Comparative Effectiveness/Safety Research With Multiple Treatment Groups

COMPARATIVE EFFECTIVENESS/SAFETY RESEARCH WITH MULTIPLE TREATMENT GROUPS

KAZUKI YOSHIDA

A Dissertation Submitted to the Faculty of
The Harvard T.H. Chan School of Public Health
in Partial Fulfillment of the Requirements
for the Degree of Doctor of Science
in the Departments of Epidemiology and Biostatistics

Harvard University
Boston, Massachusetts
March 2019

Dissertation Advisors: Dr. Sonia Hernández-Díaz and Dr. Robert J. Glynn Kazuki Yoshida

Comparative Effectiveness/Safety Research

with Multiple Treatment Groups

ABSTRACT

Thanks to the continued development of new medications in many therapeutic areas, patients and clinicians often are faced with the need to choose from multiple treatment options. All treatment decisions should ideally be informed with randomized controlled trials. However, randomized controlled trials with multiple active medications are rare. In the absence of trial evidence, comparative effectiveness/safety research utilizing observational data can play important roles.

Although propensity score methods have become a standard tool in comparative effectiveness/safety research, they are less frequently used in questions involving three or more treatment options. This is in part due to the lack of familiarity and methods. In this dissertation, we extended several existing methods that had originally been developed in the two-group setting to the multi-group setting to overcome this.

In **Chapter 1**, we extended the *matching weights*, an alternative propensity score weighting method, to the general multi-group setting. We showed its asymptotic equivalence to multi-group simultaneous propensity score matching and confirmed its similarity to three-way simultaneous matching in a simulation.

In **Chapter 2**, we applied the multi-group matching weights method to an applied question on the bone safety of analgesics. The analysis based on the initial treatment assignment showed similar changes in bone mineral density although the on-treatment analysis suggested a potentially detrimental effect of opioids.

In **Chapter 3**, we developed an *empirical equipoise* tool for the multi-group setting to address the question familiar to pharmacoepidemiologists: Are the treatment groups *similar enough*? We examined the settings in which the tool helped identify the danger of residual confounding due to dissimilar patient characteristics.

In **Chapter 4**, we proposed extensions of three existing propensity trimming methods into the multi-group setting. We examined their ability to reduce confounding due to unmeasured variables more common in the tails of the multinomial propensity score distribution.

In conclusion, we extended several existing propensity score methods to the multi-group setting. We hope these methods promote and improve comparative effectiveness/safety research with multiple treatment groups.

Table of Contents

List of Figures with Captions

Chapter 1

Chapter 2

Chapter 3

Chapter 4

List of Tables with Captions

Chapter 2

Chapter 4

Acknowledgments

I would like to express the most profound appreciation to Dr. Sonia Hernández-Díaz for allowing me to pursue my interest in methodological issues in pharmacoepidemiology and to Dr. Robert J. Glynn for his patient guidance in my research endeavor at the intersection of pharmacoepidemiology and biostatistics. I would like to thank Dr. Daniel H. Solomon for his continuous mentorship both in research and career development. I am grateful to Dr. Sebastien Haneuse for his assistance with the most technical aspects of my research.

I also received support from many people outside the dissertation committee. Drs. Jessica M. Franklin and John Jackson guided me through my first methodological paper, which became Chapter 1 of this dissertation. Dr. Joshua J. Gagne supported me to formulate the dissertation proposal by leading my oral qualifying exam committee. Drs. Seoyoung Kim, Elisabetta Patorno, Sara K. Tedeschi, Houchen Lyu, and Tzu-Chieh Lin helped me with the clinical examples and keeping manuscripts relevant. Dr. Til Stürmer assisted me extending his asymmetric trimming method. Ms. Zhi Yu, Dr. Gail A. Greendale, Dr. Kristine Ruppert, and Ms. Yinjuan Lian helped with the SWAN database.

I would also like to thank the members of my two written qualifying exam study groups: Dr. Katsiaryna Bykov, Dr. Andres Ardisson Korat, and Mr. Xeno Acharya for epidemiology and Mr. Xihao Li, Dr. Tom Chen, and Dr. Yan Wang for biostatistics.

Last but not least, I would like to thank my wife Tomoko for tolerating my protracted doctoral study and my three children, Haruka, Nanami, and Kouta for cheering me up.

Chapter 1: Matching weights to simultaneously compare three treatment groups: Comparison to three-way matching

AUTHORS: Kazuki Yoshida (1,2), Sonia Hernández-Díaz (1), Daniel H. Solomon (3,4), John W. Jackson (5,1), Joshua J. Gagne (4), Robert J. Glynn (2,4), Jessica M. Franklin (4)

AFFILIATIONS

1. Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, United States.
2. Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, United States.
3. Division of Rheumatology, Immunology and Allergy, Brigham and Women's Hospital, Boston, Massachusetts, United States.
4. Division of Pharmacoepidemiology and Pharmacoeconomics, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts, United States.
5. Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland, United States.

DISCLOSURES:
KY receives tuition support jointly from Japan Student Services Organization (JASSO) and Harvard T.H. Chan School of Public Health (partially supported by training grants from Pfizer, Takeda, Bayer and PhRMA). SHD has consulted for AstraZeneca and UCB. DHS receives salary support from institutional research grants from Eli Lilly, Amgen, Pfizer, AstraZeneca, Genentech, and Corrona. He also receives royalties from UpToDate, and serves in unpaid roles in studies funded by Pfizer and Eli Lilly. JWJ None. JJG has received salary support from institutional research grants from Novartis Pharmaceuticals Corporation. He is a consultant to Aetion, Inc., a software company. RJG None. JMF is PI of grants from PCORI and Merck. She also serves as consultant to Aetion, Inc.

FUNDING INFORMATION:
KY receives tuition support jointly from Japan Student Services Organization (JASSO) and Harvard T.H. Chan School of Public Health (partially supported by training grants from Pfizer, Takeda, Bayer and PhRMA). DHS receives salary support from NIH-K24AR055989. JWJ is funded by the Alonzo Smythe Yerby Fellowship at the Harvrad T.H. Chan School of Public Health. JMF is PI of grants from PCORI and Merck.

ABSTRACT

BACKGROUND: Propensity score matching is a commonly used tool. However, its use in settings with more than two treatment groups has been less frequent. We examined the performance of a recently developed propensity score weighting method in the three treatment group setting.

METHODS: The matching weight method is an extension of inverse probability of treatment weighting (IPTW) that reweights both exposed and unexposed groups to emulate a propensity score matched population. Matching weights can generalize to multiple treatment groups. The performance of matching weights in the three-group setting was compared via simulation to three-way 1:1:1 propensity score matching and IPTW. We also applied these methods to an empirical example that compared the safety of three analgesics.

RESULTS: Matching weights had similar bias, but better mean squared error (MSE) compared to three-way matching in all scenarios. The benefits were more pronounced in scenarios with a rare outcome, unequally sized treatment groups, or poor covariate overlap. IPTW's performance was highly dependent on covariate overlap. In the empirical example, matching weights achieved the best balance for 24 out of 35 covariates. Hazard ratios were numerically similar to matching. However, the confidence intervals were narrower for matching weights.

CONCLUSIONS: Matching weights demonstrated improved performance over three-way matching in terms of MSE, particularly in simulation scenarios where finding matched subjects was difficult. Given its natural extension to settings with even more than three groups, we recommend matching weights for comparing outcomes across multiple treatment groups, particularly in settings with rare outcomes or unequal exposure distributions.

INTRODUCTION

The emergence of multiple treatment options makes the availability of comparative effectiveness/safety evidence more important. However, head-to-head clinical trials are not common, let alone trials of multiple active treatment options. Observational studies can play an important role in filling this gap; however, confounding by indication[1] is a challenge.

Initially proposed in 1983, the propensity score[2] has become a commonly used tool to address confounding in the scientific literature. However, its use in multiple group settings has not received as much attention[3-5]. Rassen *et al*[3] explored an extension of propensity score matching to the three-group setting, developing a three-way simultaneous nearest neighbor matching algorithm (three-way matching). However, simultaneous matching in multiple dimensions is computationally burdensome and often leads to many patients being excluded because appropriate matches are unavailable. Therefore, the extension of this approach to 4 or more groups has not been achieved.

Li and Greene recently proposed a weighting analogue to pairwise 1:1 matching[6] (matching weights), and demonstrated that its estimand is asymptotically equivalent to the estimand of exact pairwise matching on the propensity score, given common support of the propensity score between treatment groups. As compared to matching, efficiency gains were seen in simulations. Therefore, we hypothesized that matching weights generalized to the setting of three treatment groups would outperform three-way matching.

In the current paper, we generalize matching weights to the setting of three or more treatment groups and present a simulation study that compares the validity and precision of matching weights, three-way matching, and inverse probability of treatment weights (IPTW). Finally, we use empirical data to demonstrate its performance in a real-life dataset.

METHODS

Matching weights

Li and Greene's proposed weight is defined as follows for the i-th subject[6]:

$$\text{Matching weight} = \frac{\min(e_i, 1 - e_i)}{Z_i e_i + (1 - Z_i)(1 - e_i)}$$

where

e_i is the propensity score

Z_i is the binary treatment indicator

The denominator is identical to that of IPTW[7], the probability of the assigned treatment given covariates. The numerator is the smallest of the propensity score or its complement, which can be thought of as a combination of the numerator for the average treatment effect on the treated weight ("treated weight")[8,9] and that for the average treatment effect on the untreated weight ("untreated weight")[9]. These weights' close relationships can be appreciated if they are expressed in the same notation as shown below.

$$\text{IPTW} = \frac{1}{Z_i e_i + (1 - Z_i)(1 - e_i)}$$

$$\text{Treated weight} = \frac{e_i}{Z_i e_i + (1 - Z_i)(1 - e_i)}$$

$$\text{Untreated weight} = \frac{1 - e_i}{Z_i e_i + (1 - Z_i)(1 - e_i)}$$

Matching weights reduce to the treated weights for those with propensity scores < 0.5, untreated weights for those with propensity scores > 0.5, and at propensity scores = 0.5, matching weights agree with both.

A simulated dataset may help intuitive understanding (**Figure 1-1**). Compared to the IPTW method, which up-weights subjects to balance the distributions of the propensity score, matching weights instead down-weight subjects to achieve balance. In this example, the treated group is as large as the untreated group, making the target of matching weights and 1:1 matching depart from the treated group. If there is a large reservoir of untreated[10], however, most observations fall below propensity score < 0.5, making both matching weights and 1:1 matching approximate the treated group similarly well (**eFigure 1-1**). Matching weights confer numerical stability compared to IPTW, which can suffer from very high weights, by focusing on treatment effects in patients with good overlap on the propensity score.[6] Compared to matching, matching weights are more efficient because they use all of the original data.

Generalization of matching weights

Unlike matching, weighting methods can naturally generalize to a non-dichotomous treatment variable, including three or more treatment groups. For matching weights under K treatment groups, the weight can be generalized as follows.

$$\text{Matching weight} = \frac{\min(e_{1i}, \dots, e_{Ki})}{\sum_{k=1}^{K} I(Z_i = k) e_{ki}}$$

where

e_{ki} is the generalized propensity score for the kth treatment

($i.e.$, probability of receiving the k-th treatment)

$Z_i \in \{1, \dots, K\}$ is a categorical treatment

$I(\cdot)$ is an indicator variable (1 if true and 0 if false)

The denominator is the probability of receiving the treatment actually received given covariates. The numerator considers probabilities for all treatment levels and selects the smallest value. For a given individual, the sum of all propensity scores must add up to 1, meaning that a single model must be fit to the data to estimate all of the propensity scores (*e.g.,* multinomial logistic regression).[4] Again the estimand of this generalized weighting method is asymptotically equivalent to the estimand of exact matching across all treatment groups if common support (*i.e.*, positivity) holds for all treatment levels (proofs in **eAppendix** page 1-9).

Simulation study

Comparison of matching weights[6], stabilized IPTW[13], and the three-way matching method developed by Rassen *et al*[3] was conducted in simulated datasets (details in **eAppendix** pages 10-14).

Data Generation

The data generation mechanism followed Franklin *et al*.[14] The outcome was binary, and the treatment took on three values. There were ten confounders (binary and continuous). Levels of covariate overlap, treatment prevalence, baseline outcome risk, treatment effects, and treatment effect modification were varied. Each dataset had 6,000 subjects. Treatment assignment ($T_i \in \{0, 1, 2\}$) was generated as a multinomial random variable based on true propensity scores. We generated all combinations of exposure prevalence {33:33:33, 10:45:45, 10:10:80} and weak (near-random treatment assignment; good covariate overlap) and strong (non-random treatment assignment; poor covariate overlap) covariate-treatment associations.

All covariates and treatment jointly determined the true probability of disease for each subject. The counterfactual probability of disease under each treatment was also recorded. To avoid non-collapsibility issues[15,16], a log-probability model was used.

$$\begin{aligned}&\log(\mathrm{P}(Y_i = 1 | T_i = t_i, \boldsymbol{X}_i = \boldsymbol{x}_i))\\ &\qquad = \beta_0 + \boldsymbol{x}_i^T \boldsymbol{\beta}_{\boldsymbol{X}} + \beta_{T1} I(t_i = 1) + \beta_{T2} I(t_i = 2) + \beta_{T1X_4} I(t_i = 1) x_{4i} + \beta_{T2X_4} I(t_i = 2) x_{4i}\end{aligned}$$

The bold $\boldsymbol{\beta}_x$ represents main effects of covariates. Treatment has two main effect terms. The last two terms are treatment-X_4 interactions. Treatment 0 served as the baseline for comparison, and Treatments 1 and 2 had no effects or protective effects. The intercept β_0 was manipulated to achieve the baseline disease risk of 5% or 20%. We controlled treatment effect heterogeneity by setting the coefficients of the interaction terms to either zeros (no heterogeneity) or negative (additional protective effect for individuals with $x_{4i} = 1$). Combining these simulation parameters, we constructed 48 simulation scenarios (**eAppendix** page 13). Each scenario was run 1,000 times.

Propensity score estimation

For each simulated dataset, the propensity score model including all covariates was fit by multinomial logistic regression[17]. For each subject, three propensity scores (e_{0i}, e_{1i}, and e_{2i}) were estimated.

Matching weight procedure

Weights were estimated from three propensity scores. Subsequent analyses, including balance metrics and risk regression (modified Poisson regression[18]), were conducted as weighted analyses[19,20]. The treatment variable was the only predictor in the outcome model. The estimation using the stabilized IPTW was conducted similarly substituting the weights. Reproducible example R code is provided in **eAppendix** (pages 15-21).

Three-way matching procedure

Using non-redundant propensity scores to define a two-dimensional propensity score space, three-way matching was conducted without replacement[3]. The Pharmacoepidemiology Toolbox version 2.4.15 (http://www.drugepi.org) was used. The caliper width was based on the perimeter of the triangle formed by three individuals in a proposed matched trio.[3] The maximum allowed perimeter was:

$$0.6 \times \sqrt{\frac{\tau_0^2 + \tau_1^2 + \tau_2^2}{3}}$$

where

$$\tau_k^2 = \frac{\mathrm{Var}(e_{0i} \mid T = k\ ) + \ \mathrm{Var}(e_{1\mathrm{i}} \mid T = k\ )}{2}$$

Modified Poisson regression[18] was conducted without stratifying on matched trios to maintain the unconditional estimand comparable to that of matching weights.

Performance assessment metrics

Several assessment metrics were used to examine validity and efficiency: weighted or matched sample size, covariate balance measured by absolute standardized mean differences,[21,22] bias in risk ratios, simulation variance, estimated variance, mean squared errors (MSE), false positive rates in null scenarios, and coverage probability of confidence intervals. Bias and covariate balance, which measures the potential for confounding bias, are measures of validity, whereas variance is a measure of efficiency.

Standardized mean differences were calculated for three pairwise contrasts and averaged for each covariate. The standardization was conducted by dividing the mean difference by the square root of the pooled within-group variance (Its definition for binary variables is explained in references).[21,22]

Bias for an effect estimate was defined as the average risk ratio estimate / the true risk ratio. The true risk ratio (estimand) was calculated as the contrast of the marginal counterfactual outcomes (average of the counterfactual probabilities of disease across individuals under each treatment). This true risk ratio calculation was conducted in the unadjusted cohort (for the average treatment effect), matching weight cohort, three-way matched cohort, and IPTW cohort (this should agree with the average treatment effect) to obtain their respective estimands. These adjusted cohorts were newly constructed using the true propensity scores to avoid the influence of the propensity score estimation model performance. The estimands themselves were also compared for their agreement under treatment effect heterogeneity.

The simulation variance is the variance of the estimator across simulation iterations, and represents the true variance of the estimator, whereas the estimated variance was calculated within each iteration and average across all iterations. The bootstrap variance was calculated for matching weights only due to computational burden. The full sequence of propensity score modeling and outcome modeling was bootstrapped[12]. For each one of 1,000 iterations of a given scenario, 1,000 bootstrap iterations were conducted. MSE combines bias and true variance (variance + bias^2). False positive rates were examined in the null scenarios where there was no treatment effect and no treatment effect heterogeneity. The confidence intervals created from the estimated variance were examined for their coverage of the aforementioned true risk ratios to see whether these intervals are conservative in nature by ignoring uncertainty in the estimated propensity score[6,11].

Empirical study

We re-analyzed Medicare data from a previously published study comparing new users of opioids, COX-2 selective inhibitors (coxibs), and non-selective non-steroidal anti-inflammatory drugs (nsNSAIDs)[3,23] for various safety outcomes. There were 35 covariates including 5 continuous variables. The propensity score model was pre-specified as a model with squared terms for the continuous variables without any interaction terms. All-cause mortality, any fracture, upper or lower gastrointestinal bleeding, and any cardiovascular events were examined.

The baseline covariates for each treatment group before and after weighting (or matching) were examined. Average standardized mean difference across all three pairwise contrasts was calculated for each variable. For the outcome analyses using Cox models, hazard ratios with corresponding 95% confidence intervals were calculated and compared between methods for each outcome event.

Computing

All analyses were conducted in R (http://cran.r-project.org) versions 3. All code for the simulation study is available online (https://github.com/kaz-yos/mw).

RESULTS

Simulation study

Sample sizes

Sample size comparison is presented in **Figure 1-2**. The matching weight sample sizes and the matched sample sizes were similar given Rassen *et al.*'s caliper configuration. They were influenced by both the treatment prevalence and covariate overlap because the size of the common support and number of 1:1:1 matches are influenced by these factors. This means their estimands are similarly affected by the characteristics of the dataset. The unmatched sample size and the stabilized IPTW sample size coincide regardless of the treatment prevalence and covariate overlap. This agrees with the fact that the stabilized IPTW estimates the effect in the entire cohort rather than a subset as in matching weights and matching.

Covariate balance

Figure 1-3 shows the covariate balance before and after balancing by the different methods. In the good covariate overlap setting where there was a minor imbalance to start with, all methods did well, making all standardized mean differences well below the conventional 0.10 threshold[21]. Among the three methods, matching weights achieved the best balance with near-zero standardized mean differences for all covariates followed by IPTW. In the poor covariate overlap setting, *i.e.*, a setting with positivity violation (some subjects exist outside the common support), IPTW broke down, indicating the entire cohort estimand is likely not estimable in this setting. In comparison, both matching weights and matching performed reasonably well, likely because of their emphasis on the effect in the common support.

Bias of estimators

eFigure 1-2 shows the biases of these methods with respect to their corresponding estimands (1.0 means unbiased). The biases were similarly small for all methods in the good covariate overlap settings. In the poor covariate overlap settings, however, their performance differed. Most noticeably IPTW sometimes gave more biased results than the unadjusted analyses, confirming the difficulty of estimating the effect in the entire cohort in such settings. Both matching weights and three-way matching performed reasonably well in all settings, although in the rare outcome setting, matching weights tended to perform better.

Comparison of estimands

eFigure 1-3 shows the estimands (true risk ratios to be estimated) of these methods in different settings. In the absence of effect modification (left half of the figure), their estimands numerically agree. In the presence of effect modification, they may differ substantially. IPTW by definition has the entire cohort as its target of inference (thus, the agreement between U and Ip in the figure). The estimands of matching weights and three-way matching agreed as expected, but they differed from the IPTW estimand particularly in the unbalanced exposure settings. On the other hand, their estimands were close to each other with good covariate overlap and the 33:33:33 exposure distribution (*i.e.*, a setting in which the matching weight or matched sample sizes are close to the entire cohort).

Variance and MSE of estimators

The matching weight estimator had smaller true variance than the three-way matching estimator, particularly in poor covariate overlap settings (**eFigure 1-4**). In these settings, matching yields a small matched cohort, whereas matching weights do leverage data from all subjects although the weighted cohort is similarly small. The difference was most striking in the poor covariate overlap, rare disease, 10:45:45 treatment distribution scenario. This difference was caused by lack of any observed events in treatment group 2 in the matched cohort in some datasets. The estimated variance (**eFigure 1-5**) showed a similar pattern but was sometimes anti-conservative for all methods in poor overlap scenarios. The bootstrap variance for matching weights was less often anti-conservative (**eFigure 1-6**). Since the bias was small, MSE (e**Figure 1-7**) also showed a similar pattern. Importantly, matching weight MSE was always smaller than matching MSE across all scenarios.

False positive rates and coverage

Matching weights had false positive rates > 0.05 for 6 scenarios whereas three-way matching had them for 5 scenarios (**eFigure 1-8**). Undercoverage (coverage < 0.94) was observed in 7 scenarios for matching weights and 3 scenarios for three-way matching (**eFigure 1-9**). For matching weights, undercoverage occurred in poor covariate overlap scenarios only, whereas two of the undercoverage scenarios for three-way matching were in good overlap scenarios.

Empirical study

In the three-group analgesic example, there were 23,647 potentially eligible patients before weighting or matching. After matching weights, the weighted sample size was 13,887.9, which was similar to the three-way matched sample size of 13,833, whereas IPTW resulted in a weighted sample size of 23,699.4, which was similar to the original cohort size. Individuals' assigned weights ranged from 0.0003 to 1 with a median of 0.577 [interquartile range: 0.318-0.897] for matching weights, and 0.241-12.938 with a median of 0.939 [interquartile range:

0.809-1.126] for stabilized IPTW. As seen in **eFigure 1-10**, matching weights achieved the best covariate balance most consistently (24 of the 35 covariates) compared to three-way matching (6 covariates) and IPTW (5 covariates). Thanks to the active comparator design[24], the covariate overlap was relatively good (relatively small standardized mean difference in the unmatched cohort), and IPTW did not break down.

The characteristics of the matching weights cohort and the matched cohorts for selected variables with most imbalances were very similar (**eTable 1-1**), again confirming the notion that matching weights are a weighting analogue to matching. As expected from the definition of the common support (overlap area of all three groups), these cohorts are most similar to the smallest group, *i.e.*, the NSAIDs group in the unmatched cohort. The IPTW cohort had somewhat different characteristics with higher morbidity levels, most closely resembling the largest group, *i.e.*, the opioids group.

The outcome model results are shown in **Table 1-1**. The hazard ratios were similar using matching weights and three-way matching, but IPTW sometimes differed. Between matching weights and three-way matching, the most noticeable difference was in the opioids-vs-nsNSAIDs comparison for the gastrointestinal bleeding outcome, which was the rarest outcome among the four considered in the current study. The standard errors were smaller for matching weights than for three-way matching or IPTW for all estimates, as reflected by the somewhat narrower confidence intervals.

DISCUSSION

We examined the usefulness of a recently proposed weighting method[6] in multiple treatment arm settings, comparing it to the previously described three-way matching method[3] as well as IPTW[25] in both simulated data and a reanalysis of a previously published empirical study.[23] Overall, matching weights provided smaller MSE than three-way matching in the scenarios studied mainly due to smaller variance. Better MSE was more pronounced in settings where matching performed poorly, such as with rare disease and poor covariate overlap. Compared to IPTW, matching weights demonstrated robustness to poor covariate overlap. The false positive rate and coverage rate for matching weights were somewhat less ideal than three-way matching, indicating the need for the bootstrap variance. In the empirical data analysis, matching weights gave similar point estimates compared to three-way matching, but with better covariate balance and narrower confidence intervals.

The strengths of matching weights are the combination of the strengths of matching and weighting. The estimand of the matching weight estimator is asymptotically equivalent to that of 1:1 exact matching. We confirmed that this approximately holds in finite datasets using nearest-neighbor matching (**eFigure 1-3**). Those who are nearly equally likely to receive all treatment choices are most represented (**Figure 1-1**). Matching weights avoid inflating weights for a small number of subjects in the extremes of the propensity score distribution treated contrary to the norm, which is a major disadvantage of typical IPTW approaches.

From weighting, matching weights inherit the maximum use of the data, *i.e.*, no one in the dataset is left out, but subjects contribute differing amounts of information depending on their weights. The efficient use of data resulted in lower variance of estimators in our simulation and narrower confidence intervals in our empirical study. As with other weighting methods, matching weights can naturally generalize to multiple treatment group settings, which we demonstrated in this paper. Currently, there appears to be no software available for 4+ group simultaneous matching, which matching weights can easily accommodate.

