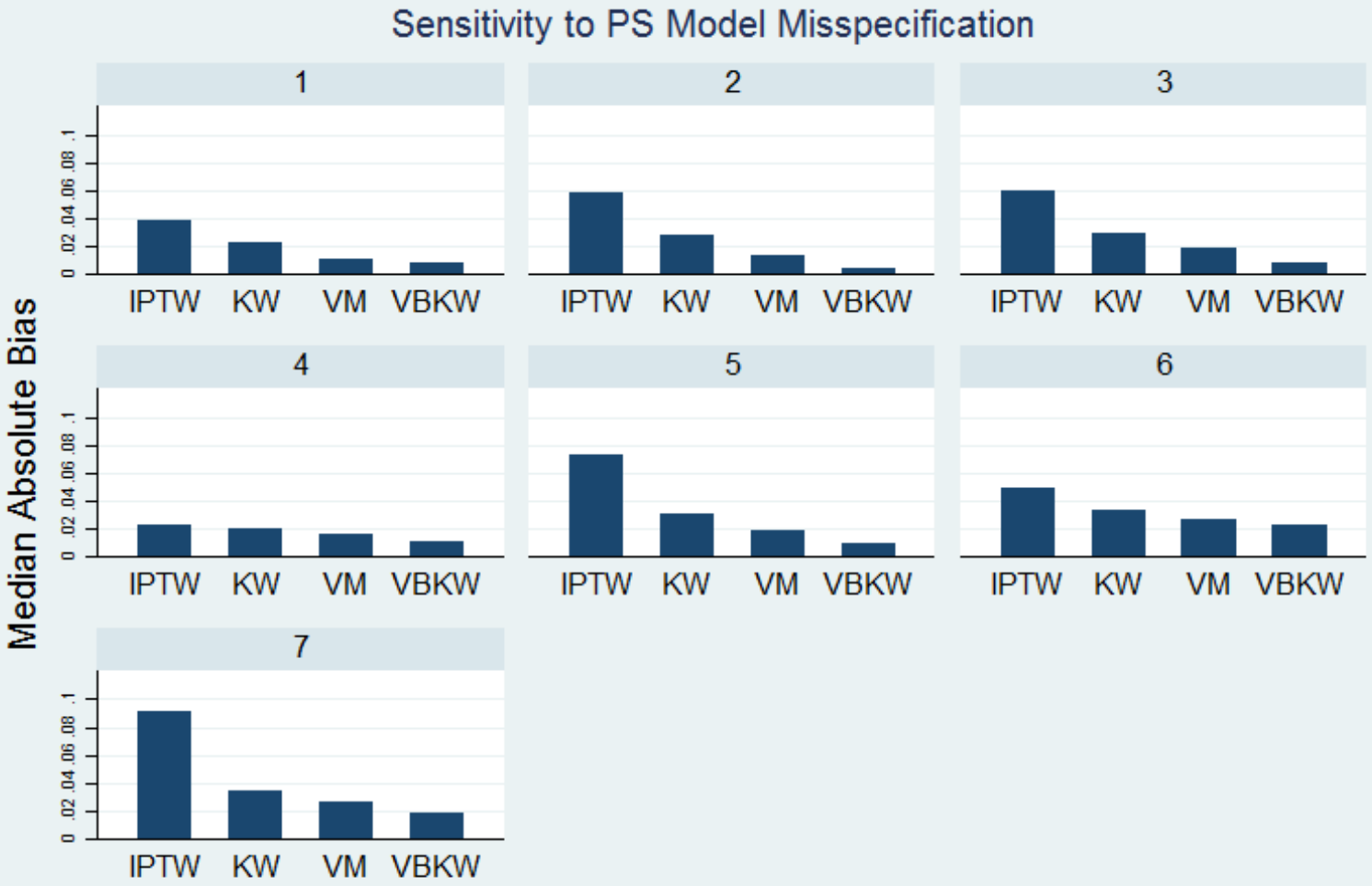
*Kang and Schafer (2007)