Matching weights outperformed IPTW in scenarios with poor covariate overlap; however, choice of a method should carefully consider both the clinical question and the data (**Table 2**). Although matching weights are an extension of IPTW, their targets of inference are different as illustrated in **Figure 1-1**. Their estimands (true risk ratios to be estimated) numerically agree if no treatment effect heterogeneity exists (left half of **eFigure 1-3**), and they nearly coincide if covariate overlap is good (first and third rows of **eFigure 1-3**). However, their estimands are not directly comparable in settings with treatment effect heterogeneity as demonstrated in the right half of **eFigure 1-3**, particularly if covariate overlap is poor (second and fourth rows). When making a decision about which propensity score method to employ, the estimand should be decided first based on the clinical question. If it is the causal effect in the entire population, IPTW is the method of choice.

Nonetheless, as seen in the poor covariate overlap simulation scenarios, the performance of IPTW degrades when positivity violations[26] exist because the effect in the entire cohort is not estimable. The IPTW cohort can be "trimmed" to drop subjects who violate positivity, but this will also reduce the effective sample size and modify the target of inference (detailed discussions of propensity score trimming in the two-group setting are in Crump *et al*[27] and Stürmer *et al*[28]). Matching weights and matching approach this problem by focusing on the patients with "empirical equipoise"[29] --*i.e.,* patients for whom all treatment options under study are appropriate. This subset is not easily definable; however, in the setting of 3+ active treatment groups, the average treatment effect on the *treated* is not uniquely defined, justifying focusing on this feasible subset. This subset is also where comparative effectiveness evidence may be most useful for decision-making. In practice, the matching weight cohort, as well as the original cohort, should be presented in the baseline table to clarify the subset of the population for which inference was made.

Another potential approach given three or more groups is to match two groups at a time, resulting in three matched cohorts with different pairs of treatment arms (*i.e.,* to separately target the populations for whom those two treatments are equally possible). These three cohorts are not directly comparable to the one cohort given by matching weights or three-way matching. Whether the former is a more appropriate method depends on the clinical question and situation. The mean matching weight (ranges 0 to 1) in the group that had the smallest unweighted sample size may be used to assess the simultaneous common support. This quantity is roughly interpretable as the fraction of the smallest treatment group in clinical equipoise with the other groups. If this fraction is close to 1, the treatment groups have reasonable overlap and the factor constraining the weighted sample size is the *number* of subjects in the smallest group. On the other hand, if the fraction is close to 0, it is the *lack of sufficient common support* that is constraining the weighted sample size. In the latter setting, the more meaningful questions may be answered by pairwise comparison. If the problem persists with pairwise matching weights, it means not enough common support exists in the data to enable comparative effectiveness research.

There are potential limitations in the current study. We employed the caliper configuration for three-way matching used in the paper by Rassen *et al.*[3] Currently, no known standard exists for caliper definitions (raw propensity score or logit of propensity score) or caliper widths for three-way matching. In the 4 or more group settings, even the distance metric is hard to define. Matching weights, on the other hand, completely avoids the use of an arbitrary caliper parameter. Investigators can instead focus on the structure of the propensity score model.

Matching methods, including three-way matching, are, by definition, protected against common support (positivity) violations at least with narrow matching calipers. Subjects with propensity scores that are not present in other treatment groups cannot match, and are excluded.

This is not true for matching weights, as everybody, even those without exactly comparable subjects in other groups, contributes to the weighted analyses. This is why the theoretical asymptotic equivalence of the estimands of matching weights and matching requires perfect common support in addition to exact matching.[6] However, poor covariate overlap did not adversely affect matching weights in comparison to IPTW, which did not perform well in poor covariate overlap scenarios.

There have been debates about whether to account for the uncertainty in the *estimated* propensity score[11], which are estimates of the true underlying propensity score. Li and Greene found that not accounting for the uncertainty (using estimated propensity scores as if they were known constants) results in conservative variance estimates[6], whereas simultaneous estimation of the propensity score and outcome model parameters gave correct variance estimates. We did not pursue this method, as the generalization to multiple treatment group settings and binary outcomes was unclear. They suggested bootstrapping as an alternative that is easier to implement. In our simulation study in the three-group setting with a binary outcome, matching weight variance estimates were somewhat anti-conservative (smaller than the true variance) in poor covariate overlap scenarios. Bootstrap variance performed more accurately and was less often anti-conservative.

In conclusion, matching weights are a viable alternative to matching, especially with three or more treatment groups. Matching weights demonstrated improved performance over three-way matching in terms of MSE. With good covariate overlap, matching weight estimates were similar to IPTW estimates, although, in such settings, IPTW may be preferable due to its clearer target of inference. Given its natural extension to settings with more than three groups, we recommend matching weights for comparing outcomes across multiple treatment groups when covariate overlap is relatively limited, outcomes are rare, or exposure distributions are unequal. For variance estimation, use of bootstrapping is preferred.

References

1. Greenland S, Neutra R. Control of confounding in the assessment of medical technology. *Int J Epidemiol*. 1980;9(4):361-367.

2. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika*. 1983;70(1):41-55. doi:10.1093/biomet/70.1.41.

3. Rassen JA, Shelat AA, Franklin JM, Glynn RJ, Solomon DH, Schneeweiss S. Matching by propensity score in cohort studies with three treatment groups. *Epidemiology*. 2013;24(3):401-409. doi:10.1097/EDE.0b013e318289dedf.

4. Imbens GW. The role of the propensity score in estimating dose-response functions. *Biometrika*. 2000;87(3):706-710. doi:10.1093/biomet/87.3.706.

5. Hernán MA, Robins JM. Estimating causal effects from epidemiological data. *J Epidemiol Community Health*. 2006;60(7):578-586. doi:10.1136/jech.2004.029496.

6. Li L, Greene T. A weighting analogue to pair matching in propensity score analysis. *Int J Biostat*. 2013;9(2):215-234. doi:10.1515/ijb-2012-0030.

7. Hernán MA, Brumback B, Robins JM. Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. *Epidemiology*. 2000;11(5):561-570.

8. Hirano K, Imbens GW. Estimation of Causal Effects using Propensity Score Weighting: An Application to Data on Right Heart Catheterization. *Health Services & Outcomes Research Methodology*. 2001;2(3-4):259-278. doi:10.1023/A:1020371312283.

9. Sato T, Matsuyama Y. Marginal structural models as a tool for standardization. *Epidemiology*. 2003;14(6):680-686. doi:10.1097/01.EDE.0000081989.82616.7d.

10. Imbens GW. Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review. *Review of Economics and Statistics*. 2004;86(1):4-29. doi:10.1162/003465304323023651.

11. Stuart EA. Matching methods for causal inference: A review and a look forward. *Stat Sci*. 2010;25(1):1-21. doi:10.1214/09-STS313.

12. Austin PC, Small DS. The use of bootstrapping when using propensity-score matching without replacement: a simulation study. *Stat Med*. 2014;33(24):4306-4319. doi:10.1002/sim.6276.

13. Xu S, Ross C, Raebel MA, Shetterly S, Blanchette C, Smith D. Use of stabilized inverse propensity scores as weights to directly estimate relative risk and its confidence intervals. *Value Health*. 2010;13(2):273-277. doi:10.1111/j.1524-4733.2009.00671.x.

14. Franklin JM, Rassen JA, Ackermann D, Bartels DB, Schneeweiss S. Metrics for covariate balance in cohort studies of causal effects. *Stat Med*. 2014;33(10):1685-1699. doi:10.1002/sim.6058.

15. Greenland S, Robins JM, Pearl J. Confounding and Collapsibility in Causal Inference. *Statist Sci*. 1999;14(1):29-46. doi:10.1214/ss/1009211805.

16. Cummings P. The relative merits of risk ratios and odds ratios. *Arch Pediatr Adolesc Med*. 2009;163(5):438-445. doi:10.1001/archpediatrics.2009.31.

17. Yee TW. *VGAM: Vector Generalized Linear and Additive Models*.; 2015. http://cran.r-project.org/web/packages/VGAM/index.html. Accessed July 13, 2015.

18. Zou G. A Modified Poisson Regression Approach to Prospective Studies with Binary Data. *Am J Epidemiol*. 2004;159(7):702-706. doi:10.1093/aje/kwh090.

19. Lumley T. *Complex Surveys: A Guide to Analysis Using R*. 1 edition. Hoboken, N.J: Wiley; 2010.

20. Horvitz DG, Thompson DJ. A Generalization of Sampling Without Replacement from a Finite Universe. *Journal of the American Statistical Association*. 1952;47(260):663-685. doi:10.1080/01621459.1952.10483446.

21. Austin PC. An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. *Multivariate Behav Res*. 2011;46(3):399-424. doi:10.1080/00273171.2011.568786.

22. Austin PC, Stuart EA. Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. *Stat Med*. 2015;34(28):3661-3679. doi:10.1002/sim.6607.

23. Solomon DH, Rassen JA, Glynn RJ, Lee J, Levin R, Schneeweiss S. The comparative safety of analgesics in older adults with arthritis. *Arch Intern Med*. 2010;170(22):1968-1976. doi:10.1001/archinternmed.2010.391.

24. Yoshida K, Solomon DH, Kim SC. Active-comparator design and new-user design in observational studies. *Nat Rev Rheumatol*. 2015;11(7):437-441. doi:10.1038/nrrheum.2015.30.

25. Robins JM, Hernán MA, Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology*. 2000;11(5):550-560.

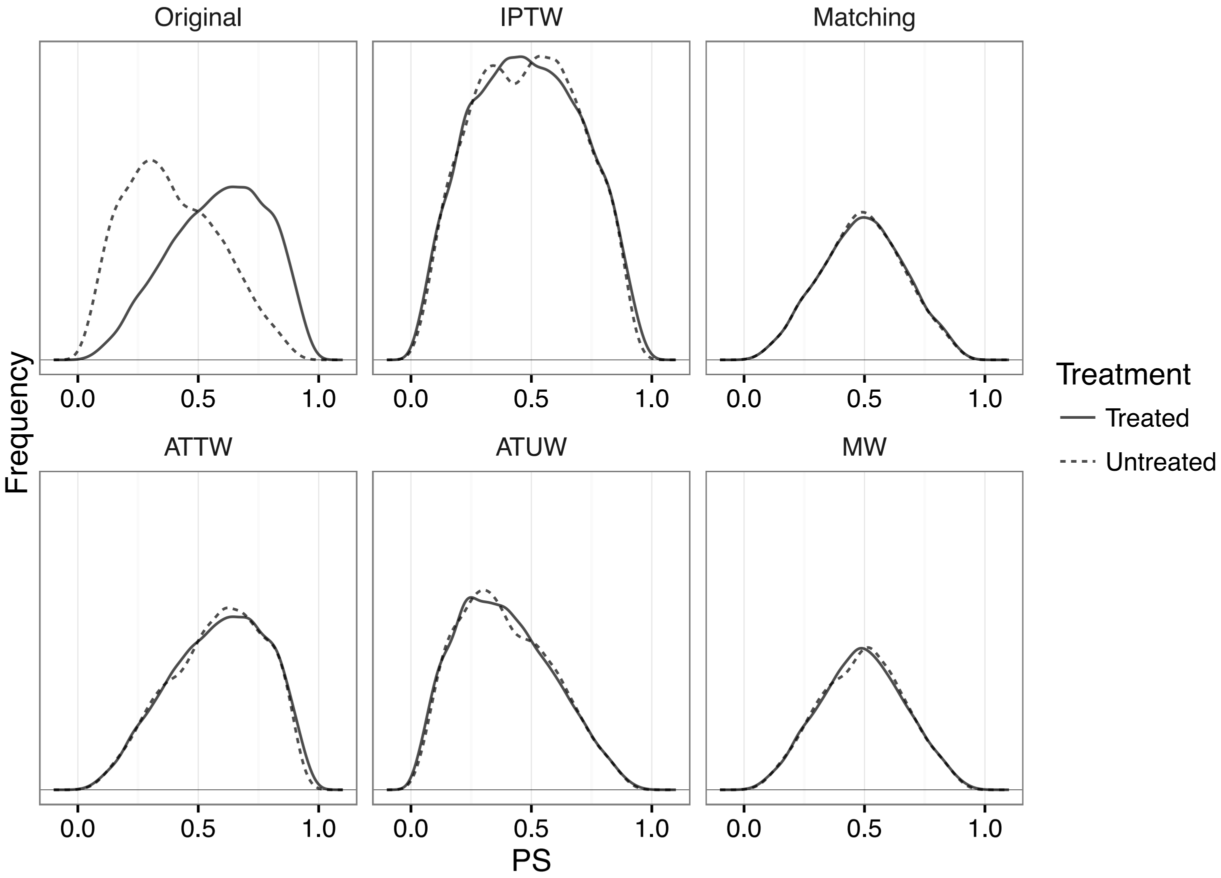
26. Petersen ML, Porter KE, Gruber S, Wang Y, van der Laan MJ. Diagnosing and responding to violations in the positivity assumption. *Stat Methods Med Res*. 2012;21(1):31-54. doi:10.1177/0962280210386207.

27. Crump RK, Hotz VJ, Imbens GW, Mitnik OA. Dealing with limited overlap in estimation of average treatment effects. *Biometrika*. 2009;96(1):187-199. doi:10.1093/biomet/asn055.

28. Stürmer T, Rothman KJ, Avorn J, Glynn RJ. Treatment effects in the presence of unmeasured confounding: dealing with observations in the tails of the propensity score distribution--a simulation study. *Am J Epidemiol*. 2010;172(7):843-854. doi:10.1093/aje/kwq198.

29. Walker A, Patrick, Lauer, et al. A tool for assessing the feasibility of comparative effectiveness research. *Comparative Effectiveness Research*. January 2013:11. doi:10.2147/CER.S40357.

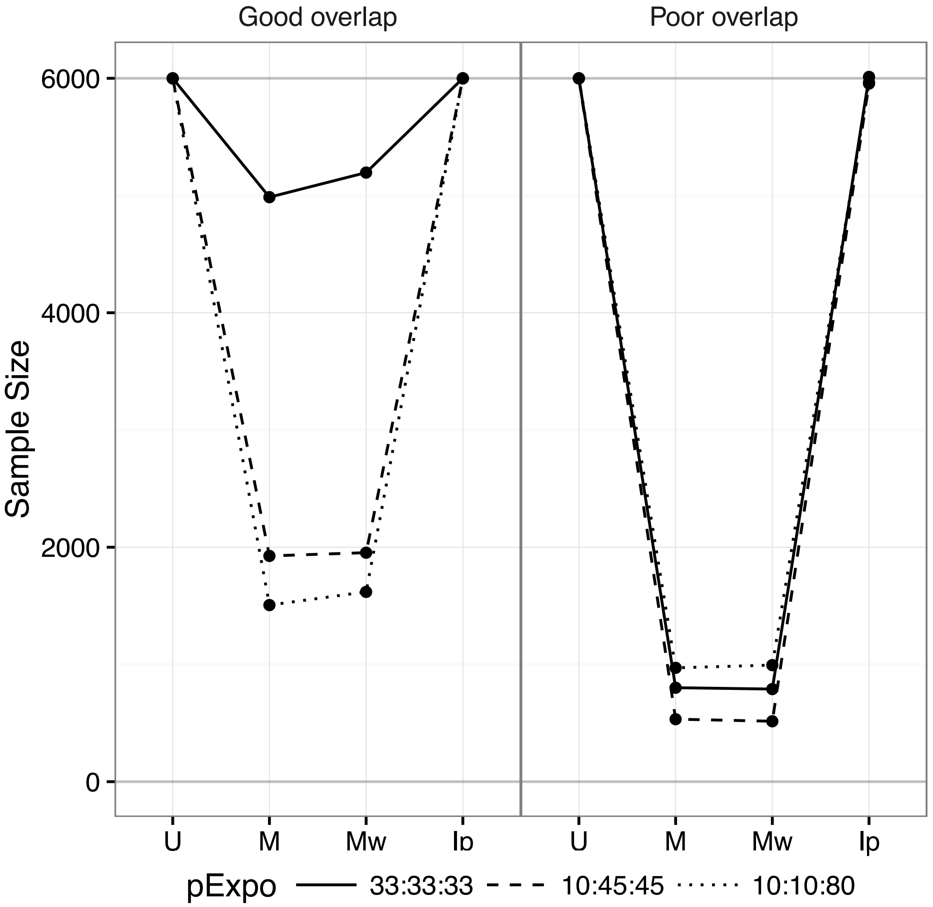
Figure 1-1. Illustration of pre- and post-weighting or post-matching distributions of propensity score when the treatment prevalence is 50%.



The solid line is the distribution of the propensity scores in the treated, and the dashed line is the distribution in the untreated. Matching and matching weight cohorts have a similar propensity score distribution, indicating that their estimands are similar. However, their distributions are substantially different from the original treated group, indicating their departure from the average treatment effect in the treated.

Abbreviations: IPTW: inverse probability of treatment weights.

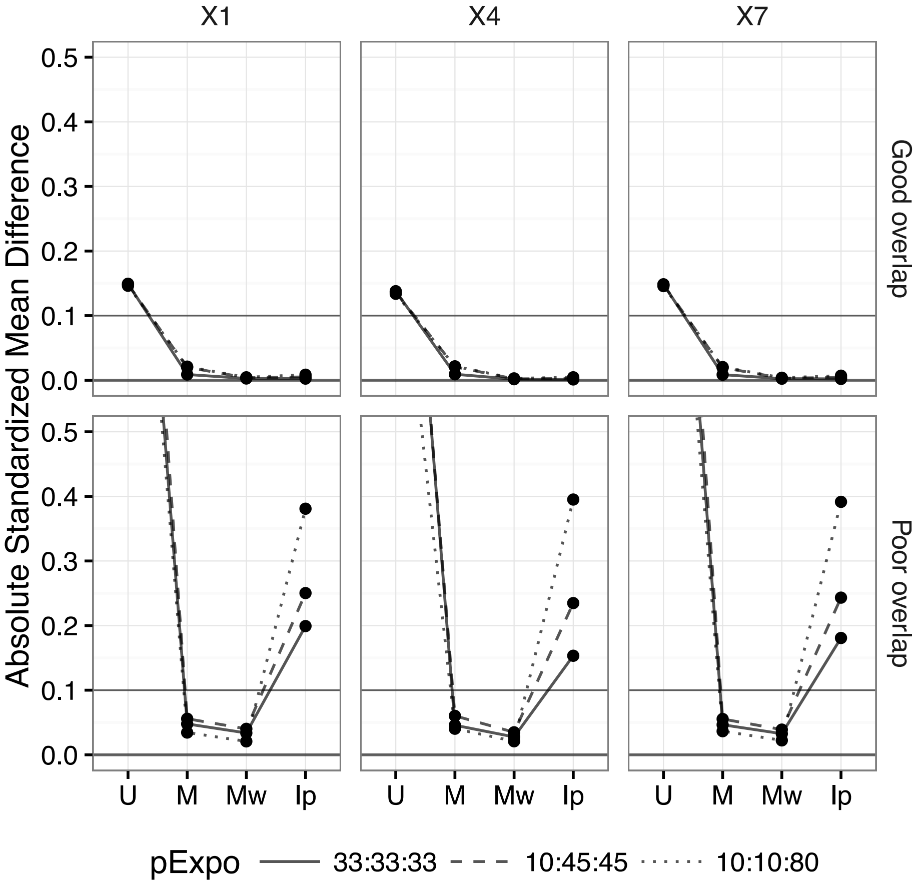
Figure 1-2. Comparison of weighted and matched sample sizes under different levels of covariate overlap.



IPTW shows a weighted sample size identical to the original cohort. Matching weights and matching are similarly affected by exposure prevalence and poor covariate overlap, indicating shifts in the target population.

Abbreviations: U: Unmatched cohort, M: Matched cohort; Mw: matching weight cohort; Ip: Inverse probability of treatment weight cohort; pExpo: Exposure prevalence

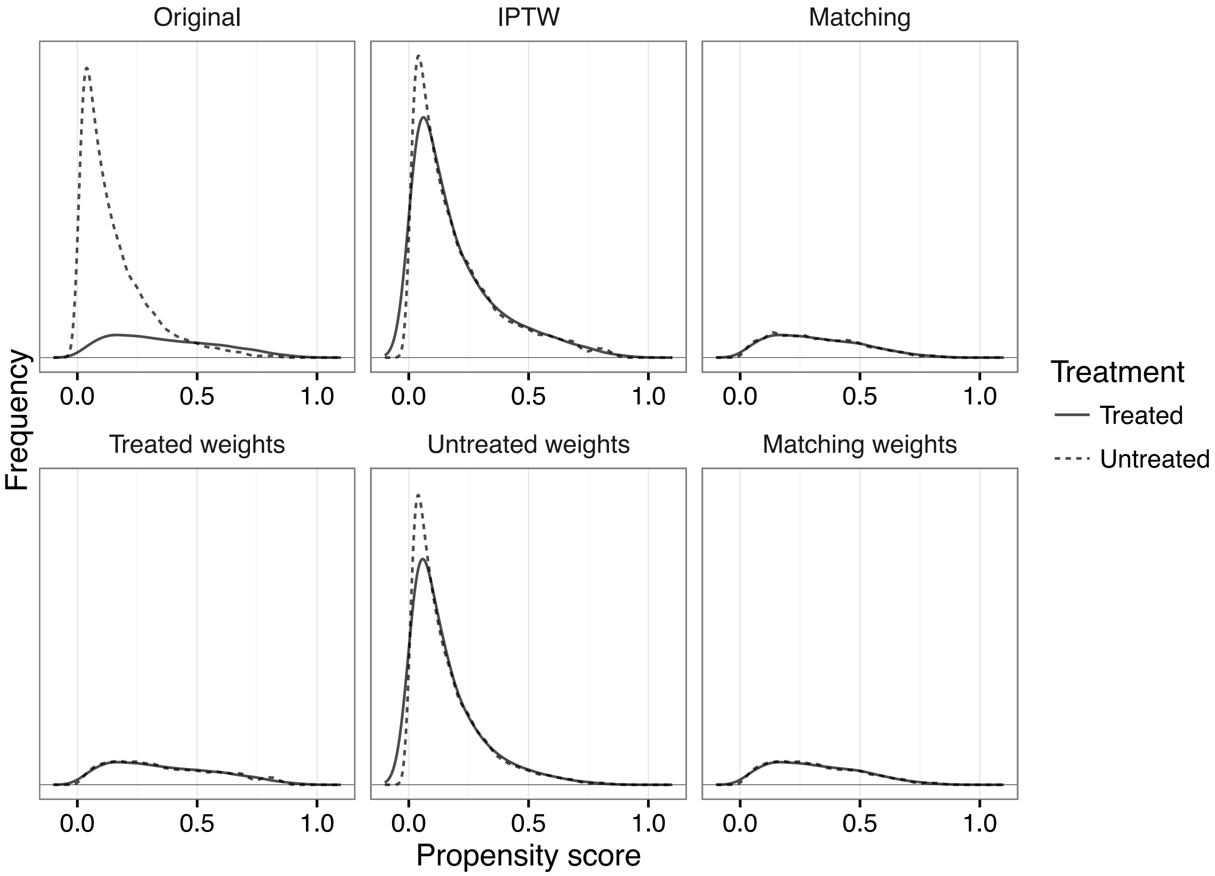
Figure 1-3. Comparison of covariate balance before and after matching or weighting by average standardized mean differences under different covariate overlap (selected covariates: X1, X4, and X7).



MW performs best in both settings, whereas IPTW only works in the good covariate setting. The other covariates showed similar patterns.

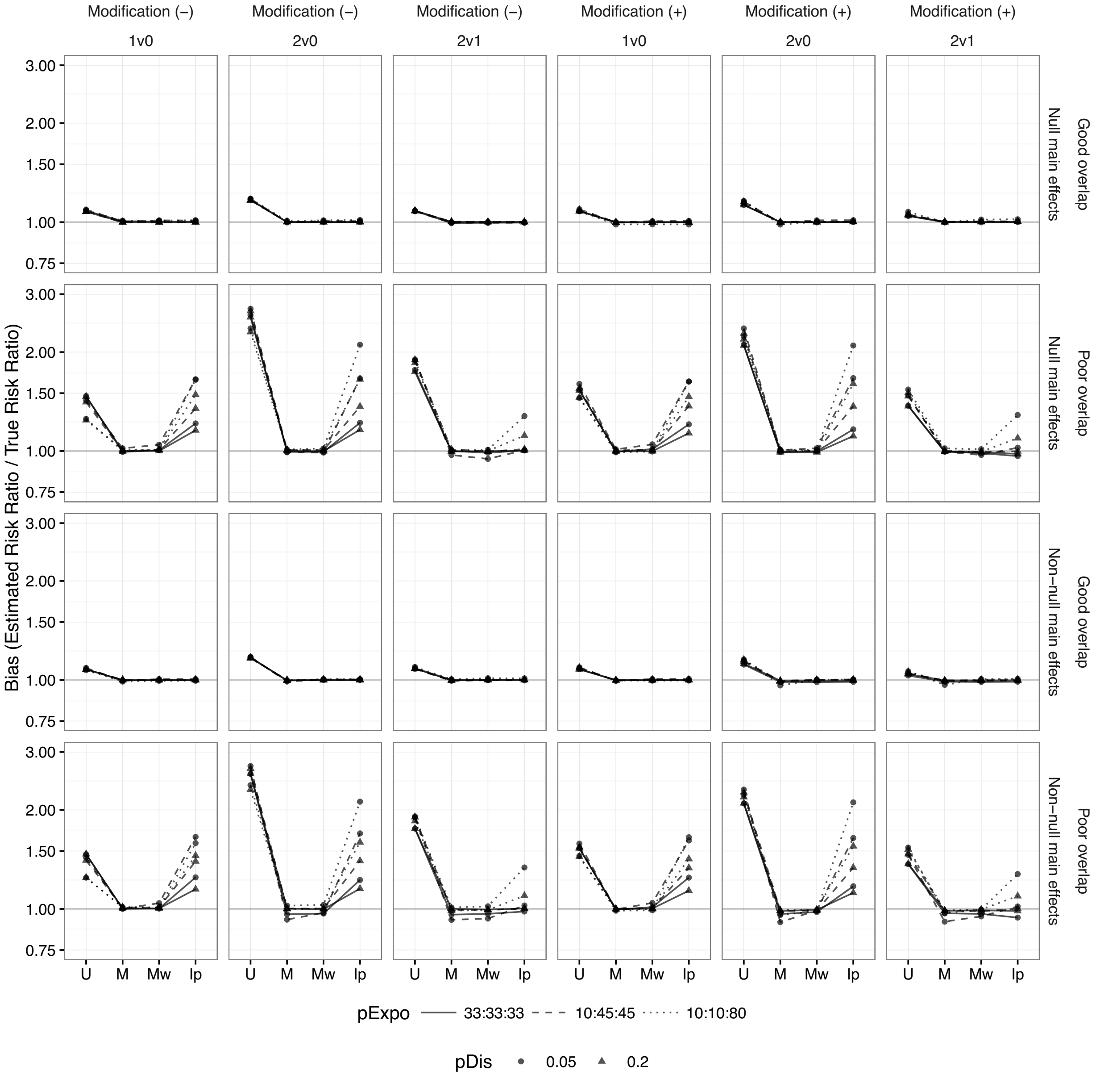
Abbreviations: U: Unmatched cohort, M: Matched cohort; Mw: matching weight cohort; Ip: Inverse probability of treatment weight cohort; pExpo: Exposure prevalence

eFigure 1-1. Illustration of pre- and post-weighting or post-matching distributions of propensity score when the treatment prevalence is 20%. The solid line is the distribution of the propensity scores in the treated, and the dashed line is the distribution in the untreated. Matching and matching weight cohorts have a similar propensity score distribution, indicating that their estimands are similar. These cohorts are very similar to the original treated group (*i.e.*, their estimands approximate the average treatment effect on the treated) although there is a minor attrition in the cohort in the high propensity score range (propensity score > 0.5).



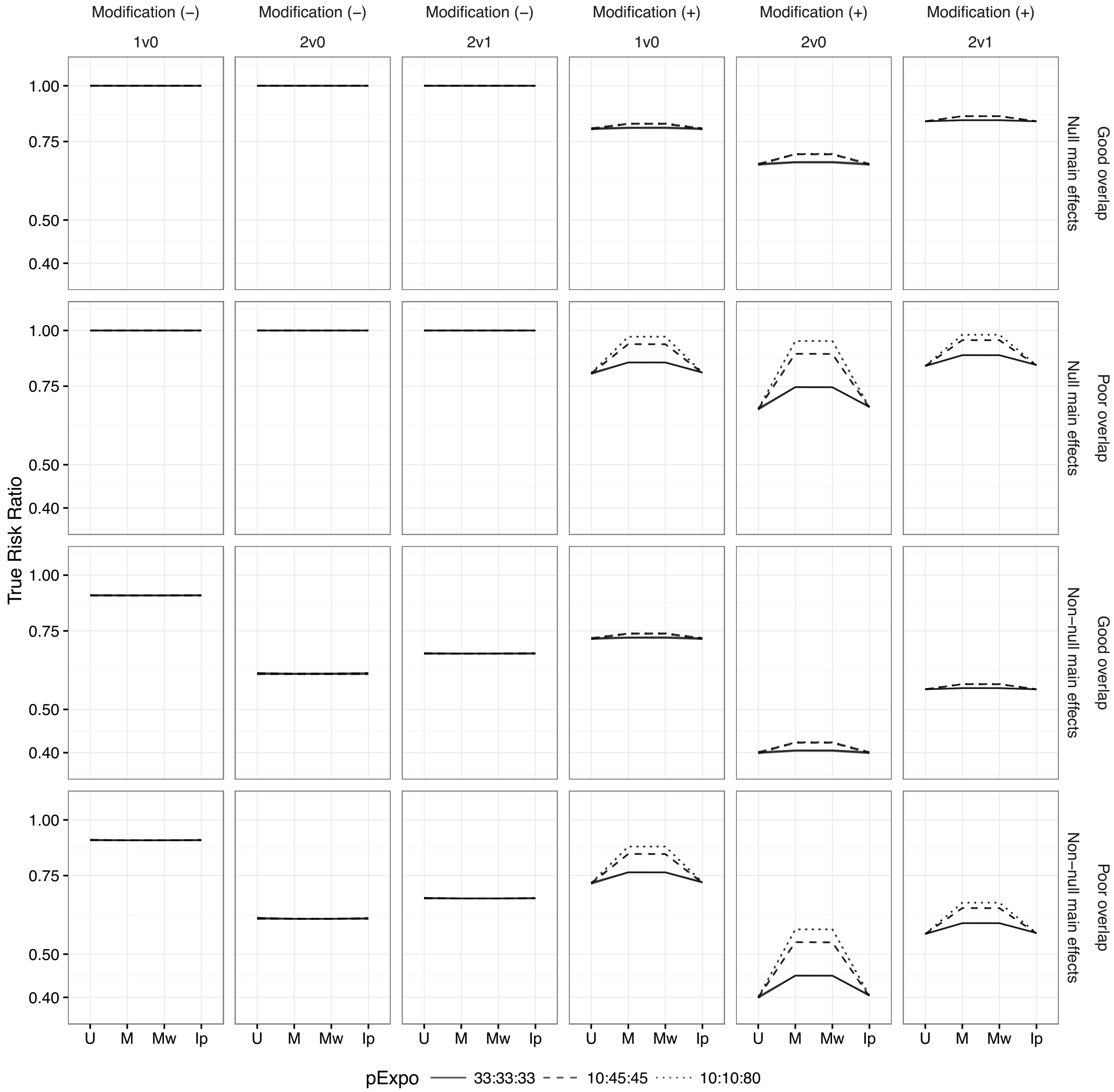
Abbreviations: IPTW: inverse probability of treatment weights.

eFigure 1-2. Comparison of bias (risk ratio / true risk ratio) between methods across 48 scenarios. The left half presents the constant treatment effect scenarios, whereas the right half presents treatment effect heterogeneity scenarios. Each three columns represent three treatment contrasts. Rows classify scenarios by good vs. poor covariate overlap levels and presence vs. absence of main effects. Each panel contains six lines classified by the exposure prevalence and the baseline risk. Matching weights and matching perform well in all scenarios; however, IPTW fails in the poor covariate overlap setting.



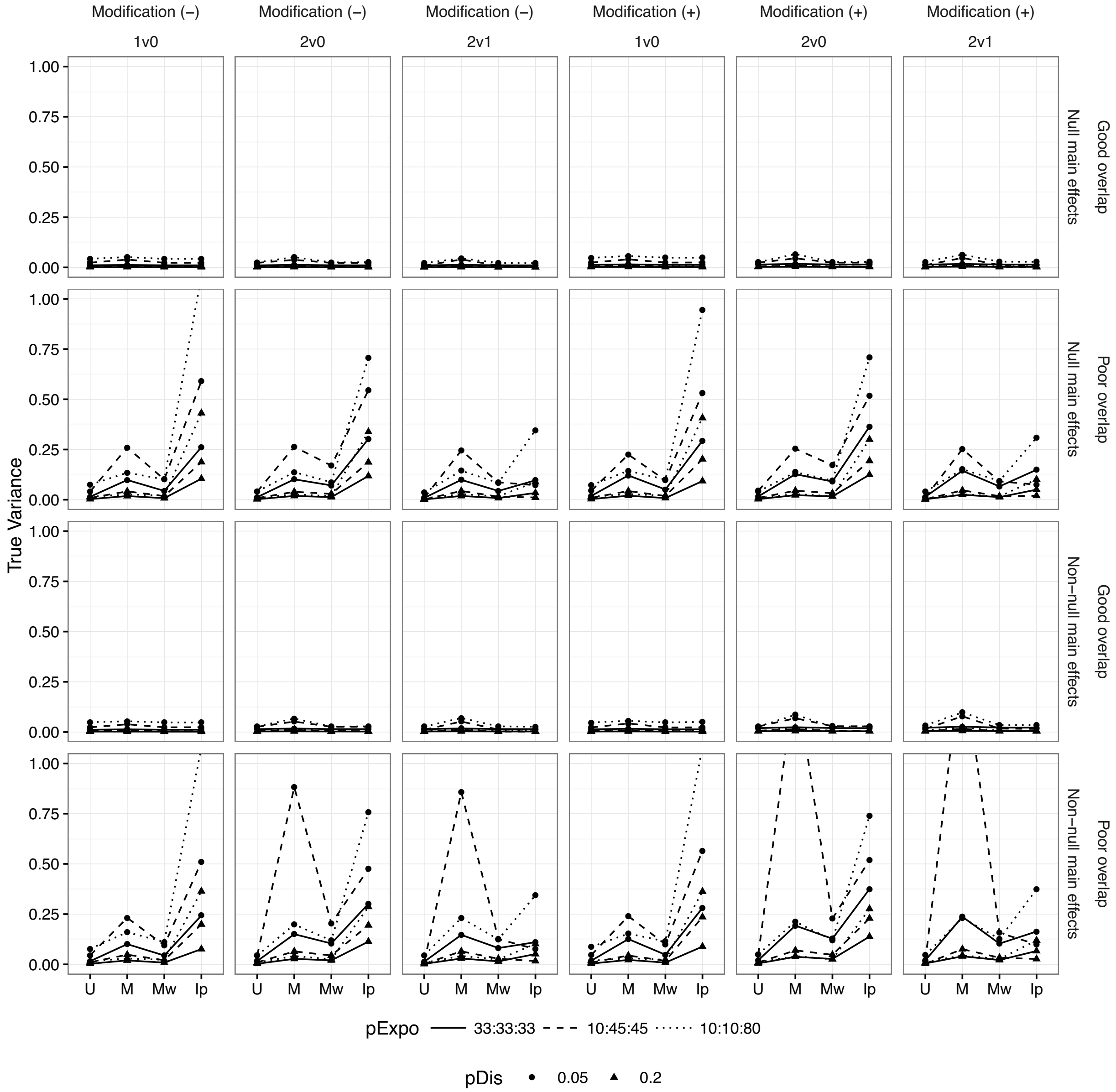
Abbreviations: U: Unmatched cohort, M: Matched cohort; Mw: matching weight cohort; Ip: Inverse probability of treatment weight cohort; pExpo: Exposure prevalence; pDis: Baseline risk of disease.

eFigure 1-3. Comparison of true risk ratios (estimands) between methods across 48 scenarios. Some scenarios have the same estimands and completely overlap. The left half presents the constant treatment effect scenarios, whereas the right half presents treatment effect heterogeneity scenarios. Each three columns represent three treatment contrasts. Rows classify scenarios by good vs. poor covariate overlap levels and presence vs. absence of main effects. Differences in estimands are only present in the treatment effect heterogeneity scenarios, particularly with poor covariate overlap and unbalanced treatment group sizes.



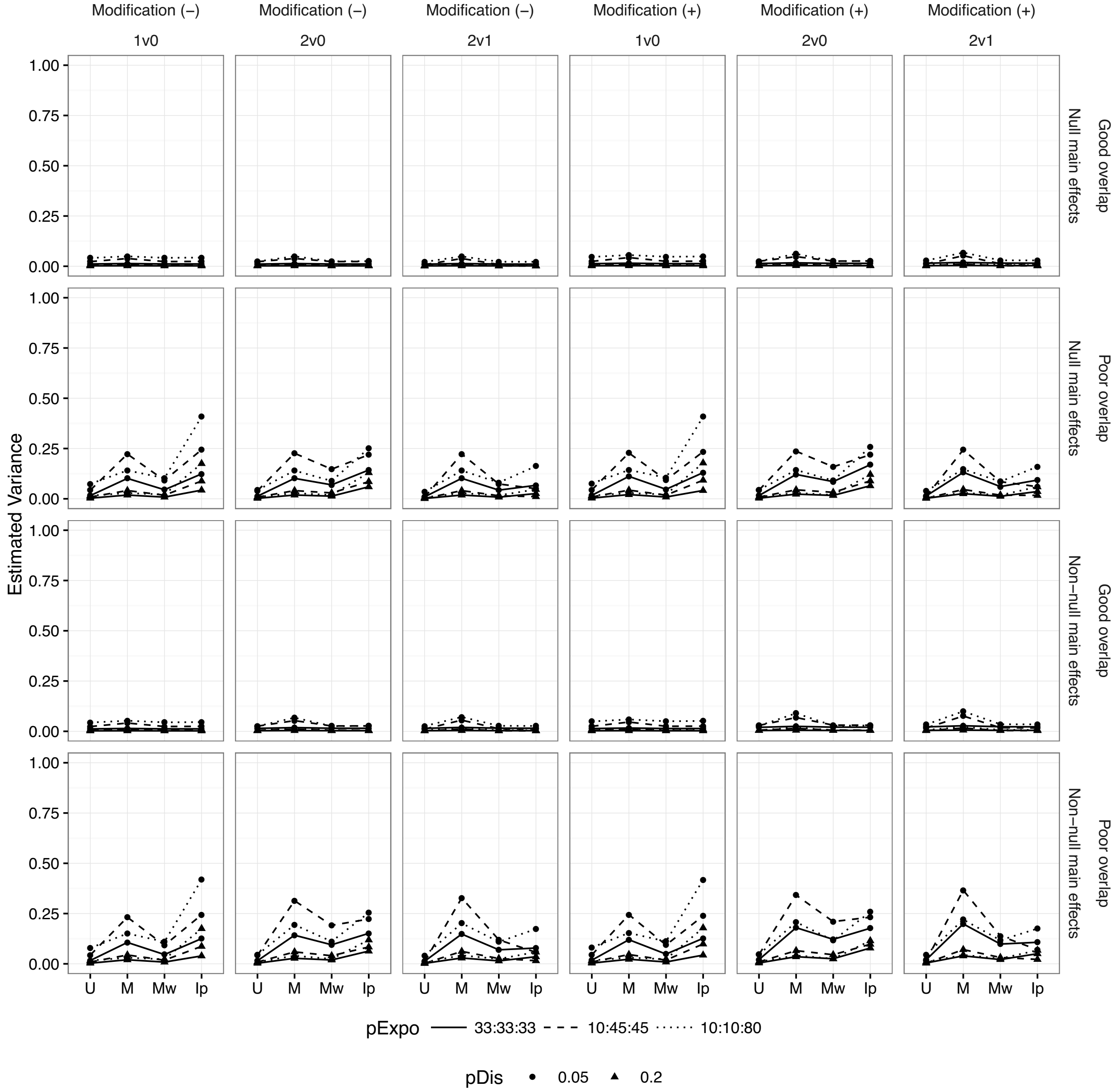
Abbreviations: U: Unmatched cohort, M: Matched cohort; Mw: matching weight cohort; Ip: Inverse probability of treatment weight cohort; pExpo: Exposure prevalence.

eFigure 1-4. Comparison of true variance of log risk ratios calculated across iterations between methods across 48 scenarios. The left half presents the constant treatment effect scenarios, whereas the right half presents treatment effect heterogeneity scenarios. Each three columns represent three treatment contrasts. Rows classify scenarios by good vs. poor covariate overlap levels and presence vs. absence of main effects. Each panel contains six lines classified by the exposure prevalence and the baseline risk. All methods performed well in the good covariate overlap scenarios; however, matching weights were most efficient in the poor covariate overlap scenarios (rows 2 and 4). Matching performed poorly in the poor covariate overlap with 10:45:45 exposure distribution, as there were often no events in Group 2 after matching.



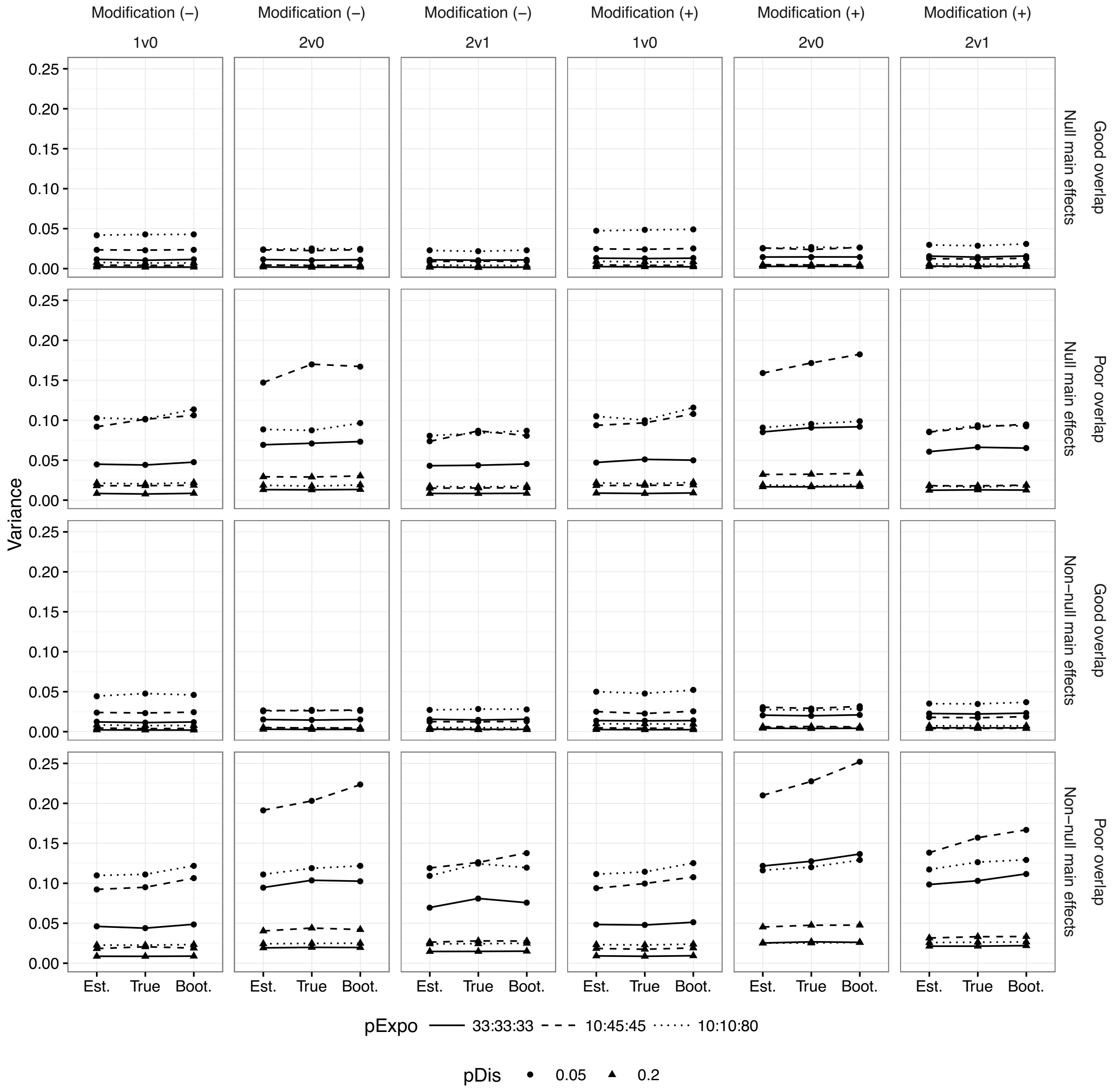
Abbreviations: U: Unmatched cohort, M: Matched cohort; Mw: matching weight cohort; Ip: Inverse probability of treatment weight cohort; pExpo: Exposure prevalence; pDis: baseline risk of disease

eFigure 1-5. Comparison of estimated variance of log risk ratios averaged across iterations between methods across 48 scenarios. The left half presents the constant treatment effect scenarios, whereas the right half presents treatment effect heterogeneity scenarios. Each three columns represent three treatment contrasts. Rows classify scenarios by good vs. poor covariate overlap levels and presence vs. absence of main effects. Each panel contains six lines classified by the exposure prevalence and the baseline risk. Results were similar to the true variance results.



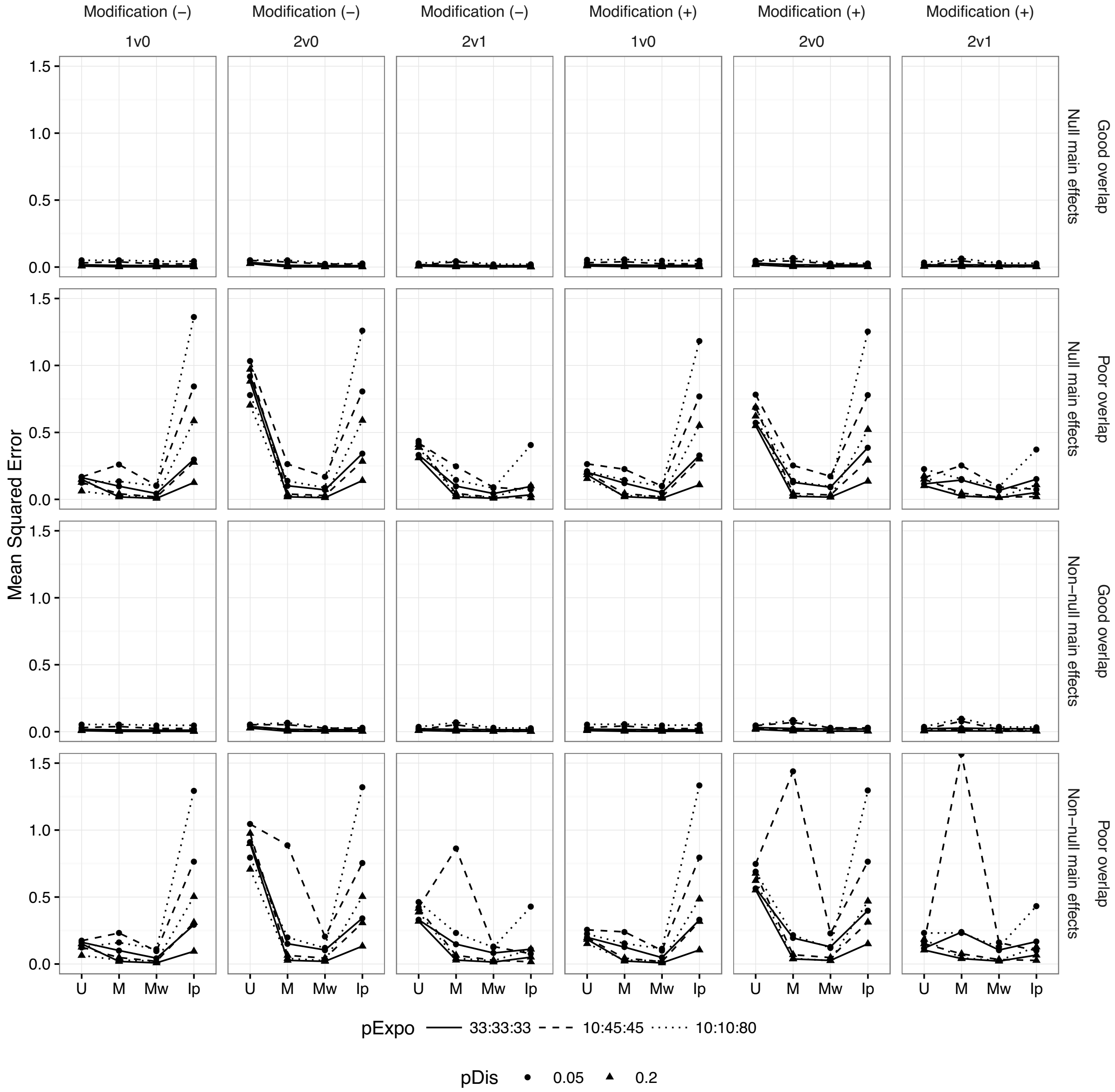
Abbreviations: U: Unmatched cohort, M: Matched cohort; Mw: matching weight cohort; Ip: Inverse probability of treatment weight cohort; pExpo: Exposure prevalence; pDis: baseline risk of disease

eFigure 1-6. Comparison of variance estimation methods for matching weights across 48 scenarios. The left half presents the constant treatment effect scenarios, whereas the right half presents treatment effect heterogeneity scenarios. Each three columns represent three treatment contrasts. Rows classify scenarios by good vs. poor covariate overlap levels and presence vs. absence of main effects. Each panel contains six lines classified by the exposure prevalence and the baseline risk. In good covariate overlap settings, the estimated variance and the bootstrap variance were both close to the true variance values. In the poor covariate overlap settings, however, the estimated variance was sometimes anti-conservative, whereas the bootstrap variance was more accurate or somewhat conservative.