VBKW led to least biased and most efficient estimates

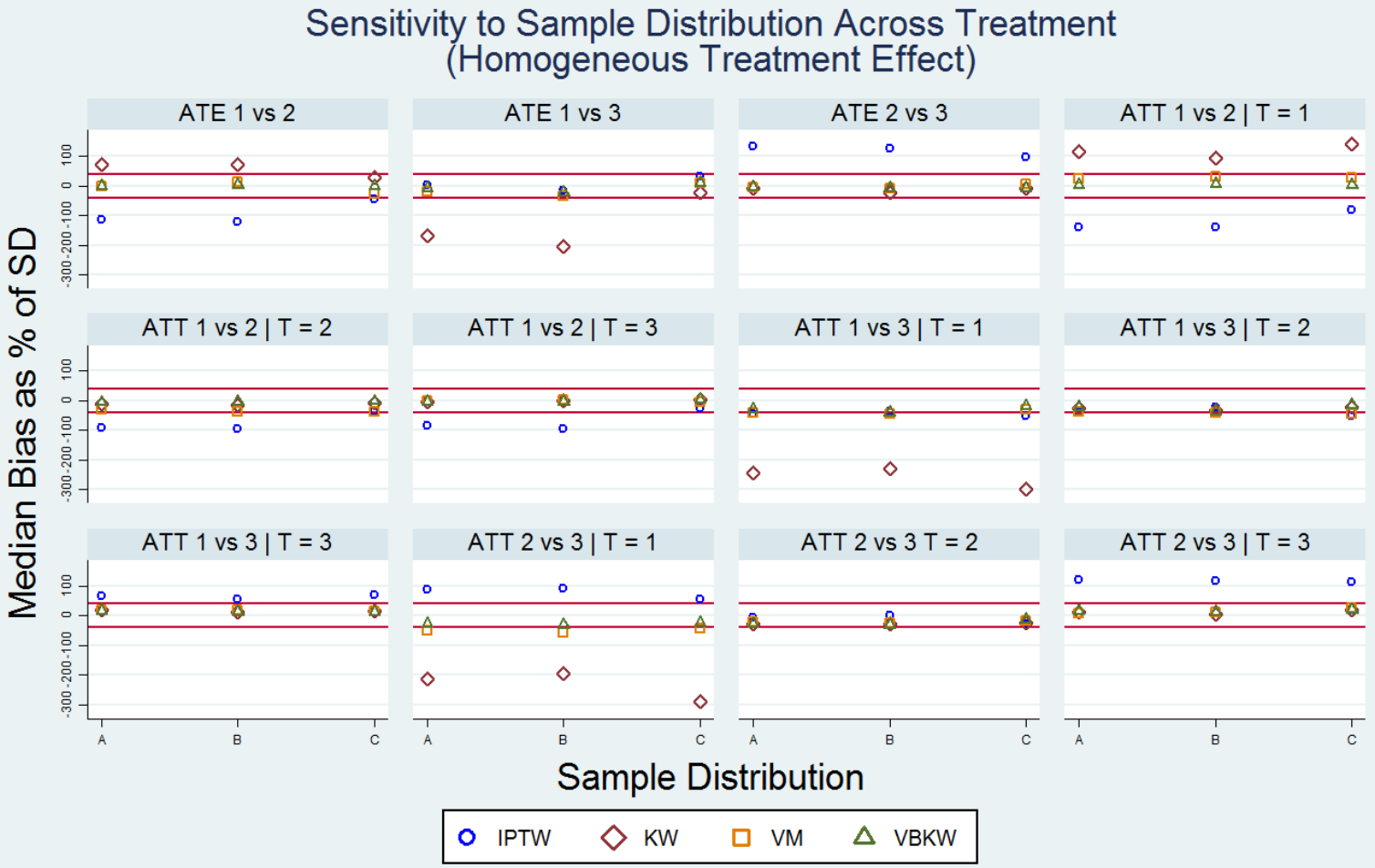
Summary of Bias and Efficiency of Estimates

	Number (%) Analytic Scenarios with < 40% Bias	Median Bias as % of SD	Median Absolute Bias	Median IQR
IPTW	221 (29)	69.626	0.051	0.095
KW	356 (47)	45.102	0.030	0.085
VM	542 (72)	26.362	0.018	0.103
VBKW	554 (73)	17.509	0.010	0.075