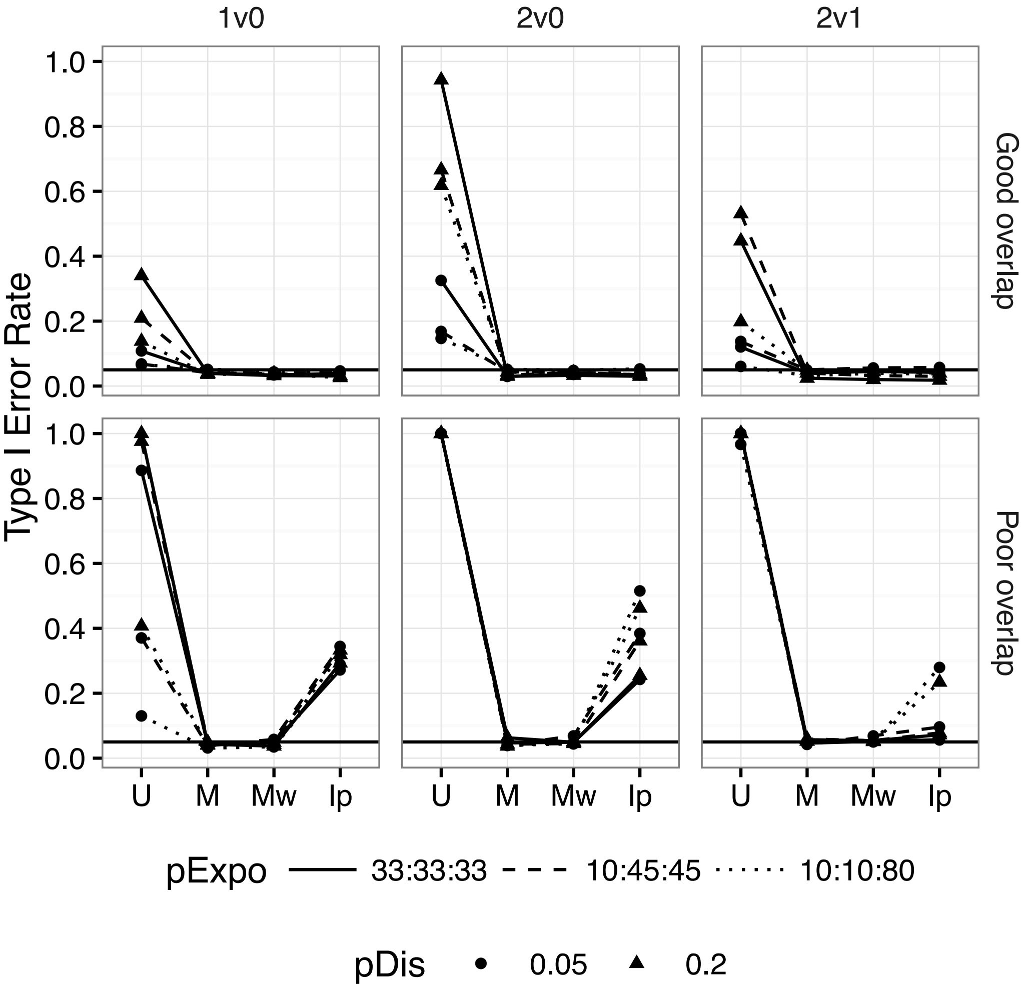
Abbreviations: Est.: Estimated variance; True: True variance calculated across iterations; Boot.: Bootstrap variance; pExpo: Exposure prevalence; pDis: baseline risk of disease.

eFigure 1-7. Comparison of mean squared error of log risk ratios between methods across 48 scenarios. The left half presents the constant treatment effect scenarios, whereas the right half presents treatment effect heterogeneity scenarios. Each three columns represent three treatment contrasts. Rows classify scenarios by good vs. poor covariate overlap levels and presence vs. absence of main effects. Each panel contains six lines classified by the exposure prevalence and the baseline risk. All methods performed well in the good covariate overlap scenarios; however, matching weights were most robust in the poor covariate overlap scenarios (rows 2 and 4). Matching performed poorly in the poor covariate overlap with 10:45:45 exposure distribution, as there were often no events in Group 2 after matching.



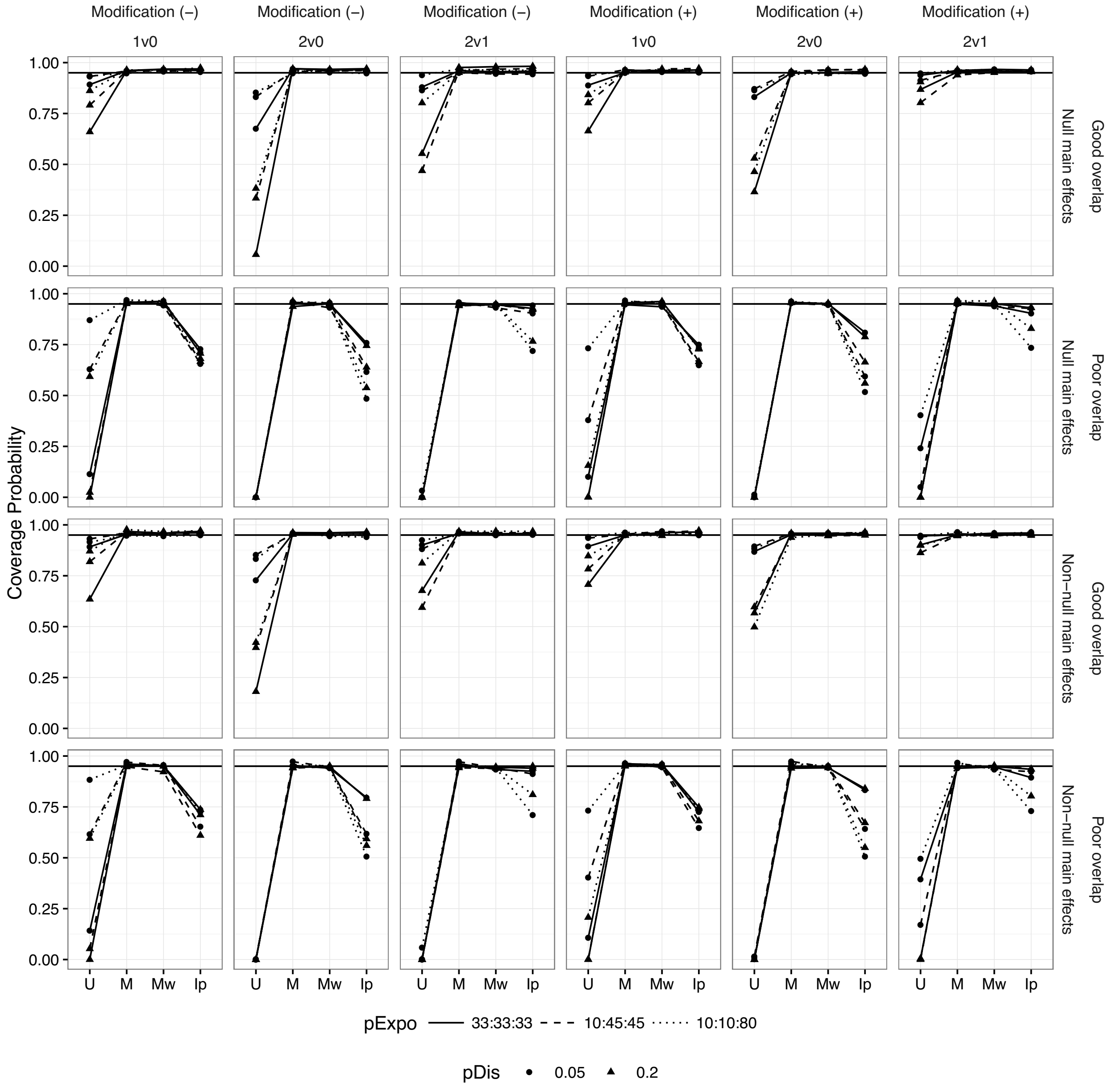
Abbreviations: U: Unmatched cohort, M: Matched cohort; Mw: matching weight cohort; Ip: Inverse probability of treatment weight cohort; pExpo: Exposure prevalence; pDis: Baseline risk of disease.

eFigure 1-8. Comparison of false positive probability in completely null treatment effect scenarios. Minor violation of the 0.05 expected false positive rate (false positive rates of 0.06-0.07) was seen in both matching weights and matching. IPTW made many false positives in the poor covariate overlap settings. These tests were based on the estimated variance.



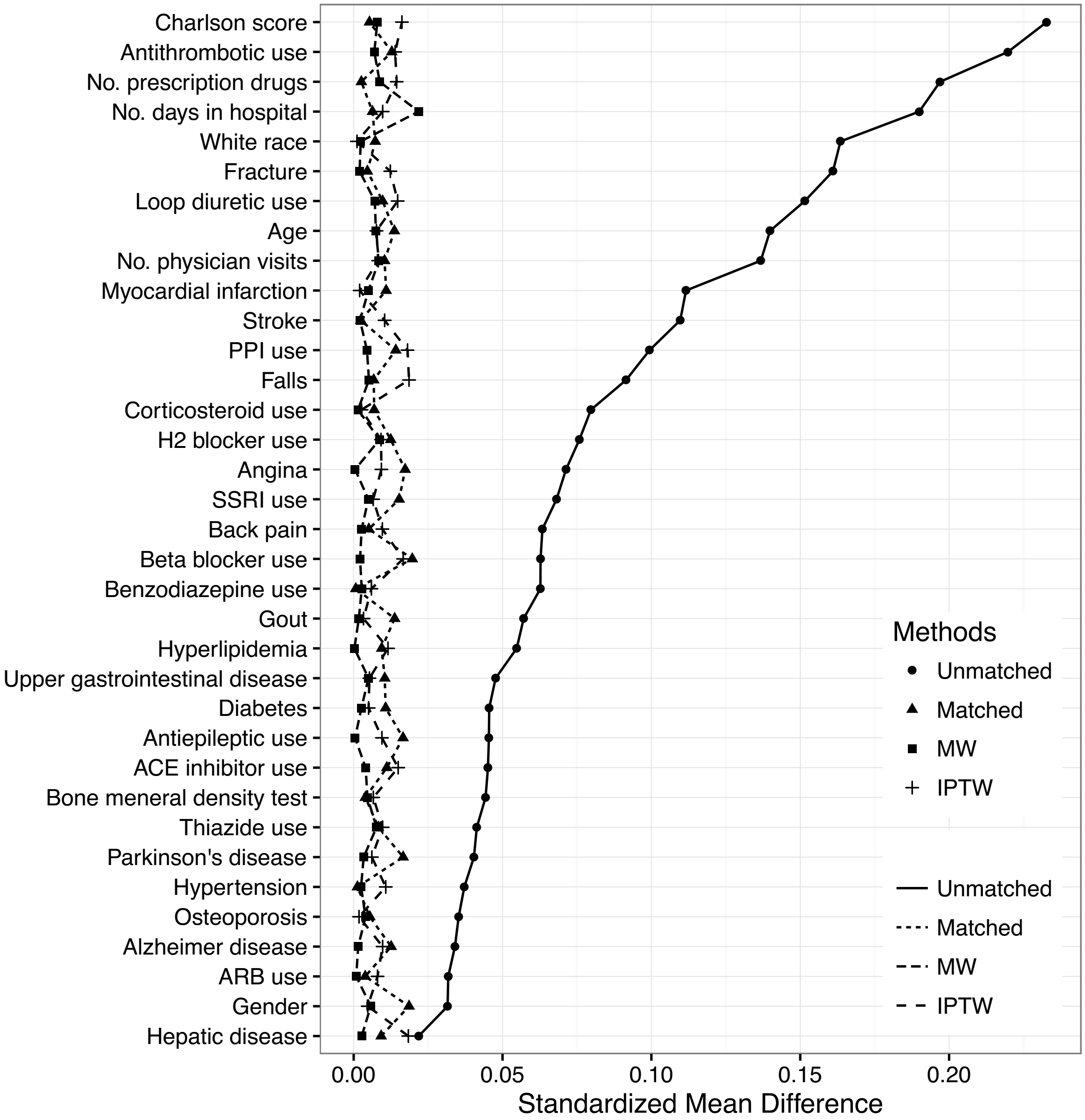
Abbreviations: U: Unmatched cohort, M: Matched cohort; Mw: matching weight cohort; Ip: Inverse probability of treatment weight cohort; pExpo: Exposure prevalence; pDis: baseline risk of disease

eFigure 1-9. Comparison of coverage probability of estimated confidence intervals between methods across 48 scenarios. The left half presents the constant treatment effect scenarios, whereas the right half presents treatment effect heterogeneity scenarios. Each three columns represent three treatment contrasts. Rows classify scenarios by good vs. poor covariate overlap levels and presence vs. absence of main effects. Each panel contains six lines classified by the exposure prevalence and the baseline risk. matching weights and matching performed similarly, whereas IPTW performed poorly in the poor covariate overlap settings. These confidence intervals were based on the estimated variance.



Abbreviations: U: Unmatched cohort, M: Matched cohort; Mw: matching weight cohort; Ip: Inverse probability of treatment weight cohort; pExpo: Exposure prevalence; pDis: baseline risk of disease.

eFigure 1-10. Standardized mean differences for each covariate averaged across three treatment contrasts in the unmatched, weighted, and matched cohort. Matching weights achieved the best covariate balance most consistently (24 of the 35 covariates) compared to three-way matching (6 covariates) and IPTW (5 covariates).



Abbreviations: PPI: proton pump inhibitor; H2: histamine-2 receptor; SSRI: selective serotonin reuptake inhibitor; ACE: angiotensin converting enzyme; ARB: angiotensin receptor blocker; MW: matching weights; IPTW: Inverse probability of treatment weights.

eTable 1-1. Characteristics of unmatched, matched, and weighted cohorts for the variables that were least balanced (average standardized mean difference > 0.1). The MW and matched cohorts were similar in characteristics, confirming the notion that MW is a weighting analogue to matching. As expected from the definition of the common support (overlap area of all three groups), these two cohorts are most similar to the smallest group, *i.e.*, the NSAIDs group in the unmatched cohort. The IPTW cohort had somewhat different characteristics with higher morbidity levels, most closely resembling the largest group, *i.e.*, the opioids group.

	nsNSAIDs	Coxibs	Opioids	SMD
Unmatched				
n	4874	6172	12601	
Charlson score, mean (SD)	1.59 (1.54)	1.72 (1.53)	2.17 (1.78)	0.233
Antithrombotic use, %	14.4	17.6	27.7	0.220
No. prescription drugs, mean (SD)	8.28 (4.69)	8.55 (4.76)	9.76 (5.38)	0.197
No. days in hospital, mean (SD)	1.85 (6.90)	2.19 (6.86)	4.18 (9.46)	0.190
White race, %	84.6	88	92.4	0.164
Fracture, %	6.5	7.2	13.7	0.161
Loop diuretic use, %	21.3	25.8	31.3	0.152
Age, mean (SD)	79.67 (7.03)	80.87 (6.99)	81.15 (7.17)	0.140
No. physician visits, mean (SD)	8.72 (6.32)	8.80 (5.99)	10.08 (7.14)	0.137
Myocardial infarction, %	5.2	5.7	9.6	0.112
Stroke, %	15.2	16.1	21.5	0.110
Matched				
n	4611	4611	4611	
Charlson score, mean (SD)	1.62 (1.54)	1.63 (1.52)	1.61 (1.52)	0.005
Antithrombotic use, %	15.1	15.5	15.8	0.013
No. prescription drugs, mean (SD)	8.34 (4.70)	8.33 (4.69)	8.32 (4.71)	0.003
No. days in hospital, mean (SD)	1.89 (6.45)	1.88 (6.54)	1.94 (6.29)	0.006
White race, %	86.9	86.7	86.6	0.007
Fracture, %	6.7	6.9	6.7	0.005
Loop diuretic use, %	22	22	22.6	0.010
Age, mean (SD)	79.97 (6.97)	79.96 (6.93)	80.11 (6.92)	0.014
No. physician visits, mean (SD)	8.76 (6.08)	8.76 (5.93)	8.66 (5.84)	0.010
Myocardial infarction, %	5.4	5.2	5.6	0.011
Stroke, %	15.5	15.6	15.7	0.002
Matching weights				
n	4633.49	4635.71	4618.71	
Charlson score, mean (SD)	1.62 (1.53)	1.61 (1.52)	1.63 (1.53)	0.008
Antithrombotic use, %	14.9	14.8	15.2	0.007
No. prescription drugs, mean (SD)	8.32 (4.70)	8.29 (4.67)	8.35 (4.71)	0.009
No. days in hospital, mean (SD)	1.87 (6.37)	1.78 (6.18)	2.00 (6.99)	0.022
White race, %	86.3	86.4	86.4	0.002

eTable 1-1 (Continued)

	nsNSAIDs	Coxibs	Opioids	SMD
Fracture, %	6.7	6.7	6.7	0.002
Loop diuretic use, %	22	21.8	22.3	0.007
Age, mean (SD)	79.97 (6.95)	79.95 (6.97)	80.02 (6.95)	0.007
No. physician visits, mean (SD)	8.72 (6.09)	8.69 (6.01)	8.76 (6.04)	0.008
Myocardial infarction, %	5.3	5.2	5.4	0.005
Stroke, %	15.4	15.4	15.5	0.002
IPTW				
n	4926.58	6187.8	12585.04	
Charlson score, mean (SD)	1.98 (1.70)	1.94 (1.68)	1.94 (1.69)	0.016
Antithrombotic use, %	23.3	22.5	22.4	0.014
No. prescription drugs, mean (SD)	9.27 (5.17)	9.15 (5.15)	9.17 (5.14)	0.014
No. days in hospital, mean (SD)	3.48 (8.96)	3.35 (8.78)	3.39 (9.82)	0.010
White race, %	89.7	89.7	89.7	0.001
Fracture, %	11.2	10.8	10.6	0.012
Loop diuretic use, %	28.9	27.9	27.9	0.015
Age, mean (SD)	80.89 (7.17)	80.82 (7.11)	80.81 (7.11)	0.008
No. physician visits, mean (SD)	9.58 (6.82)	9.49 (6.66)	9.50 (6.75)	0.008
Myocardial infarction, %	7.8	7.7	7.7	0.002
Stroke, %	19.4	18.8	18.9	0.010

Abbreviations: Matched: three-way matching; IPTW: inverse probability of treatment weights; Coxibs: COX-2 selective inhibitors; nsNSAIDs: non-selective nosteroidal anti-inflammatory drugs; SMD: standardized mean difference averaged across three pairwise contrasts

1 Proof for the two group setting

1.1 Estimand of Matching

This proof essentially follows the structure of the proof in the appendix of Li & Greene's 2013 paper[1]. The initial expression for the sample mean outcome in the matched treated group appears different from theirs, *i.e.*, $\frac{\sum_{k=1}^{K}\sum_{i=1}^{n} Y_i I(i \in \mathcal{S}_{1k})}{\sum_{k=1}^{K}\sum_{i=1}^{n} I(i \in \mathcal{S}_{1k})}$ where k is an index over discrete values of propensity scores, however both are the equivalent sample marginal mean outcome in the matched treated group. Instead of the explicit sum over k, we define a specific structure for the matched set.

The usual causal inference assumptions[2] are all required. The first is conditional exchangeability (unconfoundedness) given a function of the covariate vector $\mathbf{X}_i$ including the vector itself (finest balancing score) or the propensity score (coarsest balancing score). The latter requires no model misspecification for the propensity score model. The second is consistency, *i.e.*, $Y_i = Z_i Y_{1i} + (1 - Z_i) Y_{0i}$. That is, the observed outcome is the counterfactual potential outcome corresponding to the treatment received. This requires well-defined treatment and non-interference among individuals' potential outcomes. The third is positivity, *i.e.*, at any level of $\mathbf{X}_i$ (and thus propensity score), both treatment choices have non-zero (positive) probability. In this setting, this implies a perfect common support, *i.e.*, any propensity score values present in one of the treatment groups are also present in the other group.

Additional assumptions are required for the propensity score matching process. Matching has to be 1:1 matching without replacement. It also has to be exact matching on propensity scores (no calipers are allowed). This necessarily requires discrete propensity scores taking on a finite set of values because there has to be a positive probability of finding an exact match across two treatment groups[1]. The set of values can be arbitrarily large as long as its size is bounded and does not grow with the sample size n. When multiple untreated candidates are available for matching a treated individual at a given propensity score (< 0.5), one is selected at random. The same should apply when there are more treated individuals than untreated individuals at a given propensity score (> 0.5).

Proof: Let $l \in \{1, 2, ..., L\}$ be the index for the propensity score matched pairs. Let $\mathcal{S}_{1l}$ be the single member set of the treated subject from the l-th matched pair and the $\mathcal{S}_{0l}$ be the corresponding set of the untreated subject. Thus, $\mathcal{S}_1 = \bigcup_{l=1}^{L} \mathcal{S}_{1l}$ is the set of matched treated subjects, $\mathcal{S}_0 = \bigcup_{l=1}^{L} \mathcal{S}_{0l}$ is the set of matched untreated subjects, and $\mathcal{S} = \mathcal{S}_0 \cup \mathcal{S}_1$ is the set of the entire matched cohort. This matched cohort is balanced, *i.e.*, both groups contain the same number (L) of matched subjects. Index n is over the entire dataset before matching, thus, it includes subjects who do not match. The group mean in the matched treated group is expressed as follows. The selection indicator is effectively acting as a 0, 1 weight.

$$\frac{\sum_{i=1}^{n} Y_i I(i \in \mathcal{S}_1)}{\sum_{i=1}^{n} I(i \in \mathcal{S}_1)}$$

The numerator is examined first. The expression is multiplied by $\frac{1}{n}$, but it cancels out in the original expression as we do the same for the denominator. Y_i is the observed outcome of the i-th subject, whereas Y_{1i} is the treated counterfactual potential outcome of the i-th subject.

By consistency, the treated counterfactual is observed among the treated.

Only the treated contribute to the expression, thus, $Y_i = Y_{1i}$.

$$\frac{1}{n}\sum_{i=1}^{n} Y_i I(i \in \mathcal{S}_1) = \frac{1}{n}\sum_{i=1}^{n} Y_{1i} I(i \in \mathcal{S}_1)$$

Asymptotically, by the Weak Law of Large Number

$$\xrightarrow{p} E[Y_{1i} I(i \in \mathcal{S}_1)]$$

Rewrite as an iterative expectation.

$$= E[E[Y_{1i} I(i \in \mathcal{S}_1)|\mathbf{X}_i]]$$

Break the indicator into selection and treatment.

$$= E[E[Y_{1i} I(i \in \mathcal{S}) I(Z_i = 1)|\mathbf{X}_i]]$$

$\because$ only the treated subjects contribute to the inner expectation,

and otherwise it is zero, expectation can be taken

in the treated and weighted by its prevalence.

$$= E[E[Y_{1i} I(i \in \mathcal{S})|Z_i = 1, \mathbf{X}_i] P(Z_i = 1|\mathbf{X}_i)]$$

$\because$ given $Z_i = 1$ and within levels of $\mathbf{X}_i$, selection $(i \in \mathcal{S})$ is random,

Y_{1i} and selection indicator are conditionally independent.

$$= E[E[Y_{1i}|Z_i = 1, \mathbf{X}_i] E[I(i \in \mathcal{S})|Z_i = 1, \mathbf{X}_i] P(Z_i = 1|\mathbf{X}_i)]$$

By conditional exchangeability, $E[Y_{1i}|Z_i = 1, \mathbf{X}_i] = E[Y_{1i}|Z_i = 0, \mathbf{X}_i] = E[Y_{1i}|\mathbf{X}_i]$.

$$= E[E[Y_{1i}|\mathbf{X}_i] E[I(i \in \mathcal{S})|Z_i = 1, \mathbf{X}_i] P(Z_i = 1|\mathbf{X}_i)]$$

$\because$ expectation of a 0,1 selection indicator is the selection probability.

$$= E[E[Y_{1i}|\mathbf{X}_i] P(i \in \mathcal{S}|Z_i = 1, \mathbf{X}_i) P(Z_i = 1|\mathbf{X}_i)]$$

The last term is the propensity score by definition.

$$= E[E[Y_{1i}|\mathbf{X}_i] P\left(i \in \mathcal{S}|Z_i = 1, \mathbf{X}_i\right) e_i]$$

At a given $\mathbf{X}_i$, only the smaller group can match fully.

e_i is the fraction of the treated group at a given $\mathbf{X}_i$.

$\min(e_i, 1 - e_i)$ is the fraction of the smaller group at $\mathbf{X}_i$.

$\therefore$ among the treated group, only $\dfrac{\min(e_i, 1 - e_i)}{e_i}$ can match.

As this is a function of $\mathbf{X}_i$, conditioning is implicit.

$$= E\left[E[Y_{1i}|\mathbf{X}_i]\frac{\min(e_i, 1 - e_i)}{e_i}e_i\right]$$

$$= E\left[E[Y_{1i}|\mathbf{X}_i]\min(e_i, 1 - e_i)\right]$$

The denominator is a simplified version of the above proof.

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n} I(i \in \mathcal{S}_1) &= \frac{1}{n}\sum_{i=1}^{n} I(i \in \mathcal{S}_1) \\
&\xrightarrow{p} E[I(i \in \mathcal{S}_1)] \\
&= E[E[I(i \in \mathcal{S}_1)|\mathbf{X}_i]] \\
&= E[E[I(i \in \mathcal{S})I(Z_i = 1)|\mathbf{X}_i]] \\
&= E[E[I(i \in \mathcal{S})|Z_i = 1, \mathbf{X}_i]P\left(Z_i = 1|\mathbf{X}_i\right)]] \\
&= E[P(i \in \mathcal{S}|Z_i = 1, \mathbf{X}_i)P\left(Z_i = 1|\mathbf{X}_i\right)]] \\
&= E[P\left(i \in \mathcal{S}|Z_i = 1, \mathbf{X}_i\right)e_i] \\
&= E\left[\frac{\min(e_i, 1 - e_i)}{e_i}e_i\right] \\
&= E\left[\min(e_i, 1 - e_i)\right]
\end{aligned}$$

Therefore, the estimand of the group mean of the matched treated cohort is asymptotically the following.

$$\frac{E\left[E[Y_{1i}|\mathbf{X}_i]\min(e_i, 1 - e_i)\right]}{E\left[\min(e_i, 1 - e_i)\right]}$$

Similarly, the estimand of the group mean of the matched untreated cohort is asymptotically the following.

$$\frac{E\left[E[Y_{0i}|\mathbf{X}_i]\min(e_i, 1 - e_i)\right]}{E\left[\min(e_i, 1 - e_i)\right]}$$

Using these, the estimand of the group mean difference is

$$\begin{aligned}
\hat{\Delta}_M &= \frac{\sum_{i=1}^n Y_i I(i \in \mathcal{S}_1)}{\sum_{i=1}^n I(i \in \mathcal{S}_1)} - \frac{\sum_{i=1}^n Y_i I(i \in \mathcal{S}_0)}{\sum_{i=1}^n I(i \in \mathcal{S}_0)} \\
&= \frac{\sum_{i=1}^n Y_{1i} I(i \in \mathcal{S}_1)}{\sum_{i=1}^n I(i \in \mathcal{S}_1)} - \frac{\sum_{i=1}^n Y_{0i} I(i \in \mathcal{S}_0)}{\sum_{i=1}^n I(i \in \mathcal{S}_0)} \\
&\xrightarrow{p} \frac{E\left[E[Y_{1i}|\mathbf{X}_i]\min(e_i, 1-e_i)\right]}{E\left[\min(e_i, 1-e_i)\right]} - \frac{E\left[E[Y_{0i}|\mathbf{X}_i]\min(e_i, 1-e_i)\right]}{E\left[\min(e_i, 1-e_i)\right]} \\
&= \frac{E\left[E[Y_{1i}|\mathbf{X}_i]\min(e_i, 1-e_i) - E\left[E[Y_{0i}|\mathbf{X}_i]\min(e_i, 1-e_i)\right]\right]}{E\left[\min(e_i, 1-e_i)\right]} \\
&= \frac{E\left[(E[Y_{1i}|\mathbf{X}_i] - E[Y_{0i}|\mathbf{X}_i])\min(e_i, 1-e_i)\right]}{E\left[\min(e_i, 1-e_i)\right]} \\
&= \frac{E\left[E[Y_{1i} - Y_{0i}|\mathbf{X}_i]\min(e_i, 1-e_i)\right]}{E\left[\min(e_i, 1-e_i)\right]} \\
&= \frac{E\left[\Delta_i \min(e_i, 1-e_i)\right]}{E\left[\min(e_i, 1-e_i)\right]}
\end{aligned}$$

where Δ_i is the causal effect given covariates.

1.2 Estimand of Matching Weight

The corresponding matching weight estimator of the mean outcome in the treated is the following. The same causal inference assumptions are required except for the additional assumptions required for the matching algorithm.

$$\frac{\sum_{i=1}^n Y_i Z_i W_i}{\sum_{i=1}^n Z_i W_i}$$

where

$$W_i = \frac{\min(e_i, 1-e_i)}{Z_i e_i + (1-Z_i)(1-e_i)}$$

i.e., W_i is a function of covariates $\mathbf{X}_i$ and treatment Z_i.

The numerator has the following asymptotic characteristic.

By consistency, the treated counterfactual is observed among the treated.

Only the treated contribute to the expression, thus, $Y_i = Y_{1i}$.

$$\frac{1}{n}\sum_{i=1}^n Y_i Z_i W_i = \frac{1}{n}\sum_{i=1}^n Y_{1i} Z_i W_i$$

Asymptotically, by the Weak Law of Large Number

$$\xrightarrow{p} E[Y_{1i} Z_i W_i]$$

Rewrite as an iterative expectation.

$$= E[E[Y_{1i}Z_iW_i|\mathbf{X}_i]]$$

$\because (Y_{1i}, Y_{0i}) \perp\!\!\!\perp Z_i|\mathbf{X}_i$ implies $(Y_{1i}, Y_{0i}) \perp\!\!\!\perp f(\mathbf{X}_i, Z_i)|\mathbf{X}_i$,

the following holds $(Y_{1i}, Y_{0i}) \perp\!\!\!\perp Z_iW_i|\mathbf{X}_i$

$$= E[E[Y_{1i}|\mathbf{X}_i]E[Z_iW_i|\mathbf{X}_i]]$$

$\because$ only the treated units contribute to the second term,

and otherwise it is zero, expectation can be taken

in the treated and weighted by its prevalence.

$$= E[E[Y_{1i}|\mathbf{X}_i]E[W_i|Z_i = 1, \mathbf{X}_i]P(Z_i = 1|\mathbf{X}_i)]$$

The last term is the propensity score by definition.

Also expand the weight.

$$= E\left[E[Y_{1i}|\mathbf{X}_i]E\left[\frac{\min(e_i, 1-e_i)}{Z_ie_i + (1-Z_i)(1-e_i)}\middle|Z_i = 1, \mathbf{X}_i\right]e_i\right]$$

$\because Z_i = 1$ for the second term

$$= E\left[E[Y_{1i}|\mathbf{X}_i]\frac{\min(e_i, 1-e_i)}{e_i}e_i\right]$$

$$= E\left[E[Y_{1i}|\mathbf{X}_i]\min(e_i, 1-e_i)\right]$$

Similarly, the denominator has the following asymptotic characteristic.

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n} Z_iW_i &\xrightarrow{p} E[Z_iW_i] \\
&= E[E[Z_iW_i|\mathbf{X}_i]] \\
&= E[E[W_i|Z_i = 1, \mathbf{X}_i]P(Z_i = 1|\mathbf{X}_i)] \\
&= E\left[E\left[\frac{\min(e_i, 1-e_i)}{Z_ie_i + (1-Z_i)(1-e_i)}\middle|Z_i = 1, \mathbf{X}_i\right]e_i\right] \\
&= E\left[\frac{\min(e_i, 1-e_i)}{e_i}e_i\right] \\
&= E\left[\min(e_i, 1-e_i)\right]
\end{aligned}$$

Therefore, the estimand of matching weight estimator for the treated group mean has the same form as the corresponding matching estimator asymptotically.

$$
\begin{aligned}
\frac{\sum_{i=1}^{n} Y_i Z_i W_i}{\sum_{i=1}^{n} Z_i W_i} &= \frac{\sum_{i=1}^{n} Y_{1i} Z_i W_i}{\sum_{i=1}^{n} Z_i W_i} \\
&\xrightarrow{p} \frac{E\left[E[Y_{1i}|\mathbf{X}_i]\min(e_i, 1-e_i)\right]}{E\left[\min(e_i, 1-e_i)\right]}
\end{aligned}
$$

Because this holds similarly for the untreated group, the estimand of the treatment effect is also asymptotically equivalent.

2 Extension to 3+ group settings

In the previous proof following Li & Greene 2013, the effect estimate was compared between the matching method and the matching weight method. Proving the asymptotic equivalence of the estimand of an arbitrary group-specific mean outcome in 3+ group setting will generalize the proof. The same assumptions are required on all the treatment groups under study.

2.1 Estimand of Matching in 3+ group setting

One propensity score is defined for each treatment group. For the k-th treatment group, e_{ki} is the corresponding treatment-specific propensity score, *i.e.*, the probability of being assigned to the k-th treatment group for the i-th subject given covariates. The treatment-specific propensity scores must be formed in such a way that within an individual subject $\sum\limits_{k=1}^{K} e_{ki} = 1$ is met. This requires a single model be fit for estimation (*e.g.*, multinomial logistic regression).

The same assumptions as the two group setting are required. Regarding the matching process now it is a simultaneous $1 : 1 : ... : 1$ exact matching of K treatment groups on their K treatment-specific propensity scores without replacement. That is, K individuals with the identical propensity scores (all of the treatment-specific propensity scores, $e_{1i}, \dots, e_{Ki}$ must match up across K individuals) form a matched unit. If there are multiple candidates from a given treatment group k, the selection is random.

Proof: Let $\mathcal{S}_{kl}$ be the single member set of the subject in the k-th treatment group ($k \in \{1, 2, ..., K\}$) from the l-th propensity score matched unit ($l \in \{1, 2, ..., L\}$). Thus, $\mathcal{S}_k = \bigcup\limits_{l=1}^{L} \mathcal{S}_{kl}$ is the set of all matched subjects in the k-th treatment group, and $\mathcal{S} = \bigcup\limits_{k=1}^{K} \mathcal{S}_k$ is the set of entire matched cohort. This matched cohort is balanced, *i.e.*, each one of K treatment groups contain the same number (L) of matched subjects. Index n is still over all individuals in the dataset before matching. The treatment variable, Z_i is now a nominal variable $1, 2, ..., K$ indicating the treatment group. The group mean in the k-th group is expressed as follows.

$$\frac{\sum_{i=1}^{n} Y_i I(i \in \mathcal{S}_k)}{\sum_{i=1}^{n} I(i \in \mathcal{S}_k)}$$

The numerator is examined first. The expression is multiplied by $\frac{1}{n}$, but it cancels in the original expression as we do the same for the denominator. For the most part the proof is almost identical to the previous one.

By consistency, the k-th counterfactual is observed in the k-th group

Also only the k-th group contributes to the expression, thus, $Y_i = Y_{ki}$

$$\frac{1}{n}\sum_{i=1}^{n} Y_i I(i \in \mathcal{S}_k) = \frac{1}{n}\sum_{i=1}^{n} Y_{ki} I(i \in \mathcal{S}_k)$$

Asymptotically, by the Weak Law of Large Number

$\xrightarrow{p} E[Y_{ki} I(i \in \mathcal{S}_k)]$

Rewrite as an iterative expectation.

$= E[E[Y_{ki} I(i \in \mathcal{S}_k)|\mathbf{X}_i]]$

Break the indicator into selection and treatment.

$= E[E[Y_{ki} I(i \in \mathcal{S}) I(Z_i = k)|\mathbf{X}_i]]$

$\because$ only the k-th group contributes to the inner expectation,

and otherwise it is zero, expectation can be taken

in the k-th group and weighted by its prevalence.

$= E[E[Y_{ki} I(i \in \mathcal{S})|Z_i = k, \mathbf{X}_i] P(Z_i = k|\mathbf{X}_i)]$

$\because$ given $Z_i = k$ and within levels of $\mathbf{X}_i$, selection $(i \in \mathcal{S})$ is random,

Y_{ki} and selection indicator are conditionally independent.

$= E[E[Y_{ki}|Z_i = k, \mathbf{X}_i] E[I(i \in \mathcal{S})|Z_i = k, \mathbf{X}_i] P(Z_i = k|\mathbf{X}_i)]$

By conditional exchangeability, $E[Y_{ki}|Z_i = k, \mathbf{X}_i] = E[Y_{ki}|\mathbf{X}_i]$.

$= E[E[Y_{ki}|\mathbf{X}_i] E[I(i \in \mathcal{S})|Z_i = k, \mathbf{X}_i] P(Z_i = k|\mathbf{X}_i)]$

$\because$ expectation of a 0,1 selection indicator is the selection probability.

$= E[E[Y_{ki}|\mathbf{X}_i] P(i \in \mathcal{S}|Z_i = k, \mathbf{X}_i) P(Z_i = k|\mathbf{X}_i)]$

The last term is the PS for the k-th treatment by definition.

$= E[E[Y_{ki}|\mathbf{X}_i] P\left(i \in \mathcal{S}|Z_i = k, \mathbf{X}_i\right) e_{ki}]$

At a given $\mathbf{X}_i$, only the smallest group can match fully.

e_{ki} is the fraction of k-th group at a given $\mathbf{X}_i$.

$\min(e_{1i}, e_{2i}, ..., e_{Ki})$ is the fraction of the smallest group at $\mathbf{X}_i$.

$\therefore$ Among the k-th group, only $\dfrac{\min(e_{1i}, e_{2i}, ..., e_{Ki})}{e_{ki}}$ can match.

As this is a function of $\mathbf{X}_i$, conditioning is implicit.

$= E\left[E[Y_{ki}|\mathbf{X}_i] \dfrac{\min(e_{1i}, e_{2i}, ..., e_{Ki})}{e_{ki}} e_{ki}\right]$

$= E\left[E[Y_{ki}|\mathbf{X}_i] \min(e_{1i}, e_{2i}, ..., e_{Ki})\right]$

Similarly,

$$\frac{1}{n}\sum_{i=1}^{n} I(i \in \mathcal{S}_k) = \frac{1}{n}\sum_{i=1}^{n} I(i \in \mathcal{S}_k)$$

$$\xrightarrow{p} E\left[\min(e_{1i}, e_{2i}, ..., e_{Ki})\right]$$

Therefore, the estimand of the group mean of the matched k-th group is asymptotically the following.

$$\frac{E\left[E[Y_{ki}|\mathbf{X}_i]\min(e_{1i}, e_{2i}, ..., e_{Ki})\right]}{E\left[\min(e_{1i}, e_{2i}, ..., e_{Ki})\right]}$$

2.2 Estimand of Matching Weight in 3+ group setting

The corresponding weighted estimator of the mean outcome in the treated is the following. The denominator of the weight picks the propensity score for the assigned treatment for the i-th unit.

$$\frac{\sum_{i=1}^{n} Y_i I(Z_i = k)W_i}{\sum_{i=1}^{n} I(Z_i = k)W_i}$$

where

$$W_i = \frac{\min(e_{1i}, e_{2i}, ..., e_{Ki})}{\sum_{k=1}^{K} I(Z_i = k)e_{ki}}$$

The numerator has the following asymptotic characteristic.

By consistency, the k-th counterfactual is observed in the k-th group

Also only the k-th group contributes to the expression, thus, $Y_i = Y_{ki}$

$$\frac{1}{n}\sum_{i=1}^{n} Y_i I(Z_i = k)W_i = \frac{1}{n}\sum_{i=1}^{n} Y_{ki} I(Z_i = k)W_i$$

Asymptotically, by the Weak Law of Large Number

$$\xrightarrow{p} E[Y_{ki} I(Z_i = k)W_i]$$

Rewrite as an iterative expectation.

$$= E[E[Y_{ki} I(Z_i = k)W_i|\mathbf{X}_i]]$$

$\because Y_{ki} \perp\!\!\!\perp Z_i|\mathbf{X}_i$ implies $Y_{ki} \perp\!\!\!\perp f(\mathbf{X}_i, Z_i)|\mathbf{X}_i$,

the following holds $Y_{ki} \perp\!\!\!\perp I(Z_i = k)W_i|\mathbf{X}_i$

$$= E[E[Y_{ki}|\mathbf{X}_i]E[I(Z_i = k)W_i|\mathbf{X}_i]]$$

$\because$ only the k-th group contributes to the second term,

and otherwise it is zero, expectation can be taken in the k-th group and weighted by its prevalence.

$$= E[E[Y_{ki}|\mathbf{X}_i]E[W_i|Z_i = k, \mathbf{X}_i]P(Z_i = k|\mathbf{X}_i)]$$

The last term is the propensity score for the k-th treatment.

Also expand the weight.

$$= E\left[E[Y_{ki}|\mathbf{X}_i]E\left[\frac{\min(e_{1i}, e_{2i}, ..., e_{Ki})}{\sum_{k=1}^{K} I(Z_i = k)e_{ki}}\middle|Z_i = k, \mathbf{X}_i\right] e_{ki}\right]$$

$\because Z_i = k$ for the second term

$$= E\left[E[Y_{ki}|\mathbf{X}_i]\frac{\min(e_{1i}, e_{2i}, ..., e_{Ki})}{e_{ki}}e_{ki}\right]$$

$$= E\left[E[Y_{ki}|\mathbf{X}_i]\min(e_{1i}, e_{2i}, ..., e_{Ki})\right]$$

Similarly,

$$\frac{1}{n}\sum_{i=1}^{n} I(Z_i = k)W_i \xrightarrow{p} E[I(Z_i = k)W_i]$$

$$= E\left[\min(e_{1i}, e_{2i}, ..., e_{Ki})\right]$$

Therefore, the estimand of matching weight estimator has the same form as the matching estimator asymptotically.

$$\frac{\sum_{i=1}^{n} Y_i I(Z_i = k)W_i}{\sum_{i=1}^{n} I(Z_i = k)W_i} = \frac{\sum_{i=1}^{n} Y_{ki} I(Z_i = k)W_i}{\sum_{i=1}^{n} I(Z_i = k)W_i}$$

$$\xrightarrow{p} \frac{E\left[E[Y_{ki}|\mathbf{X}_i]\min(e_{1i}, e_{2i}, ..., e_{Ki})\right]}{E\left[\min(e_{1i}, e_{2i}, ..., e_{Ki})\right]}$$

Because this holds true for each treatment group, the estimand of any two group contrast effect is also asymptotically equivalent between the multi-way matching method and the matching weight method.

References

[1] L. Li and T. Greene, “A weighting analogue to pair matching in propensity score analysis,” *The International Journal of Biostatistics*, vol. 9, no. 2, pp. 215–234, 2013.

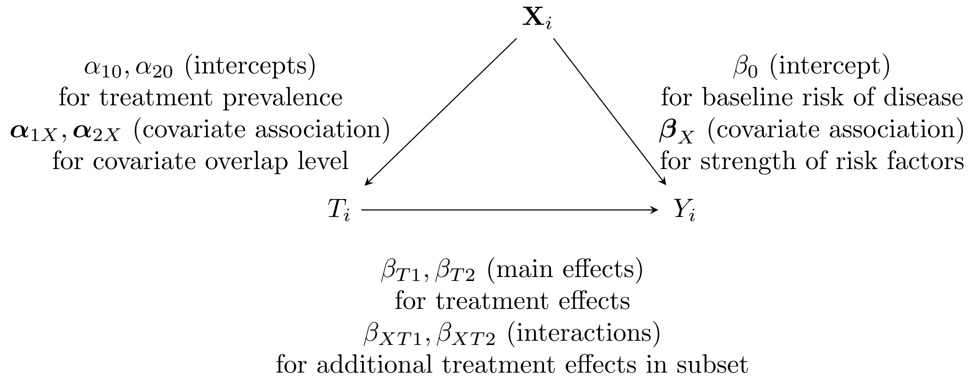
[2] M. A. Hernan and J. M. Robins, *Causal Inference.* Chapman & Hall/CRC, 2016.

1 Data Generation Mechanism DAG

The covariate were generated following the data generation process of Franklin *et al*[1]. The treatment assignment process also followed that of Franklin *et al*[1], but was extended to the three treatment group setting using a multinomial logistic model[2, 3]. The outcome model was a log-probability model to avoid non-collapsibility issues[4, 5].

1.1 Annotated Directed Acyclic Graph

$\mathbf{X}_i$ is a vector of ten covariates for the i-th individual, $T_i \in \{0, 1, 2\}$ is the treatment level, and $Y_i \in \{0, 1\}$ is the binary outcome.


$\mathbf{X}_i$
α_{10}, α_{20} (intercepts)
for treatment prevalence
$\boldsymbol{\alpha}_{1X}, \boldsymbol{\alpha}_{2X}$ (covariate association)
for covariate overlap level
β_0 (intercept)
for baseline risk of disease
$\boldsymbol{\beta}_X$ (covariate association)
for strength of risk factors
T_i
Y_i
β_{T1}, β_{T2} (main effects)
for treatment effects
β_{XT1}, β_{XT2} (interactions)
for additional treatment effects in subset

1.2 Covariate Generation

The covariate vector for the i-th individual, $\mathbf{X}_i$ had the following random elements[1].

Variable	Generation Process
X_{1i}	Normal$(0, 1^2)$
X_{2i}	Log-Normal$(0, 0.5^2)$
X_{3i}	Normal$(0, 10^2)$
X_{4i}	Bernoulli$(p_i = e^{2X_{1i}}/(1 + e^{2X_{1i}}))$ where $E[p_i] = 0.5$
X_{5i}	Bernoulli$(p = 0.2)$
X_{6i}	Multinomial$(\mathbf{p} = (0.5, 0.3, 0.1, 0.05, 0.05)^T)$
X_{7i}	$\sin(X_{1i})$
X_{8i}	X_{2i}^2
X_{9i}	$X_{3i} \times X_{4i}$
X_{10i}	$X_{4i} \times X_{5i}$

1.3 Treatment Generating Model

As there were three treatment groups, two relative probabilities were jointly modeled by two simultaneous models (essentially multinomial logistic model).

$$\eta_{T1i} = \log\left(\frac{P(T_i = 1|\mathbf{X}_i = \mathbf{x}_i)}{P(T_i = 0|\mathbf{X}_i = \mathbf{x}_i)}\right) = \alpha_{10} + \boldsymbol{\alpha}_{1X}^T \mathbf{x}_i$$

$$\eta_{T2i} = \log\left(\frac{P(T_i = 2|\mathbf{X}_i = \mathbf{x}_i)}{P(T_i = 0|\mathbf{X}_i = \mathbf{x}_i)}\right) = \alpha_{20} + \boldsymbol{\alpha}_{2X}^T \mathbf{x}_i$$

where

α_{10}, α_{20} determine treatment prevalence

$\boldsymbol{\alpha}_{1X}, \boldsymbol{\alpha}_{2X}$ determine covariate-treatment association

Importantly, the covariate-treatment association is inversely correlated with the covariate overlap in these model. This is because if patient characteristics play more important roles in treatment decision, the treatment assignment is less random.

To obtain the three predicted probabilities (true propensity scores) from the two linear predictors, we conducted the following normalization process[2, 3].

$$\begin{aligned}
e_{0i} &= P(T_i = 0|\mathbf{X}_i = \mathbf{x}_i) = \frac{1}{q_i} \\
e_{1i} &= P(T_i = 1|\mathbf{X}_i = \mathbf{x}_i) = \frac{\exp(\eta_{T1i})}{q_i} \\
e_{2i} &= P(T_i = 2|\mathbf{X}_i = \mathbf{x}_i) = \frac{\exp(\eta_{T2i})}{q_i} \\
&\text{where } q_i = 1 + \exp(\eta_{T1i}) + \exp(\eta_{T2i})
\end{aligned}$$

Finally, the treatment level was assigned in a multinomial random number generating process.

$$T_i \sim \text{Multinomial}\left(n = 1, \mathbf{p} = (e_{0i}, e_{1i}, e_{2i})^T\right)$$

1.4 Outcome Generating Model

The log probability of disease was generated using a log-linear (log-probability) model to avoid the non-collapsibility issue of the logistic model.

$$\begin{aligned}
\eta_{Yi} = \log(P(Y_i = 1|T_i = t_i, \mathbf{X}_i = \mathbf{x}_i)) &= \beta_0 + \boldsymbol{\beta}_X^T \mathbf{x}_i + \beta_{T1} I(t_i = 1) + \beta_{T2} I(t_i = 2) \\
&+ \beta_{XT1} x_{4i} I(t_i = 1) + \beta_{XT2} x_{4i} I(t_i = 2) \\
\text{where}& \\
t_i =&\ \text{Assigned treatment} \\
\beta_0 =&\ \text{Intercept determining baseline disease risk} \\
\boldsymbol{\beta}_X =&\ \text{Effects of ten covariates (risk factors) on disease risk} \\
\beta_{T1} =&\ \text{Main effect of Treatment 1 compared to Treatment 0} \\
\beta_{T2} =&\ \text{Main effect of Treatment 2 compared to Treatment 0} \\
\beta_{XT1} =&\ \text{Additional effect for Treatment 1 vs 0 among } X_{4i} = 1 \\
\beta_{XT2} =&\ \text{Additional effect for Treatment 2 vs 0 among } X_{4i} = 1
\end{aligned}$$

Using this linear predictor, the probability of disease was calculated as follows.

$$p_{Yi} = P(Y_i = 1|T_i = t_i, \mathbf{X}_i = \mathbf{x}_i) = \exp(\eta_{Yi})$$

Then the outcome was assigned using a Bernoulli random number generating process.

$$Y_i \sim \text{Bernoulli}\,(p_{Yi})$$

The counterfactual probability of disease under each treatment was defined as follows.

$$\begin{aligned}
p_{Yi}(0) &= P(Y_i = 1 | T_i = 0, \mathbf{X}_i = \mathbf{x}_i) \\
p_{Yi}(1) &= P(Y_i = 1 | T_i = 1, \mathbf{X}_i = \mathbf{x}_i) \\
p_{Yi}(2) &= P(Y_i = 1 | T_i = 2, \mathbf{X}_i = \mathbf{x}_i)
\end{aligned}$$

1.5 Parameter Settings

The parameters were assinged as follows.

1.5.1 Treatment Generating Model

All possible combinations of three treatment prevalences and two levels of covariate overlap (inverse of covariate-treatment association) were generated as follows (6 combinations).