VBKW less sensitive to PS model misspecification

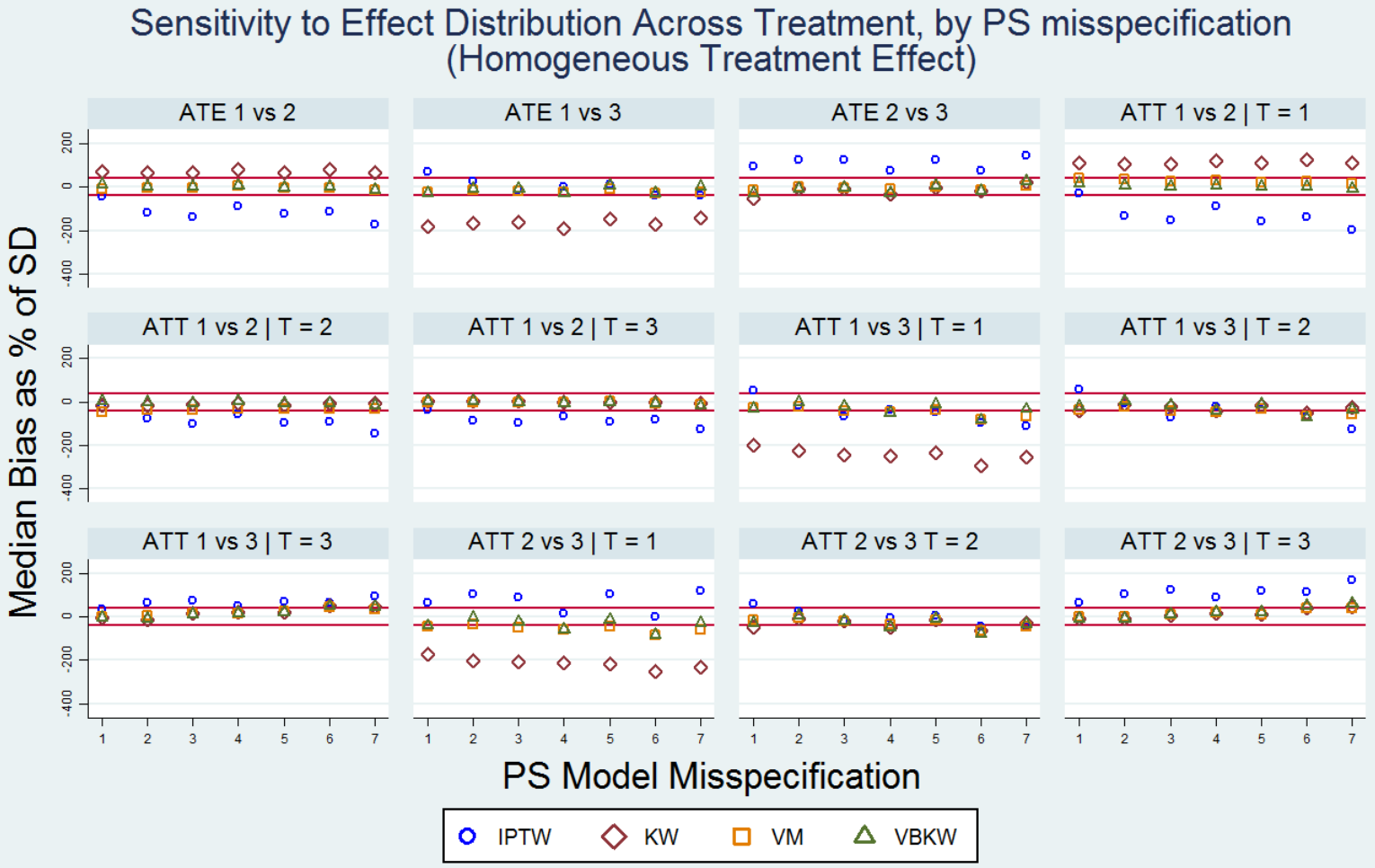


When treatment effect is homogeneous, IPTW & KW are most likely to be biased



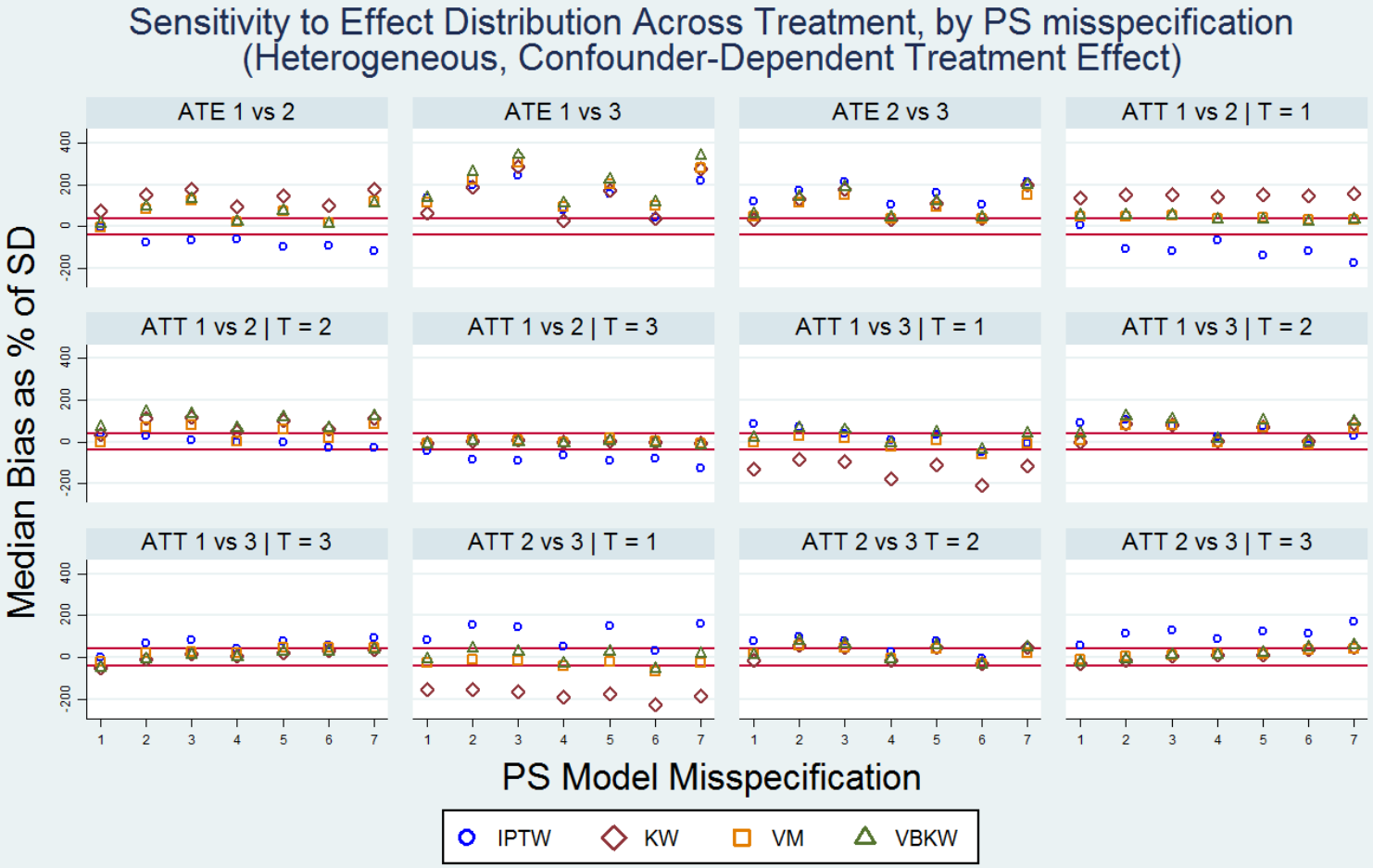
Red lines indicate range of values within 40% of SD
A = sample equally distributed amongst treatment groups
B = 50% of sample in one treatment group, equal distribution of sample across remaining groups
C = 10% of sample in one group, equal distribution of sample across remaining groups