	Treatment Prevalence											
	33:33:33				10:45:45				10:10:80			
	Covariate Overlap											
	Good		Poor		Good		Poor		Good		Poor	
	$\boldsymbol{\alpha}_1$	$\boldsymbol{\alpha}_2$	$\boldsymbol{\alpha}_1$	$\boldsymbol{\alpha}_2$	$\boldsymbol{\alpha}_1$	$\boldsymbol{\alpha}_2$	$\boldsymbol{\alpha}_1$	$\boldsymbol{\alpha}_2$	$\boldsymbol{\alpha}_1$	$\boldsymbol{\alpha}_2$	$\boldsymbol{\alpha}_1$	$\boldsymbol{\alpha}_2$
Intercept	-0.13	-0.26	-0.75	-3.75	1.30	1.18	1.55	-0.65	-0.10	1.87	0.60	1.70
X_1	0.05	0.10	0.80	1.60	0.05	0.10	0.80	1.60	0.05	0.10	0.80	1.60
X_2	0.00	0.01	0.06	0.12	0.00	0.01	0.06	0.12	0.00	0.01	0.06	0.12
X_3	0.00	0.01	0.06	0.12	0.00	0.01	0.06	0.12	0.00	0.01	0.06	0.12
X_4	0.09	0.19	1.50	3.00	0.09	0.19	1.50	3.00	0.09	0.19	1.50	3.00
X_5	0.09	0.19	1.50	3.00	0.09	0.19	1.50	3.00	0.09	0.19	1.50	3.00
X_6	0.03	0.05	0.40	0.80	0.03	0.05	0.40	0.80	0.03	0.05	0.40	0.80
X_7	0.05	0.10	0.80	1.60	0.05	0.10	0.80	1.60	0.05	0.10	0.80	1.60
X_8	0.00	0.01	0.04	0.08	0.00	0.01	0.04	0.08	0.00	0.01	0.04	0.08
X_9	0.01	0.01	0.08	0.16	0.01	0.01	0.08	0.16	0.01	0.01	0.08	0.16
X_{10}	0.06	0.12	1.00	2.00	0.06	0.12	1.00	2.00	0.06	0.12	1.00	2.00

where $\boldsymbol{\alpha}_1 = (\alpha_{10}, \boldsymbol{\alpha}_{1X}^T)^T$ and $\boldsymbol{\alpha}_2 = (\alpha_{20}, \boldsymbol{\alpha}_{2X}^T)^T$.

1.5.2 Outcome Generating Model

The outcome generating model parameters were the following.

Two types of baseline risks

$\beta_0 \in \{\log(0.05), \log(0.20)\}$, *i.e.*, 5% and 20% baseline risk

One type of covariate-outcome association

$$\boldsymbol{\beta}_X = (0.160, 0.012, 0.012, 0.300, 0.300, 0.080, 0.160, 0.008, 0.016, 0.200)^T$$

Null or non-null treatment (main) effects

$$\boldsymbol{\beta}_T = (\beta_{T1}, \beta_{T2})^T \in \left\{(0,0)^T, (\log(0.9), \log(0.6))^T\right\}$$

For the non-null case:

relative risk of 0.9 comparing Treatment 1 vs 0

relative risk of 0.6 comparing Treatment 2 vs 0

$\Longrightarrow$ relative risk of 6/9 comparing Treatment 2 vs 1

$$\begin{aligned}
&\text{Null or non-null treatment effect modification}\\
\boldsymbol{\beta}_{XT} &= (\beta_{XT1}, \beta_{XT2})^T \in \left\{(0,0)^T, (\log(0.7), \log(0.5))^T\right\}\\
&\text{For the non-null case:}\\
&\text{additional } 0.7\times \text{ risk reduction among } X_{5i} = 1 \text{ for Treatment 1 vs 0}\\
&\text{additional } 0.5\times \text{ risk reduction among } X_{5i} = 1 \text{ for Treatment 2 vs 0}\\
\implies &\text{additional } 5/7\times \text{ risk reduction among } X_{5i} = 1 \text{ for Treatment 2 vs 1}
\end{aligned}$$

There are thus, $2 \times 1 \times 2 \times 2 = 8$ combinations of the outcome generating model parameters

1.6 Simulation scenarios

There are $6 \times 8 = 48$ total simulation scenarios. The columns in the table denote the following settings.

- Scenario: Scenario number
- N: Total sample size
- EM: Treatment effect modification by covariate
- Main effects: Treatment main effects
- Risk: Baseline risk in Treatment 0
- Group sizes: Relative sizes of three treatment groups
- Covariate overlap: Level of covariate overlap

Scenario	N	EM	Main effects	Risk	Group sizes	Covariate overlap
1	6000	Modification (-)	Null main effects	0.05	33:33:33	Good overlap
2	6000	Modification (-)	Null main effects	0.05	33:33:33	Poor overlap
3	6000	Modification (-)	Null main effects	0.05	10:45:45	Good overlap
4	6000	Modification (-)	Null main effects	0.05	10:45:45	Poor overlap
5	6000	Modification (-)	Null main effects	0.05	10:10:80	Good overlap
6	6000	Modification (-)	Null main effects	0.05	10:10:80	Poor overlap
7	6000	Modification (-)	Null main effects	0.2	33:33:33	Good overlap
8	6000	Modification (-)	Null main effects	0.2	33:33:33	Poor overlap
9	6000	Modification (-)	Null main effects	0.2	10:45:45	Good overlap
10	6000	Modification (-)	Null main effects	0.2	10:45:45	Poor overlap
11	6000	Modification (-)	Null main effects	0.2	10:10:80	Good overlap
12	6000	Modification (-)	Null main effects	0.2	10:10:80	Poor overlap
13	6000	Modification (-)	Non-null main effects	0.05	33:33:33	Good overlap
14	6000	Modification (-)	Non-null main effects	0.05	33:33:33	Poor overlap
15	6000	Modification (-)	Non-null main effects	0.05	10:45:45	Good overlap
16	6000	Modification (-)	Non-null main effects	0.05	10:45:45	Poor overlap
17	6000	Modification (-)	Non-null main effects	0.05	10:10:80	Good overlap
18	6000	Modification (-)	Non-null main effects	0.05	10:10:80	Poor overlap
19	6000	Modification (-)	Non-null main effects	0.2	33:33:33	Good overlap
20	6000	Modification (-)	Non-null main effects	0.2	33:33:33	Poor overlap
21	6000	Modification (-)	Non-null main effects	0.2	10:45:45	Good overlap
22	6000	Modification (-)	Non-null main effects	0.2	10:45:45	Poor overlap
23	6000	Modification (-)	Non-null main effects	0.2	10:10:80	Good overlap
24	6000	Modification (-)	Non-null main effects	0.2	10:10:80	Poor overlap
25	6000	Modification (+)	Null main effects	0.05	33:33:33	Good overlap
26	6000	Modification (+)	Null main effects	0.05	33:33:33	Poor overlap
27	6000	Modification (+)	Null main effects	0.05	10:45:45	Good overlap
28	6000	Modification (+)	Null main effects	0.05	10:45:45	Poor overlap
29	6000	Modification (+)	Null main effects	0.05	10:10:80	Good overlap
30	6000	Modification (+)	Null main effects	0.05	10:10:80	Poor overlap
31	6000	Modification (+)	Null main effects	0.2	33:33:33	Good overlap
32	6000	Modification (+)	Null main effects	0.2	33:33:33	Poor overlap
33	6000	Modification (+)	Null main effects	0.2	10:45:45	Good overlap
34	6000	Modification (+)	Null main effects	0.2	10:45:45	Poor overlap
35	6000	Modification (+)	Null main effects	0.2	10:10:80	Good overlap
36	6000	Modification (+)	Null main effects	0.2	10:10:80	Poor overlap
37	6000	Modification (+)	Non-null main effects	0.05	33:33:33	Good overlap
38	6000	Modification (+)	Non-null main effects	0.05	33:33:33	Poor overlap
39	6000	Modification (+)	Non-null main effects	0.05	10:45:45	Good overlap
40	6000	Modification (+)	Non-null main effects	0.05	10:45:45	Poor overlap
41	6000	Modification (+)	Non-null main effects	0.05	10:10:80	Good overlap
42	6000	Modification (+)	Non-null main effects	0.05	10:10:80	Poor overlap
43	6000	Modification (+)	Non-null main effects	0.2	33:33:33	Good overlap
44	6000	Modification (+)	Non-null main effects	0.2	33:33:33	Poor overlap
45	6000	Modification (+)	Non-null main effects	0.2	10:45:45	Good overlap
46	6000	Modification (+)	Non-null main effects	0.2	10:45:45	Poor overlap
47	6000	Modification (+)	Non-null main effects	0.2	10:10:80	Good overlap
48	6000	Modification (+)	Non-null main effects	0.2	10:10:80	Poor overlap

References

[1] J. M. Franklin, J. A. Rassen, D. Ackermann, D. B. Bartels, and S. Schneeweiss, "Metrics for covariate balance in cohort studies of causal effects," *Statistics in Medicine*, vol. 33, pp. 1685–1699, May 2014.

[2] A. Linden, S. D. Uysal, A. Ryan, and J. L. Adams, "Estimating causal effects for multivalued treatments: a comparison of approaches," *Statistics in Medicine*, Oct. 2015.

[3] M. D. Cattaneo, D. M. Drukker, and A. D. Holland, "Estimation of multivalued treatment effects under conditional independence," vol. 13, no. 3, pp. 407–450, 2013.

[4] P. Cummings, "The relative merits of risk ratios and odds ratios," *Archives of Pediatrics & Adolescent Medicine*, vol. 163, pp. 438–445, May 2009.

[5] S. Greenland, J. M. Robins, and J. Pearl, "Confounding and Collapsibility in Causal Inference," *Statistical Science*, vol. 14, pp. 29–46, Feb. 1999.

1 Aim

This document provides a step-by-step guide for implementation of matching weight method in practice. The example is in the three-group setting. However, the essentially the same code can be used in the two-group setting or settings where there are more than three groups. The example is written in R, but it can be implemented in any statistical environment that has (multinomial) logistic regression and weighted data analysis capabilities.

2 Dataset

The tutoring dataset included in the TriMatch R package is used. The exposure is the treat variable, which takes one of Treat1, Treat2, and Control. These represent the tutoring method each student received. The outcome is the Grade ordinal variable, which takes one of 0, 1, 2, 3, or 4. Pre-treatment potential confounders include gender, ethnicity, military service status of the student, non-native English speaker status, education level of the subject's mother (ordinal), education level of the subject's father (ordinal), age of the student, employment status (no, part-time, full-time), household income (ordinal), number of transfer credits, grade point average. The dataset does not contain any missing values. See ?tutoring for details. The employment categorical variable is coded numerically. Thus, it is converted to a factor.

```
## Load data
library(TriMatch)
data(tutoring)
summary(tutoring)

##      treat        Course              Grade         Gender    Ethnicity    Military
##  Control:918   Length:1142        Min.   :0.000   FEMALE:627   Black:211   Mode :logical
##  Treat1 :134   Class :character   1st Qu.:2.000   MALE  :515   Other:193   FALSE:783
##  Treat2 : 90   Mode  :character   Median :4.000                White:738   TRUE :359
##                                   Mean   :2.891
##                                   3rd Qu.:4.000
##                                   Max.   :4.000
##     ESL             EdMother        EdFather          Age          Employment        Income
##  Mode :logical   Min.   :1.000   Min.   :1.000   Min.   :20.00   Min.   :1.000   Min.   :1.000
##  FALSE:1049      1st Qu.:3.000   1st Qu.:3.000   1st Qu.:30.00   1st Qu.:3.000   1st Qu.:3.000
##  TRUE :93        Median :3.000   Median :3.000   Median :37.00   Median :3.000   Median :5.000
##                  Mean   :3.785   Mean   :3.684   Mean   :36.92   Mean   :2.667   Mean   :5.059
##                  3rd Qu.:5.000   3rd Qu.:5.000   3rd Qu.:43.00   3rd Qu.:3.000   3rd Qu.:7.000
##                  Max.   :8.000   Max.   :9.000   Max.   :65.00   Max.   :3.000   Max.   :9.000
##     Transfer           GPA         GradeCode           Level           ID
##  Min.   :  3.00   Min.   :0.000   Length:1142        Lower:988   Min.   :   1.0
##  1st Qu.: 36.66   1st Qu.:2.890   Class :character   Upper:154   1st Qu.: 286.2
##  Median : 48.31   Median :3.215   Mode  :character               Median : 571.5
##  Mean   : 52.12   Mean   :3.166                                  Mean   : 571.5
##  3rd Qu.: 65.00   3rd Qu.:3.518                                  3rd Qu.: 856.8
##  Max.   :126.00   Max.   :4.000                                  Max.   :1142.0

## Make employment categorical
tutoring$Employment <- factor(tutoring$Employment, levels = 1:3,
                              labels = c("no","part-time","full-time"))
```

3 Pre-weighting balance assessment

The tableone package can be utilized for covariate balance assessment using standardized mean differences (SMD). SMD greater than 0.1 is often regarded as a substantial imbalance. The SMD shown in the table is the average of all possible pairwise SMDs.

```
## Examine covariate balance
library(tableone)
covariates <- c("Gender", "Ethnicity", "Military", "ESL",
                "EdMother", "EdFather", "Age", "Employment",
                "Income", "Transfer", "GPA")
tab1Unadj <- CreateTableOne(vars = covariates, strata = "treat", data = tutoring)
print(tab1Unadj, test = FALSE, smd = TRUE)
```

```
##                       Stratified by treat
##                        Control       Treat1        Treat2        SMD
##   n                      918           134            90
##   Gender = MALE (%)      449 (48.9)     38 (28.4)     28 (31.1)   0.287
##   Ethnicity (%)                                                   0.095
##      Black               166 (18.1)     24 (17.9)     21 (23.3)
##      Other               157 (17.1)     23 (17.2)     13 (14.4)
##      White               595 (64.8)     87 (64.9)     56 (62.2)
##   Military = TRUE (%)    309 (33.7)     32 (23.9)     18 (20.0)   0.208
##   ESL = TRUE (%)          76 ( 8.3)      8 ( 6.0)      9 (10.0)   0.100
##   EdMother (mean (sd))  3.80 (1.49)   3.78 (1.51)   3.67 (1.54)   0.057
##   EdFather (mean (sd))  3.68 (1.65)   3.66 (1.73)   3.78 (1.73)   0.044
##   Age (mean (sd))      36.75 (8.95)  37.10 (9.41)  38.41 (9.49)   0.119
##   Employment (%)                                                  0.248
##      no                   95 (10.3)     24 (17.9)     18 (20.0)
##      part-time            75 ( 8.2)     20 (14.9)     11 (12.2)
##      full-time           748 (81.5)     90 (67.2)     61 (67.8)
##   Income (mean (sd))    5.10 (2.24)   5.04 (2.60)   4.69 (2.51)   0.111
##   Transfer (mean (sd)) 51.40 (24.38) 57.37 (25.10) 51.61 (26.39)  0.158
##   GPA (mean (sd))       3.16 (0.58)   3.16 (0.46)   3.24 (0.58)   0.097

## Examine all pairwise SMDs
ExtractSmd(tab1Unadj)

##               average      1 vs 2      1 vs 3     2 vs 3
## Gender     0.28718081 0.431825669 0.369462797 0.06025398
## Ethnicity  0.09475231 0.004540496 0.137619463 0.14209699
## Military   0.20773590 0.217301900 0.312032587 0.09387322
## ESL        0.09955894 0.089842245 0.059753148 0.14908142
## EdMother   0.05735067 0.014182489 0.086066827 0.07180268
## EdFather   0.04433253 0.007919139 0.059274560 0.06580389
## Age        0.11889003 0.038429226 0.179969129 0.13827175
## Employment 0.24838203 0.332479394 0.324590337 0.08807636
## Income     0.11113230 0.025003114 0.171951403 0.13644238
## Transfer   0.15777889 0.241327245 0.008454888 0.22355453
## GPA        0.09651297 0.009213587 0.128230886 0.15209444
```

4 Propensity score modeling

As the exposure is a three-category variable, the propensity score model can be modeled using multinomial logistic regression. In R, the VGAM (vector generalized linear and additive models) package provides a flexible framework for this. Because the sample size of the treatment 2 group is small, making flexible modeling difficult, the ordinal variables are used only as linear terms. Predicting the "response" gives predicted probabilities of each treatment as a (sample size) × 3 matrix, which then can be added to the dataset. The following AddGPS function can be used to ease this process. Three propensity scores (one for each treatment category) are added to the dataset.

```
## Function to add generalized PS to dataset
AddGPS <- function(data, formula, family = multinomial(), psPrefix = "PS_") {
    library(VGAM)
    ## Fit multinomial logistic regression
    resVglm <- vglm(formula = formula, data = data, family = family)
    ## Calculate PS
    psData <- as.data.frame(predict(resVglm, type = "response"))
    names(psData) <- paste0(psPrefix, names(psData))
    cbind(data, psData)
}

tutoring <- AddGPS(data = tutoring, # dataset
                   ## Propensity score model for multinomial regression
                   formula = treat ˜ Gender + Ethnicity + Military +
                                 ESL + EdMother + EdFather + Age +
                                 Employment + Income + Transfer + GPA)
```

5 Weight creation

As mentioned in the text, the matching weight is defined as follows.

$$\begin{aligned} MW_i &= \frac{\text{Smallest PS}}{\text{PS of assigned treatment}} \\ &= \frac{\min(e_{1i}, ..., e_{Ki})}{\sum_{k=1}^{K} I(Z_i = k) e_{ki}} \end{aligned}$$

where e_{ki} is the i-th individual's probability of being assigned to the k-th treatment category given the covariate pattern, $Z_i \in \{1, ..., K\}$ is the categorical variable indicating the i-th individual's treatment assignment.

The following function can be used to add matching weight to the dataset. Individuals' matching weights have a range of [0,1], where as the inverse probability treatment weights have a range of [1,∞].

```r
## Function to add matching weight as mw to dataset
AddMwToData <- function(data, txVar, txLevels, psPrefix = "PS_") {
    ## Treatment indicator data frame (any number of groups allowed)
    dfAssign <- as.data.frame(lapply(txLevels, function(tx_k) {
        as.numeric(data[txVar] == tx_k)
    }))
    ## Name of PS variables
    psVars <- paste0(psPrefix, txLevels)
    ## Pick denominator (PS for assigned treatment)
    data$PS_assign <- rowSums(data[psVars] * dfAssign)
    ## Pick numerator
    data$PS_min <- do.call(pmin, data[psVars])
    ## Calculate the IPTW
    data$iptw <- 1 / data$PS_assign
    ## Calculate the matching weight
    data$mw <- exp(log(data$PS_min) - log(data$PS_assign))
    ## Return the whole data
    data
}

## Add IPTW and MW
tutoring <- AddMwToData(data = tutoring, # dataset
                        txVar = "treat", # treatment variable name
                        tx = c("Control", "Treat1", "Treat2")) # treatment levels

## Check how weights are defined
head(tutoring[c("treat","PS_Control","PS_Treat1","PS_Treat2",
                "PS_assign","PS_min","iptw","mw")], 20)

##      treat PS_Control  PS_Treat1  PS_Treat2  PS_assign     PS_min      iptw         mw
## 3  Control  0.8192816 0.11440448 0.06631388 0.81928164 0.06631388  1.220581 0.08094149
## 4  Control  0.8313205 0.10516348 0.06351606 0.83132046 0.06351606  1.202906 0.07640383
## 11 Control  0.6346235 0.22597339 0.13940309 0.63462352 0.13940309  1.575737 0.21966266
## 12 Control  0.7203265 0.11853269 0.16114082 0.72032649 0.11853269  1.388259 0.16455412
## 14 Control  0.6759314 0.15931947 0.16474916 0.67593137 0.15931947  1.479440 0.23570361
## 16  Treat1  0.7278386 0.18054526 0.09161616 0.18054526 0.09161616  5.538777 0.50744155
## 17 Control  0.7963014 0.09228518 0.11141339 0.79630143 0.09228518  1.255806 0.11589227
## 18 Control  0.7963014 0.09228518 0.11141339 0.79630143 0.09228518  1.255806 0.11589227
## 19 Control  0.4011609 0.29293705 0.30590201 0.40116094 0.29293705  2.492765 0.73022327
## 23 Control  0.7980564 0.14170696 0.06023666 0.79805638 0.06023666  1.253044 0.07547920
## 28  Treat2  0.7696177 0.11208565 0.11829667 0.11829667 0.11208565  8.453323 0.94749620
## 31  Treat1  0.7876534 0.11912070 0.09322587 0.11912070 0.09322587  8.394847 0.78261688
## 32 Control  0.7602112 0.13218394 0.10760486 0.76021120 0.10760486  1.315424 0.14154600
## 34  Treat2  0.6994628 0.12694918 0.17358797 0.17358797 0.12694918  5.760768 0.73132478
## 38  Treat1  0.6359332 0.24401948 0.12004734 0.24401948 0.12004734  4.098034 0.49195804
## 39 Control  0.7523881 0.15006473 0.09754713 0.75238814 0.09754713  1.329101 0.12965001
## 40 Control  0.8281320 0.11921012 0.05265789 0.82813199 0.05265789  1.207537 0.06358635
## 49  Treat1  0.7963180 0.09950924 0.10417277 0.09950924 0.09950924 10.049318 1.00000000
## 50 Control  0.8929612 0.06199434 0.04504442 0.89296124 0.04504442  1.119869 0.05044387
## 51 Control  0.6910650 0.16455995 0.14437500 0.69106505 0.14437500  1.447042 0.20891666

## Check weight distribution
summary(tutoring[c("mw","iptw")])

##        mw                iptw
##  Min.   :0.01025   Min.   : 1.052
##  1st Qu.:0.05546   1st Qu.: 1.154
##  Median :0.09410   Median : 1.258
##  Mean   :0.21706   Mean   : 3.066
##  3rd Qu.:0.17721   3rd Qu.: 1.465
##  Max.   :1.00000   Max.   :46.446
```

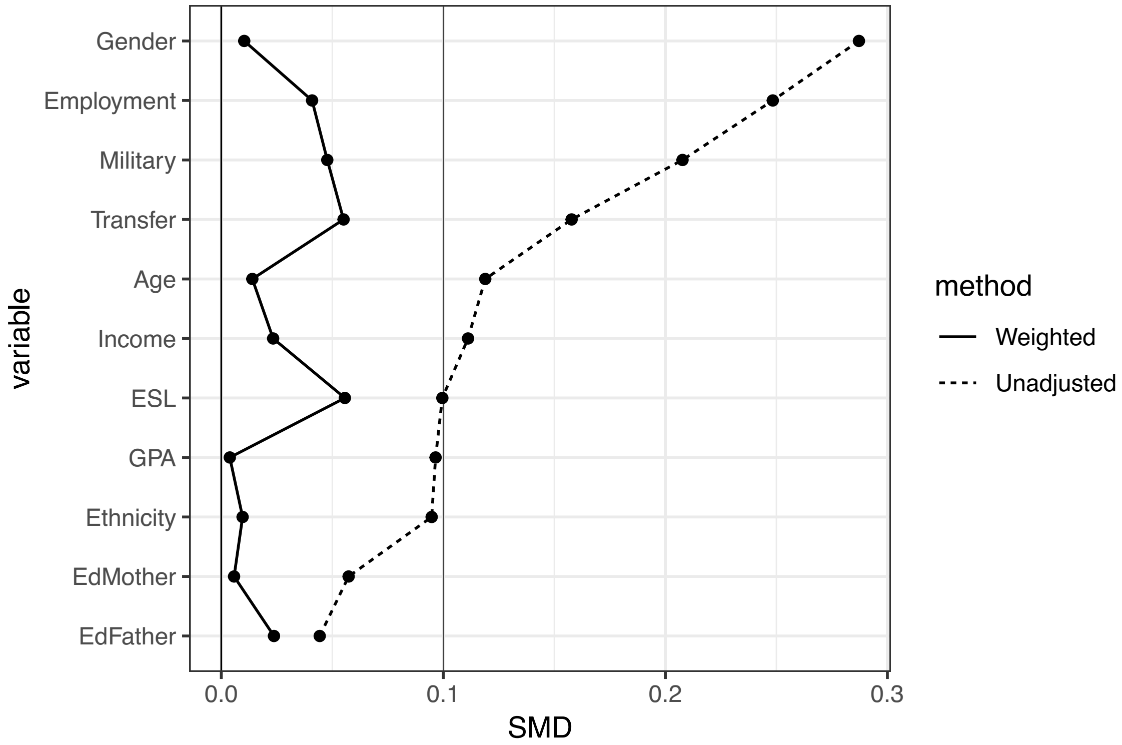
6 Post-weighting balance assessment

All analyses afterward should be proceeded as weighted analyses. In R, this is most easily achieved by using the survey package. Firstly, a survey design object must be created with svydesign function. The resulting object is then used as the dataset. The weighted covariate table can be constructed with the tableone package. All SMDs are less than 0.1 after weighting, indicating better covariate balance.

```
## Created weighted data object
library(survey)
tutoringSvy <- svydesign(ids = ~ 1, data = tutoring, weights = ~ mw)

## Weighted table with tableone
tab1Mw <- svyCreateTableOne(vars = covariates, strata = "treat", data = tutoringSvy)
print(tab1Mw, test = FALSE, smd = TRUE)

##                         Stratified by treat
##                          Control       Treat1        Treat2        SMD
##   n                       82.8          82.6          82.5
##   Gender = MALE (%)       24.9 (30.1)   25.0 (30.3)   24.4 (29.6)   0.010
##   Ethnicity (%)                                                     0.010
##      Black                18.9 (22.9)   19.2 (23.3)   18.8 (22.8)
##      Other                11.7 (14.1)   11.3 (13.7)   11.6 (14.1)
##      White                52.2 (63.0)   52.1 (63.1)   52.0 (63.0)
##   Military = TRUE (%)     17.2 (20.8)   19.7 (23.8)   17.4 (21.1)   0.048
##   ESL = TRUE (%)           6.1 ( 7.4)    6.4 ( 7.7)    8.1 ( 9.8)   0.056
##   EdMother (mean (sd))    3.66 (1.49)   3.65 (1.47)   3.65 (1.55)   0.006
##   EdFather (mean (sd))    3.71 (1.70)   3.66 (1.75)   3.73 (1.70)   0.024
##   Age (mean (sd))        38.13 (9.68)  38.21 (9.63)  38.01 (9.38)   0.014
##   Employment (%)                                                    0.041
##      no                   16.3 (19.7)   15.6 (18.9)   15.2 (18.4)
##      part-time            10.2 (12.3)    9.2 (11.2)   10.5 (12.7)
##      full-time            56.3 (68.0)   57.7 (69.9)   56.8 (68.9)
##   Income (mean (sd))      4.76 (2.35)   4.72 (2.47)   4.80 (2.47)   0.023
##   Transfer (mean (sd)) 52.46 (24.04) 51.39 (25.02) 53.48 (26.19)   0.055
##   GPA (mean (sd))         3.21 (0.49)   3.21 (0.45)   3.21 (0.59)   0.004

## All pairwise SMDs
ExtractSmd(tab1Mw)

##                average      1 vs 2       1 vs 3       2 vs 3
## Gender     0.010336859 0.004393687 0.0111115330 0.0155053556
## Ethnicity  0.009595945 0.013881066 0.0006174629 0.0142893048
## Military   0.047738733 0.071609306 0.0067821033 0.0648247896
## ESL        0.055666107 0.010019487 0.0834804231 0.0734984115
## EdMother   0.005765913 0.008755059 0.0082762793 0.0002663992
## EdFather   0.023721214 0.024874520 0.0107632204 0.0355259006
## Age        0.013982735 0.008033386 0.0128645704 0.0210502478
## Employment 0.040896810 0.043102022 0.0330741322 0.0465142771
## Income     0.023351441 0.019691181 0.0157469189 0.0346162234
## Transfer   0.055073782 0.043293809 0.0406028456 0.0813246930
## GPA        0.003834104 0.006104611 0.0018132523 0.0035844491
```