When treatment effect is homogeneous, IPTW & KW are most likely to be biased



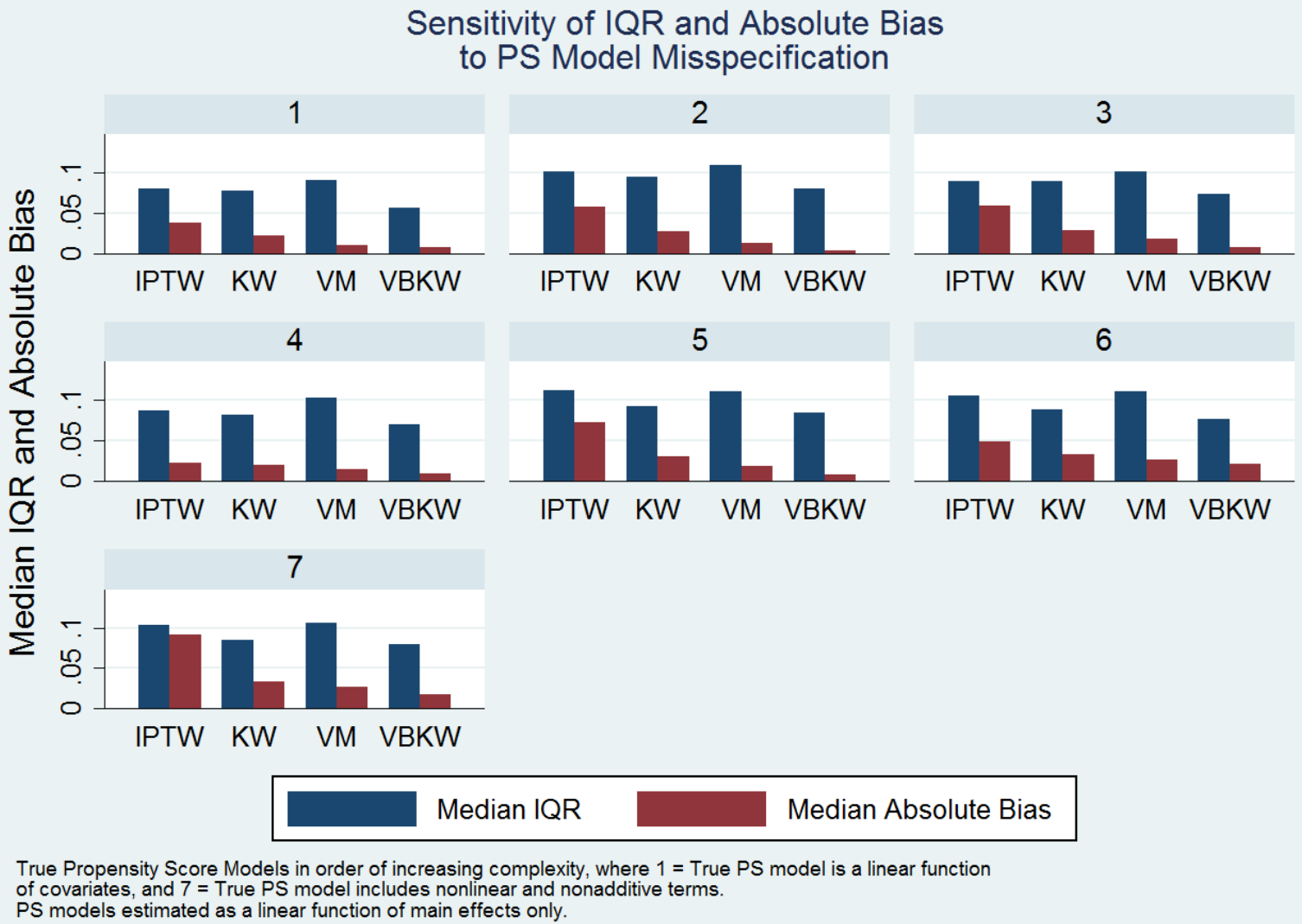
Red lines indicate range of values within 40% of SD
True Propensity Score Models in order of increasing complexity, where 1 = True PS model is a linear function of covariates, and 7 = True PS model includes nonlinear and nonadditive terms.
PS models estimated as linear function of main effects only.

In the presence of heterogeneous, confounder-dependent treatment effects, all strategies likely to produce biased ATEs



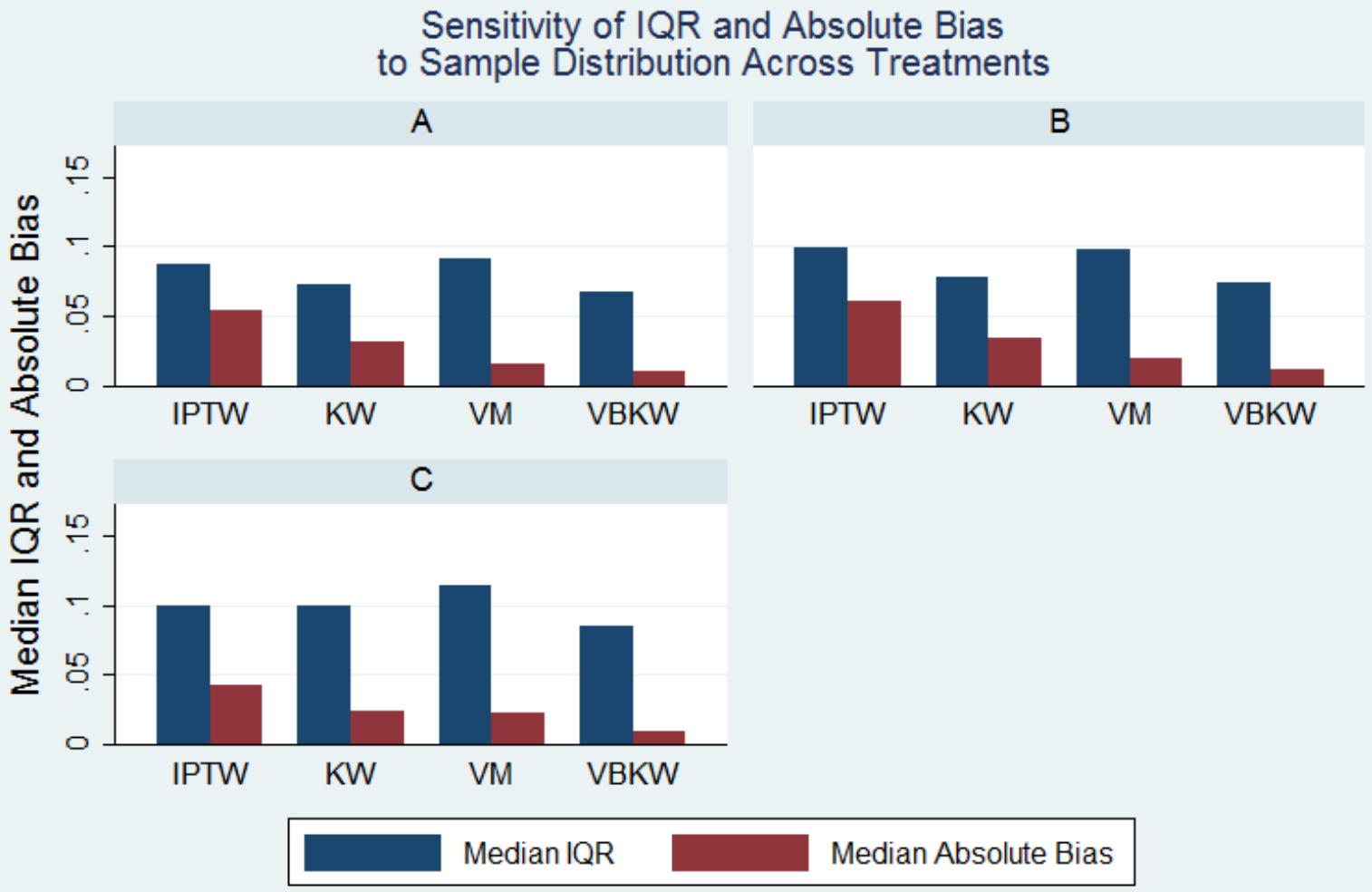
Red lines indicate range of values within 40% of SD
True Propensity Score Models in order of increasing complexity, where 1 = True PS model is a linear function of covariates, and 7 = True PS model includes nonlinear and nonadditive terms.
PS models estimated as linear function of main effects only.