Visualizing the covariate balance before and after weighting can sometimes be helpful. Extracted SMD data can be fed to a plotting function (here ggplot2).

```
## Create SMD data frame
dataPlot <- data.frame(variable   = rownames(ExtractSmd(tab1Unadj)),
                       Unadjusted = ExtractSmd(tab1Unadj)[,"average"],
                       Weighted   = ExtractSmd(tab1Mw)[,"average"])
## Reshape to long format
library(reshape2)
dataPlotMelt <- melt(data          = dataPlot,
                     id.vars       = "variable",
                     variable.name = "method",
                     value.name    = "SMD")
## Variables names ordered by unadjusted SMD values
varsOrderedBySmd <- rownames(dataPlot)[order(dataPlot[,"Unadjusted"])]
## Reorder factor levels
dataPlotMelt$variable <- factor(dataPlotMelt$variable,
                                levels = varsOrderedBySmd)
dataPlotMelt$method <- factor(dataPlotMelt$method,
                              levels = c("Weighted","Unadjusted"))
## Plot
library(ggplot2)
ggplot(data = dataPlotMelt,
       mapping = aes(x = variable, y = SMD, group = method, linetype = method)) +
    geom_line() +
    geom_point() +
    geom_hline(yintercept = 0, size = 0.3) +
    geom_hline(yintercept = 0.1, size = 0.1) +
    coord_flip() +
    theme_bw() + theme(legend.key = element_blank())
```



7 Outcome analysis

The outcome analyses should also be proceeded as weighted analyses. Most functions in the survey package is named svy⋆ with ⋆ being the name of the unweighted counterpart.

The outcome was handled as a continuous outcome for simplicity. In weighted linear regression, both treatments appear superior to the control without tutoring regarding the course grade assuming the propensity score model was correctly specified. The mean difference was 0.45 [0.23, 0.67] for treatment 1 vs control and 0.67 [0.45, 0.89] for treatment 2 vs control.

```
## Weighted group means of Grade
svyby(formula = ~ Grade, by = ~ treat, design = tutoringSvy, FUN = svymean)

##           treat    Grade         se
## Control Control 2.792759 0.06648740
## Treat1   Treat1 3.244832 0.09179853
## Treat2   Treat2 3.463329 0.09070431

## Group difference tested in weighted regression
modelOutcome1 <- svyglm(formula = Grade ~ treat, design = tutoringSvy)
summary(modelOutcome1)

##
## Call:
## svyglm(formula = Grade ~ treat, design = tutoringSvy)
##
## Survey design:
## svydesign(ids = ~1, data = tutoring, weights = ~mw)
##
## Coefficients:
##             Estimate Std. Error t value      Pr(>|t|)
## (Intercept)  2.79276    0.06649  42.004       < 2e-16 ***
## treatTreat1  0.45207    0.11335   3.988 0.00007076303 ***
## treatTreat2  0.67057    0.11246   5.963 0.00000000331 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 1.394533)
##
## Number of Fisher Scoring iterations: 2

## ShowRegTable in tableone may come in handy
ShowRegTable(modelOutcome1, exp = FALSE)

##             coef [confint]    p
## (Intercept) 2.79 [2.66, 2.92] <0.001
## treatTreat1 0.45 [0.23, 0.67] <0.001
## treatTreat2 0.67 [0.45, 0.89] <0.001
```

8 Bootstrapping

As discussed in the text, bootstrapping may provide better variance estimates than model-based inference. The `boot` package is a general purpose bootstrapping package. The following context-specific wrapper functions can be used to simplify the process. In this specific example, the bootstrap confidence intervals for the treatment effects were somewhat narrower.

```
## Define a function for each bootstrap step
BootModelsConstructor <- function(formulaPs, formulaOutcome, OutcomeRegFun, ...) {
    ## Obtain treatment variable name
    txVar <- as.character(formulaPs[[2]])
    ## Return a function
    function(data, i) {
        ## Obtain treatment levels
        txLevels <- names(table(data[,txVar]))
        ## Add generalized propensity scores
        dataB <- AddGPS(data = data[i,], formula = formulaPs)
        ## Add matching weight
        dataB <- AddMwToData(data = dataB, txVar = txVar, txLevels = txLevels)
        ## Weighted analysis (lm() ok as only the estimates are used)
        lmWeighted <- OutcomeRegFun(formula = formulaOutcome, data = dataB,
                                    weights = mw, ...)
        ## Extract coefs
        coef(lmWeighted)
    }
}

## Define a function to summarize bootstrapping
BootSummarize <- function(data, R, BootModels, level = 0.95, ...) {
    ## Use boot library
    library(boot)
    ## Run bootstrapping
    outBoot       <- boot(data = data, statistic = BootModels, R = R, ...)
    out           <- outBoot$t
    colnames(out) <- names(outBoot$t0)
    ## Confidence intervals
    lower <- apply(out, MARGIN = 2, quantile, probs = (1 - level) / 2)
    upper <- apply(out, MARGIN = 2, quantile, probs = (1 - level) / 2 + level)
    outCi <- cbind(lower = lower, upper = upper)
    ## Variance of estimator
    outVar <- apply(out, MARGIN = 2, var)
    outSe  <- sqrt(outVar)
    ## Return as a readable table
    cbind(est = outBoot$t0, outCi, var = outVar, se = outSe)
}

## Construct a custom bootstrap function with specific formulae
## formulaPs is propensity score model
BootModels <- BootModelsConstructor(formulaPs = treat ~ Gender + Ethnicity + Military +
                                                ESL + EdMother + EdFather + Age +
                                                Employment + Income + Transfer + GPA,
                                    ## Outcome model
                                    formulaOutcome = Grade ~ treat,
                                    ## Regression function for outcome model
                                    OutcomeRegFun = lm)

## Use a clean dataset without PS and weight variables
data(tutoring)
## Make employment categorical
tutoring$Employment <- factor(tutoring$Employment, levels = 1:3,
                              labels = c("no","part-time","full-time"))
## Run bootstrap
set.seed(201508131)
system.time(bootOut1 <- BootSummarize(data = tutoring, R = 2000, BootModels = BootModels,
                                      parallel = "multicore", ncpus = 12))

##    user  system elapsed
## 159.201  13.593  17.688

bootOut1

##                   est     lower     upper         var         se
## (Intercept) 2.7927593 2.6130814 2.9872607 0.008972568 0.09472364
## treatTreat1 0.4520730 0.2325361 0.6577786 0.011831058 0.10877067
## treatTreat2 0.6705692 0.4626595 0.8484488 0.009776627 0.09887683

## Show naive confidence interval again
ShowRegTable(modelOutcome1, exp = FALSE, digits = 7)

##             coef [confint]                   p
## (Intercept) 2.7927593 [2.6624464, 2.9230722] <0.001
## treatTreat1 0.4520730 [0.2299169, 0.6742290] <0.001
## treatTreat2 0.6705692 [0.4501465, 0.8909920] <0.001
```

Chapter 2: Effects of Analgesics on Bone Mineral Density: a Longitudinal Analysis of the Prospective SWAN Cohort with Three-group Matching Weights

AUTHORS: Kazuki Yoshida(1,2,3), Zhi Yu(3,4), Gail A. Greendale(4), Kristine Ruppert(5), Yinjuan Lian(5), Sara K. Tedeschi(3), Tzu-Chieh Lin(3), Sebastien Haneuse(2), Robert J. Glynn(2, 6), Sonia Hernandez-Diaz(1), Daniel H. Solomon(3,6)

AFFILIATIONS

1. Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, United States.
2. Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, United States.
3. Division of Rheumatology, Immunology and Allergy, Brigham and Women's Hospital, Boston, Massachusetts, United States.
4. Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA
5. Division of Geriatrics, David Geffen School of Medicine at UCLA, Los Angeles, California, United States
6. Department of Epidemiology, University of Pittsburgh, Pittsburgh, Pennsylvania, United States
7. Division of Pharmacoepidemiology and Pharmacoeconomics, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts, United States.

DISCLOSURES:

KY receives tuition support jointly from Japan Student Services Organization (JASSO) and Harvard T.H. Chan School of Public Health (partially supported by training grants from Pfizer, Takeda, Bayer and PhRMA). RJG received research support in the form of grants to his institution for clinical trial design, monitoring, and analysis from AstraZeneca, Kowa, Novartis, and Pfizer. DHS receives salary support from institutional research grants from Eli Lilly, Amgen, Pfizer, AstraZeneca, Genentech, and Corrona. He also receives royalties from UpToDate, and serves in unpaid roles in studies funded by Pfizer and Eli Lilly.

FUNDING INFORMATION:

KY receives tuition support jointly from Japan Student Services Organization (JASSO) and Harvard T.H. Chan School of Public Health (partially supported by training grants from Pfizer, Takeda, Bayer and PhRMA). SKT receives salary support from the Lupus Foundation of America and NIH-L30 AR070514. DHS receives salary support from NIH-K24AR055989.

SWAN ACKNOWLEDGEMENT:

The Study of Women's Health Across the Nation (SWAN) has grant support from the National Institutes of Health (NIH), DHHS, through the National Institute on Aging (NIA), the National Institute of Nursing Research (NINR) and the NIH Office of Research on Women’s Health (ORWH) (Grants U01NR004061; U01AG012505, U01AG012535, U01AG012531, U01AG012539, U01AG012546, U01AG012553, U01AG012554, U01AG012495). The content of this article is solely the responsibility of the authors and does not necessarily represent the official views of the NIA, NINR, ORWH or the NIH.

Clinical Centers: *University of Michigan, Ann Arbor – Siobán Harlow, PI 2011 – present, MaryFran Sowers, PI 1994-2011*; *Massachusetts General Hospital, Boston, MA – Joel Finkelstein, PI 1999 – present*; *Robert Neer, PI 1994 – 1999; Rush University, Rush University Medical Center, Chicago, IL – Howard Kravitz, PI 2009 – present*; *Lynda Powell, PI 1994 – 2009; University of California, Davis/Kaiser – Ellen Gold, PI*; *University of California, Los Angeles – Gail Greendale, PI*; *Albert Einstein College of Medicine, Bronx, NY – Carol Derby, PI 2011 – present, Rachel Wildman, PI 2010 – 2011; Nanette Santoro, PI 2004 – 2010; University of Medicine and Dentistry – New Jersey Medical School, Newark – Gerson Weiss, PI 1994 – 2004;* and the *University of Pittsburgh, Pittsburgh, PA – Karen Matthews, PI.*

NIH Program Office: *National Institute on Aging, Bethesda, MD – Chhanda Dutta 2016-present; Winifred Rossi 2012–2016; Sherry Sherman 1994 – 2012; Marcia Ory 1994 – 2001; National Institute of Nursing Research, Bethesda, MD – Program Officers.*

Central Laboratory: *University of Michigan, Ann Arbor – Daniel McConnell* (Central Ligand Assay Satellite Services).

Coordinating Center: *University of Pittsburgh, Pittsburgh, PA – Maria Mori Brooks, PI 2012 - present; Kim Sutton-Tyrrell, PI 2001 – 2012; New England Research Institutes, Watertown, MA - Sonja McKinlay, PI 1995 – 2001.*

Steering Committee: Susan Johnson, Current Chair
Chris Gallagher, Former Chair

We thank the study staff at each site and all the women who participated in SWAN.

ABSTRACT

PURPOSE: To examine the effects of analgesics on bone mineral density (BMD), which have not been examined in a longitudinal study with multiple measurements.

METHODS: We investigated changes in BMD associated with new use of analgesics in a prospective longitudinal cohort of mid-life women. BMD and medication use were measured annually. We compared BMD among new users of acetaminophen, NSAIDs, and opioids. Adjustment for baseline covariates was conducted through propensity score matching weights. On-treatment analysis was conducted with inverse probability of censoring weights. Analysis based on the initial treatment group was also conducted to provide insights into selection bias. Repeated BMD measurements were examined with generalized estimating equations.

RESULTS: We identified 71 acetaminophen new users, 659 NSAID new users, and 84 opioid new users among 2,365 participants. In the on-treatment analysis, the opioid group in comparison to the acetaminophen group had an additional average BMD decline of -0.06% [-1.24, 1.11] per year in the spine and -0.45% [-1.51, 0.61] per year in the femoral neck. BMD mean trajectories over time suggested a fifth-year decline in the opioid persistent users compared to other two groups. In the initial treatment group analysis, all three groups showed similar trajectories.

CONCLUSION: The BMD decline over time was similar among the three groups. However, five years of continuous opioid use may be associated with a greater BMD decline than five years on other analgesics. Further studies examining the relationship between very long term persistent opioid use and BMD are warranted.

INTRODUCTION

Many classes of drugs have been linked to increased risk of fractures or reduced bone mineral density (BMD). Histamine H2 receptor antagonists[1], opioids, and anticonvulsants[2] were among them. The cross-sectional nature of these studies, however, limit the assessment of temporality. Additionally, long-term effects of analgesics on bone health are not well understood.

Opioids have been associated with increased fracture risks in multiple longitudinal studies[3–8]. The increased risks occurring soon after initiation[4,6] suggest the primary mechanism is through acute neurologic effects, such as gait imbalance. However, chronic opioid use may also have indirect effects via endocrine changes[9–11]; for example, hypogonadotropic hypogonadism has been found in patients receiving methadone maintenance therapy[12]. Several studies also suggest lower BMD in opioid users[13–15]. However, these studies were cross-sectional, had limited control of confounding, and focused on a particular subset of chronic opioid users (*i.e.*, former heroin addicts on methadone maintenance).

NSAID use has been associated with higher BMD in two cross-sectional studies[16,17] adjusted for potential confounders such as body weight, although a more recent study found increased fracture risk among NSAID users despite stable BMD[5]. Both selective and non-selective NSAIDs exhibit anti-inflammatory properties through inhibiting cyclooxygenase(COX)-2, an enzyme that plays a role in prostaglandins synthesis. Prostaglandins, in turn, play important roles in both bone formation and bone resorption[18].

The Study of Women's Health Across the Nation (SWAN)[19,20] allows for rigorous assessment of the effects of analgesics on BMD because of its longitudinal design and repeatedly measured BMD. We hypothesized that opioids and NSAIDs are associated with BMD reduction compared to acetaminophen (active control). Presence of three treatment groups of quite different sizes as well as frequent treatment changes posed challenges in analysis. Thus, we used recently proposed matching weights in a multiple group setting[21,22] along with inverse probability of censoring weights over time[23].

METHODS

Study population and design

SWAN[19,20], a prospective longitudinal, community-based cohort study of mid-life women, enrolled participants in their pre-menopause between 1996 and 1997 from 8 U.S. sites to observe the natural history of menopause. Eligibility criteria included age 42-52 years old, at least one menstrual period within the past three months, and no hormonal medication use within the last three months. The SWAN BMD substudy enrolled 2,365 women of four racial/ethnic groups (1,177 Caucasian, 665 African American, 273 Japanese, and 250 Chinese) with approximately annual BMD measurement. Longitudinal follow-up is still ongoing, and SWAN data collection consists of physical measures, fasting morning blood draw, interviewer-administered and self-administered questionnaires (completed at home or in clinic). Participants gave written informed consent and study sites obtained institutional review board approval.

Exposure assessment

The exposure of interest was the type of analgesic --opioid, NSAID including COX-2 selective inhibitors, and/or acetaminophen-- that participants took at ≥ 2 consecutive annual visits. The individual-specific baseline visit was defined as the visit immediately before the first of these consecutive visits. Medication use, including both prescription and over the counter (OTC), was ascertained through interviewer-administered questionnaire for medications used twice or more per week during the past month and was then verified by inspection of medication containers. The exposure definition was constructed *hierarchically* (**eTable 2-1**): opioid user if an opioid is used regardless of the other two; NSAID user if an NSAID is used but not opioids regardless of acetaminophen; and acetaminophen user if it is the only analgesic used. Participants who transitioned between these exposure categories were assigned the exposure status at the time when they first met the eligibility criteria.

Outcome assessment

Details of the BMD measurements have been described in previous studies using SWAN[24–26]. BMD (g/cm^2) was measured in the lumbar spine and femoral neck at each study visit. Raw BMD measurements were converted to baseline-normalized %BMD values for interpretability, as regularly done in major osteoporosis clinical trials[27–30]. That is, for each individual, the outcome was defined as 100% at the individual-specific baseline visit when the covariates were ascertained (year 0), and subsequent values were described in relation to this baseline value (e.g., 96% of the baseline value at year 4). Follow-up was truncated at year 5 because very few people remained in the initial treatment categories beyond that point.

Covariate assessment

Covariates were assessed at the individual-specific baseline visit. Body mass index (BMI) was calculated from height and weight at the study baseline. The demographic variables included age, race/ethnicity, self-reported annual income (low [≤ $19,999], medium [$20,000-49,999], and high [≥ $50,000-]), and college education (yes/no). Alcohol intake (none/low [< 1 drink/month], moderate [up to 1/week], and high [≥ 2/week]), current tobacco use (yes/no), and physical activity measures were available as lifestyle variables. Physical activity was measured using the modified Baecke Physical Activity Questionnaire (range 3–15, with lower scores indicating less exercise)[31,32]. Self-reported comorbidities included thyroid disease, diabetes, and history of cancer. Self-reported pain-related quality of life (range 0-100, with 100 indicating excellent quality of life[33]), vasomotor symptoms, and overall perception of health were also reported. Medications included hormone therapy for menopause, bisphosphonates, calcium supplements, vitamin D supplements, and oral glucocorticoids. Menopause transition (MT) stage was defined based on menstrual cycles[25] (**eTable 2-2**). We created four categories of MT stages for the main analysis: pre- or early perimenopause; late perimenopause; postmenopause; and unknown (**eTable 2-2**). We also conducted a subgroup analysis among those who had a known date of the final menstrual period (FMP), using MT stages based on time prior to or after the FMP (**eTable 2-3**).[24]

Statistical analyses

Participant characteristics at the study baseline were summarized within each exposure group. To examine between group imbalance in the unmatched cohort of patients, the standardized mean differences (SMD)[34] were calculated in each pairwise treatment contrast and then averaged across all three contrasts. The SMD represents how different groups are for a given covariate. Covariates that have SMD ≤ 0.1 are considered reasonably balanced[34]. We multiply imputed missing covariates via the *mice* R package.[35,36]

Multinomial logistic regression was used for the propensity score (PS) model because the exposure status had three categories (acetaminophen, NSAIDs, or opioids)[37], resulting in one PS for each exposure category. All baseline covariates listed in the baseline Table 2-(**Table 2-1**) were included as explanatory variables. We used the PSs as *matching weights* (MW; **eAppendix Methods**), a PS weighting method proposed by Li and Greene[21]. A recent study generalized MW to multiple treatment group settings[22]. Compared to 1:1:1 PS matching, MW allows for retention of all subjects, which is a potential advantage when the group sizes are dissimilar. Compared to the conventional inverse probability of treatment weights (IPTW), the target of inference focuses on those who are in clinical equipoise among all drugs (i.e., similar estimand to PS matching). This clinical equipoise estimand was more stably estimated in the settings where baseline covariates were more different among groups[22].

MW, as it is known currently, is only applicable to time-invariant exposure. However, a drug exposure is typically time-varying. Therefore, we used the *on-treatment analysis* and *initial treatment group analysis* to make treatment group assignment effectively time-invariant. The main analysis was *on-treatment* analysis of those who remained in the initial treatment category (adherers). That is, those who deviated from their initial category were censored at the time of deviation, making the treatment assignment effectively time-invariant among uncensored time points remaining in the analysis dataset. We additionally censored patients at the initiation of hormone therapy for menopause or bisphosphonate or cancer diagnosis. Such censoring of participants who deviate from the initial treatment status or started bone active medications can introduce selection bias -- those who are censored and retained may not share the same risks for BMD changes. Thus, we additionally assigned time-varying inverse probability of censoring weights (IPCW)[23] to ameliorate this selection bias issue using the same set of covariates as the time-invariant MW model, but updated for each time point. A final weight for a given time point was constructed as the product of the individual-specific time-invariant MW and the individual-specific, time point-specific time-varying IPCW and was normalized to represent the sample size of each treatment group at each time point.[38] This approach should estimate the effect of continuous treatment,[23,39] assuming both MW for baseline confounding by indication and IPCW for selection bias introduced by artificial censoring are successful. We also conducted an alternative analysis based on the initial treatment category at the study baseline (*initial treatment* *group* analysis). Participants remained in their original treatment category regardless of subsequent medication changes in this analysis, also making the treatment variable time-invariant. This approach is an observational analogue of the intention-to-treat analysis used in clinical trials, and should estimate the effect of assigned treatment[39], assuming MW for baseline confounding by indication is successful. Censoring also occurred administratively because some subjects started analgesics late in the SWAN study, thus, reaching the latest SWAN visit (visit 13) before having the fifth-year visit after analgesic initiation. This type of administrative censoring was assumed non-informative.

The mean baseline-normalized %BMD over time for the spine and femoral neck were plotted in both the on-treatment analysis and initial treatment group analysis. We used the generalized estimating equation with the auto-regressive correlation structure to account for weighting and the clustering of repeated BMD measurements within each individual during follow-up. Confidence intervals were calculated based on robust sandwich standard error estimates. The time effect on the mean baseline-normalized %BMD was modeled as a linear term to provide average yearly change estimates. The slope differences of interest, NSAIDs versus acetaminophen, and opioids versus acetaminophen, were incorporated into the model as time-group interaction terms. We repeated the analyses in the FMP subgroup. We also repeated the main analysis after excluding an outlying data point as a sensitivity analysis. Another sensitivity analysis for the outcome model further adjusted for variables that had SMD > 0.1 after balancing by MW.

RESULTS

Study population

Among 2,365 participants in the SWAN BMD cohort, 71 acetaminophen new users, 659 NSAID new users, and 84 opioid new users were identified (**eFigure 2-1**; break down by generic names in **eTable 2-4**). Their unadjusted baseline characteristics are shown in **Table 2-1**. The most prominent baseline differences were noted for pain-related quality of life (QoL), ethnic composition, income, overall perception of health, BMI, femoral neck BMD, and physical activity. The pain-related QoL was lower for the opioid users (48.8) compared to the other two groups that had scores around 70. Femoral neck BMD was higher in the opioid group than the other groups likely associated with their higher BMI. Physical activity was highest among NSAID users and was lowest among opioid users. Twenty-six percent of NSAID users were also exposed to acetaminophen. Opioid users also had substantial concurrent exposure (acetaminophen 80% and NSAIDs 66%). Matching weights reduced group imbalance at the baseline (**Table 2-2**), even in comparison to other PS methods (**eFigure 2-2**)[40,41]. The mean follow-up durations were similar across treatment groups (**eTable 2-5**).