VBKW most efficient across various PS model misspecifications

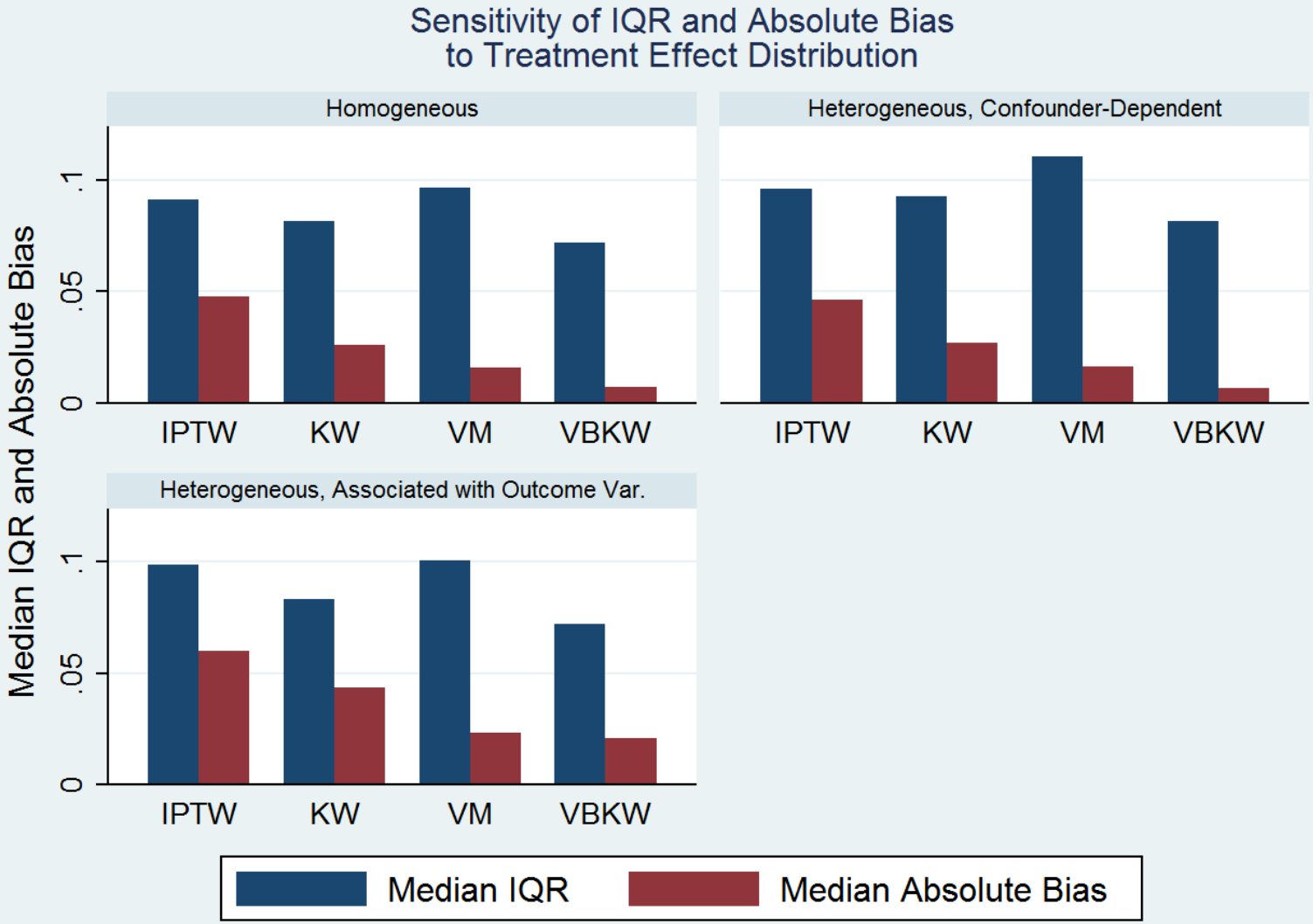


True Propensity Score Models in order of increasing complexity, where 1 = True PS model is a linear function of covariates, and 7 = True PS model includes nonlinear and nonadditive terms.
PS models estimated as a linear function of main effects only.

VBKW most efficient across sample distributions



VBKW most efficient across effect distributions



Limitations/ Future directions

- Simulations based on imposed rather than empirical DGP. Future work will include plasmode simulations based on empirical DGP.

- Our estimated PS model contained only main effects to test robustness to misspecification. Researchers should ensure propensity score leads to adequate covariate balance.

- We did not test sensitivity of results to observed covariate choice or covariate measurement errors, nor do we examine doubly-robust estimates.

- Future work: plasmode simulations, generalized boosted models, covariate balancing propensity scores, variable bandwidth, assessment of covariate balance, robustness to unobserved confounding. Creation of **vbkw** command.

Discussion

- Simulation results suggest VBKW less sensitive to PS model misspecification & sample distribution across treatment groups than other methods. VBKW only slightly better than VM, but simpler to implement.

- IPTW & KW not well suited to produce unbiased estimates of transitive effects.

- None of the estimators led to consistent unbiased estimation of heterogeneous treatment effect due to confounder.

Contact Information

Thank you!

For comments, questions, or suggestions, or to request a copy of the working paper:

Jessica Lum

Jessica.Lum2@va.gov

1. Project supported by VA HSR&D IIR 16-140 (PI: Garrido)
2. The views expressed here are those of the authors and not necessarily those of the Department of Veterans Affairs or United States Government.