Adjusted main analysis using menstrual period-defined stages

Figure 2-1 shows the mean baseline-normalized BMD over the five-year follow-up period for each treatment group (n = 814) as well as the treatment group contrasts from the generalized estimating equation (see **eFigure 2-3** for unadjusted counterpart). The mean annual change in each treatment group as well as group differences in slopes are shown in **Table 2-3**.

The on-treatment analysis (**Figure 2-1, left panels**) was suggestive of a greater decline in BMD in the opioid group compared to the acetaminophen group, principally at the fifth year.

The opioid group in comparison to the acetaminophen group had an additional mean BMD decline of -0.06% [-1.24, 1.11] per year in the spine and -0.45% [-1.51, 0.61] per year in the femoral neck. The initial treatment group analysis, on the other hand, demonstrated more similar trajectories for all three groups (**Figure 2-1, right panels**). The difference between the opioid group and the acetaminophen group diminished to -0.06% [-0.66, 0.78] in the spine and to 0.08% [-0.65, 0.82] in the femoral neck.

Adjusted final menstrual period-based analysis

eFigure 2-4 shows the corresponding outcome analysis in the subgroup of women with a known FMP date (n = 471). The adjustment for the menopause transition stages at individual-specific baseline visit was based on the time prior to or after FMP (pre-transmenopause, transmenopause, or postmenopause; **eTable 2-3**)[24]. The baseline characteristics before propensity score weighting are in **eTable 2-6**. Propensity score weighing improved covariates balance, but to a lesser extent than in the main cohort (**eTable 2-7**). The mean trajectories were less stable due to the smaller sample size, particularly in the on-treatment analyses. The mean annual change in each treatment group is shown in **eTable 2-8**. The on-treatment analyses exhibited overlapping mean trajectories (**eFigure 2-4, left panels**). The initial treatment group analyses (**eFigure 2-4, right panels**) produced trajectories with more separation than the main initial treatment group analyses (**Figure 2-1, right panels**).

Sensitivity analysis

As the main on-treatment analysis showed a strong fifth-year deflection in the trajectory, we examined for the presence of outliers. One subject with probable thyroid disease exhibited an outlying decline trajectory. This subject remained in the opioid category for the full five years without meeting any of the censoring criteria, thus, she was gradually up-weighted over time via

IPCW, becoming more influential. Reanalysis excluding this subject (**eFigure 2-5**) resulted in a less prominent decline in the fifth year, although the opioid group remained the lowest group at the fifth year. Outcome analysis further adjusting for the sub-optimally balanced variables gave similar estimates of group differences in slopes (**eTable 2-9**).

DISCUSSION

In the current study, we examined the association between analgesic use and BMD decline over time in a well-established cohort of mid-life women, with a focus on the contrasts between opioids and acetaminophen as well as NSAIDs and acetaminophen. We used three-group MW for baseline covariate balancing and time-varying IPCW to reduce selection bias by artificial censoring over time. To our knowledge, the current study is the first instance of MW used in conjunction with IPCW in the multiple treatment group setting. The average slope differences were not statistically significant in both on-treatment analysis and initial treatment group analysis. However, the on-treatment analysis was suggestive of a potentially greater decline in the BMD in the opioid group compared to the acetaminophen group after five years of continuous use. The trajectory of BMD decline in the NSAID group was similar to the acetaminophen group. Between-group differences were not clearly observed in the initial treatment group analysis.

There is no established gold standard for the clinically meaningful group difference in BMD changes over time, however, several clinical trials were summarized in **eTable 2-10** to give some idea[27–30]. In the FIT study[27], which demonstrated hip fracture reduction, the annual slope difference in the femoral neck BMD was +1.0% / year in the alendronate group compared to the placebo group. Our study found that the annual slope difference was -0.45% / year [-1.51, 0.61] for the femoral neck BMD comparing the opioid new users to the acetaminophen new users

(**Figure 2-1**), which was not statistically significant, but did not rule out 1.0% difference in annual slopes. The five-year difference in BMD comparing the opioid group to the acetaminophen group was close to -10% in the on-treatment analysis although the difference was negligible in the initial treatment group analysis (**Figure 2-1**). The noticeable discrepancy between the on-treatment analysis and the initial treatment group analysis suggests the contribution of residual selection bias that was not fully controlled by IPCW, likely due to the small size of the opioid arm that remained on treatment, in addition to the exposure misclassification in the initial treatment group analysis. However, even in the sensitivity analysis removing an outlying observation, some group difference in the range of -3 to -5% remained, which may suggest a potentially greater decline in BMD among persistent opioid users.

Although the longitudinal association of opioid use and fractures has been well documented in multiple studies[3–8], the association of opioid use and lower BMD has been shown only in cross-sectional studies[13–15]. To our knowledge, only one study has examined the longitudinal effect of opioids on BMD[5] and reported no clinically relevant longitudinal association based on BMD measurements ten years apart. Our study provides additional insight into the potential effect of opioids on BMD by providing more granular follow-up, although this study alone is not conclusive. Past cross-sectional studies that demonstrated an association between opioid use and lower BMD were among former opioid abusers undergoing methadone therapy, whereas the current study was among community-dwelling healthy women.

Several studies have suggested potentially beneficial effects of NSAIDs on BMD. Bauer *et al*. found a cross-sectional association between higher BMD and current frequent NSAID use compared to infrequent use and non-use in their 1996 study on community-dwelling women aged at least 65 years old[16]. Carbone *et al*.[17] examined the cross-sectional association between

NSAID use and BMD in the Health ABC study among community-dwelling men and women 70-79 years of age. They found that current users of COX-2 selective NSAIDs with concurrent aspirin use had higher BMD than non-users. A 10-year longitudinal study by Vestergaard *et al.*[5], which also examined acetaminophen, NSAIDs, and opioids, found a very minor (clinically insignificant) increase in spine and whole body BMD among NSAID users compared to non-users. The current study showed essentially identical BMD trajectories between NSAID users and acetaminophen users in both the on-treatment analysis and the initial treatment group analysis.

SWAN was designed to characterize the biological, symptomatic, and psychosocial changes that occur during the menopausal transition and their effects on women's health and well-being. Thus, our findings may not generalize to men, or to women in different age ranges. SWAN did not specifically enroll analgesic users, thus, the number of users was small, limiting our ability to draw firm conclusions. Also SWAN does not have reliable medication dosage information. Doses of opioids can be highly variable among opioid users due to the highly individualized nature of these prescriptions[42]. However, high-dose opioid use is unlikely in this population cohort of generally healthy mid-life women.

Our longitudinal study design has some unique strengths compared to the prior cross-sectional studies on this topic. Use of acetaminophen as a comparator medication --active comparator design[43]-- ensured that all three treatment groups had at least some pain. Non-users -- individuals who do not use analgesics -- are expected to have much less pain than analgesic users, thus using such a comparator group without pain could induce a spurious association between BMD changes and medication use[44], which can be difficult to control for. We also used a new user design[43], which examines subjects starting the medication of interest, in an attempt to parallel the design of a hypothetical clinical trial[45] and ensures that the baseline covariates were measured before medication initiation.

As a safety outcome study, the primary effect of interest is the on-treatment effect[39], that is the effect of medication on the outcome if subjects were made to adhere to the regimen[23]. However, the naïve on-treatment analysis that simply censors subjects who do not follow the initial regimen of interest often introduces selection bias[46]. Therefore, we used IPCW to account for selection. The study revealed a difficulty of IPCW in the presence of small number of subjects in each arm. One of the few persistent opioid users happened to have an outlying decline in BMD, thereby exerting increasing influence at later time points because of progressively greater IPCW. Some of the differences in BMD trajectories, however, persisted after excluding this subject. Examination of the very long-term on-treatment effect beyond 5 years was not possible due to the very few adherers, potentially limiting the scope of the study.

In conclusion, the average BMD slope differences over a five-year period were not statistically significant among mid-life female analgesic new users. However, five years of persistent opioid use may be associated with a greater BMD decline. It is important to remember that chronic opioid use, although becoming common, is not a well-justified practice in the setting of non-cancer pain[47,48]. Further studies examining the relationship between very long term persistent opioid use and BMD as well as their dose response are warranted.

References

1. Kinjo M, Setoguchi S, Solomon DH. Antihistamine therapy and bone mineral density: analysis in a population-based US sample. *Am J Med* 2008; **121**: 1085–1091. doi:10.1016/j.amjmed.2008.06.036.

2. Kinjo M, Setoguchi S, Schneeweiss S, Solomon DH. Bone mineral density in subjects using central nervous system-active medications. *Am J Med* 2005; **118**: 1414. doi:10.1016/j.amjmed.2005.07.033.

3. Teng Z, Zhu Y, Wu F, *et al.* Opioids contribute to fracture risk: a meta-analysis of 8 cohort studies. *PLoS ONE* 2015; **10**: e0128232. doi:10.1371/journal.pone.0128232.

4. Li L, Setoguchi S, Cabral H, Jick S. Opioid use for noncancer pain and risk of fracture in adults: a nested case-control study using the general practice research database. *Am J Epidemiol* 2013; **178**: 559–569. doi:10.1093/aje/kwt013.

5. Vestergaard P, Hermann P, Jensen J-EB, Eiken P, Mosekilde L. Effects of paracetamol, non-steroidal anti-inflammatory drugs, acetylsalicylic acid, and opioids on bone mineral density and risk of fracture: results of the Danish Osteoporosis Prevention Study (DOPS). *Osteoporos Int* 2012; **23**: 1255–1265. doi:10.1007/s00198-011-1692-0.

6. Miller M, Stürmer T, Azrael D, Levin R, Solomon DH. Opioid analgesics and the risk of fractures in older adults with arthritis. *J Am Geriatr Soc* 2011; **59**: 430–438. doi:10.1111/j.1532-5415.2011.03318.x.

7. Saunders KW, Dunn KM, Merrill JO, *et al.* Relationship of opioid use and dosage levels to fractures in older chronic pain patients. *J Gen Intern Med* 2010; **25**: 310–315. doi:10.1007/s11606-009-1218-z.

8. Solomon DH, Rassen JA, Glynn RJ, Lee J, Levin R, Schneeweiss S. The comparative safety of analgesics in older adults with arthritis. *Arch Intern Med* 2010; **170**: 1968–1976. doi:10.1001/archinternmed.2010.391.

9. Coluzzi F, Pergolizzi J, Raffa RB, Mattia C. The unsolved case of “bone-impairing analgesics”: the endocrine effects of opioids on bone metabolism. *Ther Clin Risk Manag* 2015; **11**: 515–523. doi:10.2147/TCRM.S79409.

10. Katz N, Mazer NA. The impact of opioids on the endocrine system. *Clin J Pain* 2009; **25**: 170–175. doi:10.1097/AJP.0b013e3181850df6.

11. Merza Z. Chronic use of opioids and the endocrine system. *Horm Metab Res* 2010; **42**: 621–626. doi:10.1055/s-0030-1254099.

12. Hallinan R, Byrne A, Agho K, McMahon CG, Tynan P, Attia J. Hypogonadism in men receiving methadone and buprenorphine maintenance treatment. *International Journal of Andrology* 2009; **32**: 131–139. doi:10.1111/j.1365-2605.2007.00824.x.

13. Grey A, Rix-Trott K, Horne A, Gamble G, Bolland M, Reid IR. Decreased bone density in men on methadone maintenance therapy. *Addiction* 2011; **106**: 349–354. doi:10.1111/j.1360-0443.2010.03159.x.

14. Kim TW, Alford DP, Malabanan A, Holick MF, Samet JH. Low bone density in patients receiving methadone maintenance treatment. *Drug Alcohol Depend* 2006; **85**: 258–262. doi:10.1016/j.drugalcdep.2006.05.027.

15. Milos G, Gallo LM, Sosic B, *et al.* Bone mineral density in young women on methadone substitution. *Calcif Tissue Int* 2011; **89**: 228–233. doi:10.1007/s00223-011-9510-4.

16. Bauer DC, Orwoll ES, Fox KM, *et al.* Aspirin and NSAID use in older women: effect on bone mineral density and fracture risk. Study of Osteoporotic Fractures Research Group. *J Bone Miner Res* 1996; **11**: 29–35. doi:10.1002/jbmr.5650110106.

17. Carbone LD, Tylavsky FA, Cauley JA, *et al.* Association between bone mineral density and the use of nonsteroidal anti-inflammatory drugs and aspirin: impact of cyclooxygenase selectivity. *J Bone Miner Res* 2003; **18**: 1795–1802. doi:10.1359/jbmr.2003.18.10.1795.

18. Blackwell KA, Raisz LG, Pilbeam CC. Prostaglandins in bone: bad cop, good cop? *Trends Endocrinol Metab* 2010; **21**: 294–301. doi:10.1016/j.tem.2009.12.004.

19. Sowers M, Crawford SL, Sternfeld B, *et al.* SWAN: a multi-center, multi-ethnic, community-based cohort study of women and the menopausal transition. In: Lobo RA, Kelsey J, Marcus R (eds.) *Menopause: Biology and Pathobiology*. Academic Press, 2000; 175–188. Available at: https://www.researchgate.net/publication/43196481_SWAN_A_Multicenter_Multiethnic_Community-Based_Cohort_Study_of_Women_and_the_Menopausal_Transition. Accessed January 17, 2017.

20. Study of Women's Health Across the Nation (SWAN). *SWAN Study*. Available at: http://www.swanstudy.org. Accessed September 22, 2015.

21. Li L, Greene T. A weighting analogue to pair matching in propensity score analysis. *Int J Biostat* 2013; **9**: 215–234. doi:10.1515/ijb-2012-0030.

22. Yoshida K, Hernandez-Diaz S, Solomon DH, *et al.* Matching Weights to Simultaneously Compare Three Treatment Groups: Comparison to Three-way Matching. *Epidemiology* 2017; **28**: 387–395. doi:10.1097/EDE.0000000000000627.

23. Toh S, Hernán MA. Causal inference from longitudinal studies with baseline randomization. *Int J Biostat* 2008; **4**: Article 22. doi:10.2202/1557-4679.1117.

24. Greendale GA, Sowers M, Han W, *et al.* Bone mineral density loss in relation to the final menstrual period in a multiethnic cohort: results from the Study of Women's Health Across the Nation (SWAN). *J Bone Miner Res* 2012; **27**: 111–118. doi:10.1002/jbmr.534.

25. Solomon DH, Diem SJ, Ruppert K, *et al.* Bone mineral density changes among women initiating proton pump inhibitors or H2 receptor antagonists: a SWAN cohort study. *J Bone Miner Res* 2015; **30**: 232–239. doi:10.1002/jbmr.2344.

26. Solomon DH, Ruppert K, Zhao Z, *et al.* Bone mineral density changes among women initiating blood pressure lowering drugs: a SWAN cohort study. *Osteoporos Int* 2015. doi:10.1007/s00198-015-3332-6.

27. Black DM, Cummings SR, Karpf DB, *et al.* Randomised trial of effect of alendronate on risk of fracture in women with existing vertebral fractures. Fracture Intervention Trial Research Group. *Lancet* 1996; **348**: 1535–1541.

28. Black DM, Delmas PD, Eastell R, *et al.* Once-yearly zoledronic acid for treatment of postmenopausal osteoporosis. *N Engl J Med* 2007; **356**: 1809–1822. doi:10.1056/NEJMoa067312.

29. Cummings SR, San Martin J, McClung MR, *et al.* Denosumab for prevention of fractures in postmenopausal women with osteoporosis. *N Engl J Med* 2009; **361**: 756–765. doi:10.1056/NEJMoa0809493.

30. Neer RM, Arnaud CD, Zanchetta JR, *et al.* Effect of parathyroid hormone (1-34) on fractures and bone mineral density in postmenopausal women with osteoporosis. *N Engl J Med* 2001; **344**: 1434–1441.

doi:10.1056/NEJM200105103441904.

31. Baecke JA, Burema J, Frijters JE. A short questionnaire for the measurement of habitual physical activity in epidemiological studies. *Am J Clin Nutr* 1982; **36**: 936–942.

32. Sternfeld B, Ainsworth BE, Quesenberry CP. Physical activity patterns in a diverse population of women. *Prev Med* 1999; **28**: 313–323. doi:10.1006/pmed.1998.0470.

33. Ware JE, Sherbourne CD. The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. *Med Care* 1992; **30**: 473–483.

34. Austin PC. An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. *Multivariate Behav Res* 2011; **46**: 399–424. doi:10.1080/00273171.2011.568786.

35. Buuren S van, Groothuis-Oudshoorn K. mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software* 2011; **45**: 1–67.

36. Schafer JL, Yucel RM. Computational Strategies for Multivariate Linear Mixed-Effects Models With Missing Values. *Journal of Computational and Graphical Statistics* 2002; **11**: 437–457. doi:10.1198/106186002760180608.

37. Imbens GW. The role of the propensity score in estimating dose-response functions. *Biometrika* 2000; **87**: 706–710. doi:10.1093/biomet/87.3.706.

38. Xu S, Ross C, Raebel MA, Shetterly S, Blanchette C, Smith D. Use of stabilized inverse propensity scores as weights to directly estimate relative risk and its confidence intervals. *Value Health* 2010; **13**: 273–277. doi:10.1111/j.1524-4733.2009.00671.x.

39. Hernán MA, Hernández-Díaz S. Beyond the intention-to-treat in comparative effectiveness research. *Clin Trials* 2012; **9**: 48–55. doi:10.1177/1740774511420743.

40. Rassen JA, Shelat AA, Franklin JM, Glynn RJ, Solomon DH, Schneeweiss S. Matching by propensity score in cohort studies with three treatment groups. *Epidemiology* 2013; **24**: 401–409. doi:10.1097/EDE.0b013e318289dedf.

41. Robins JM, Hernán MA, Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology* 2000; **11**: 550–560.

42. Chou R, Fanciullo GJ, Fine PG, *et al.* Clinical guidelines for the use of chronic opioid therapy in chronic noncancer pain. *J Pain* 2009; **10**: 113–130. doi:10.1016/j.jpain.2008.10.008.

43. Yoshida K, Solomon DH, Kim SC. Active-comparator design and new-user design in observational studies. *Nat Rev Rheumatol* 2015; **11**: 437–441. doi:10.1038/nrrheum.2015.30.

44. Walker AM. Confounding by indication. *Epidemiology* 1996; **7**: 335–336.

45. Hernán MA, Sauer BC, Hernández-Díaz S, Platt R, Shrier I. Specifying a target trial prevents immortal time bias and other self-inflicted injuries in observational analyses. *J Clin Epidemiol* 2016. doi:10.1016/j.jclinepi.2016.04.014.

46. Hernán MA, Hernández-Díaz S, Robins JM. A structural approach to selection bias. *Epidemiology* 2004; **15**: 615–625.

47. Franklin GM, American Academy of Neurology. Opioids for chronic noncancer pain: a position paper of the American Academy of Neurology. *Neurology* 2014; **83**: 1277–1284. doi:10.1212/WNL.0000000000000839.

48. Berthelot J-M, Darrieutort-Lafitte C, Le Goff B, Maugars Y. Strong opioids for noncancer pain due to musculoskeletal diseases: Not more effective than acetaminophen or NSAIDs. *Joint Bone Spine* 2015; **82**: 397–401. doi:10.1016/j.jbspin.2015.08.003.

Table 2-1. Baseline characteristics of analgesics new users *before* propensity score weighting.

	APAP users	NSAID users	Opioid users	SMD
N	71	659	84	
Age (mean (SD))	49.34 (4.33)	49.43 (4.02)	50.63 (4.42)	0.200
Ethnicity (%)				0.493
Caucasian	33 (46.5)	385 (58.4)	40 (47.6)	
Black	25 (35.2)	191 (29.0)	43 (51.2)	
Asian	13 (18.3)	83 (12.6)	1 (1.2)	
Income (%)				0.383
Low (-19k)	9 (15.5)	46 (7.9)	13 (20.0)	
Middle (20k-49k)	16 (27.6)	203 (34.7)	28 (43.1)	
High (50k-)	33 (56.9)	336 (57.4)	24 (36.9)	
College education (%)	26 (36.6)	302 (46.2)	23 (27.7)	0.260
BMI (mean (SD))	28.43 (8.23)	29.16 (7.19)	32.43 (7.25)	0.354
Physical activity [3-15] (mean (SD))	7.49 (1.51)	7.87 (1.65)	7.08 (2.08)	0.295
Vasomotor symptoms (%)	34 (49.3)	341 (52.2)	47 (56.0)	0.089
Menopause transition stage (%)				0.318
Pre/Early Peri	45 (63.4)	458 (69.5)	41 (48.8)	
Late Peri	3 (4.2)	43 (6.5)	6 (7.1)	
Post	16 (22.5)	108 (16.4)	26 (31.0)	
Unknown	7 (9.9)	50 (7.6)	11 (13.1)	
Lumbar spine BMD g/cm2 (mean (SD))	1.04 (0.14)	1.08 (0.15)	1.11 (0.16)	0.291
Femoral neck BMD g/cm2 (mean (SD))	0.81 (0.13)	0.85 (0.14)	0.88 (0.14)	0.340
Pain-related QoL [0-100] (mean (SD))	70.44 (18.26)	69.75 (19.74)	48.77 (25.35)	0.647
Overall perception of health (%)				0.386
Same	28 (42.4)	270 (43.1)	30 (40.0)	
Better	31 (47.0)	297 (47.4)	22 (29.3)	
Worse	7 (10.6)	60 (9.6)	23 (30.7)	
Alcohol (%)				0.177
None/Low	34 (56.7)	285 (47.1)	34 (50.7)	
Moderate	16 (26.7)	168 (27.8)	21 (31.3)	
High	10 (16.7)	152 (25.1)	12 (17.9)	
Current smoker (%)	14 (20.3)	96 (14.7)	21 (25.0)	0.173
Thyroid disease (%)	9 (13.4)	66 (10.1)	10 (11.9)	0.069
Diabetes (%)	6 (8.5)	34 (5.2)	16 (19.0)	0.292
Calcium supplement (%)	21 (29.6)	216 (32.8)	18 (21.4)	0.171
Vitamin D supplement (%)	15 (21.1)	105 (15.9)	8 (9.5)	0.218
Oral glucocorticoids (%)	3 (4.2)	14 (2.1)	2 (2.4)	0.080

Missing proportions: BMI 7%; Income 13%; College education 1%; Physical activity 52% (not measured at every visit by design); Vasomotor symptoms 1%; BMD 10%; Pain-related QoL 15%; Alcohol 11%; Smoking 1%; Cancers 1%; Thyroid disease 1%. **Abbreviations**: APAP: acetaminophen; NSAID: non-steroidal anti-inflammatory drug; SMD: standardized mean difference; BMI: body mass index; Menopausal status: menopausal status define by menstrual cycles (See **eTable 2-2**); BMD: bone mineral density in g/cm^2; QoL: quality of life.

Table 2-2. Baseline characteristics of analgesics new users *after* propensity score weighting.

	APAP users	NSAID users	Opioid users	SMD
Age (mean (SD))	49.25 (3.72)	49.66 (4.23)	49.73 (4.27)	0.086
Ethnicity (%)				0.103
Caucasian	53.9	54.0	50.6	
Black	41.7	43.4	46.9	
Asian	4.4	2.6	2.5	
Income (%)				0.121
Low (-19k)	16.3	19.4	20.2	
Middle (20k-49k)	33.7	34.2	37.6	
High (50k-)	50.0	46.5	42.2	
College education (%)	37.0	32.0	31.1	0.085
BMI (mean (SD))	30.79 (7.87)	30.56 (7.35)	30.66 (6.54)	0.031
Physical activity [3-15] (mean (SD))	7.25 (1.63)	7.26 (1.67)	7.45 (1.92)	0.089
Vasomotor symptoms (%)	46.7	54.6	55.7	0.125
Menopause transition stage (%)				0.155
Pre/Early Peri	61.8	58.8	57.1	
Late Peri	7.2	5.3	9.2	
Post	21.5	22.0	21.2	
Unknown	9.5	13.9	12.6	
Lumbar spine BMD g/cm2 (mean (SD))	1.08 (0.14)	1.07 (0.15)	1.08 (0.14)	0.066
Femoral neck BMD g/cm2 (mean (SD))	0.85 (0.13)	0.84 (0.13)	0.86 (0.12)	0.074
Pain-related QoL [0-100] (mean (SD))	65.17 (17.45)	62.66 (19.84)	63.84 (20.99)	0.088
Overall perception of health (%)				0.092
Same	40.4	44.8	39.7	
Better	42.7	38.5	43.8	
Worse	16.9	16.7	16.5	
Alcohol (%)				0.074
None/Low	52.0	50.0	49.2	
Moderate	32.8	32.2	34.0	
High	15.3	17.8	16.8	
Current smoker (%)	19.8	26.1	23.9	0.102
Thyroid disease (%)	10.2	14.6	12.4	0.091
Diabetes (%)	12.8	13.4	12.2	0.030
Calcium supplement (%)	16.5	22.6	21.8	0.106
Vitamin D supplement (%)	8.4	12.7	12.6	0.098
Oral glucocorticoids (%)	2.9	2.3	1.3	0.079

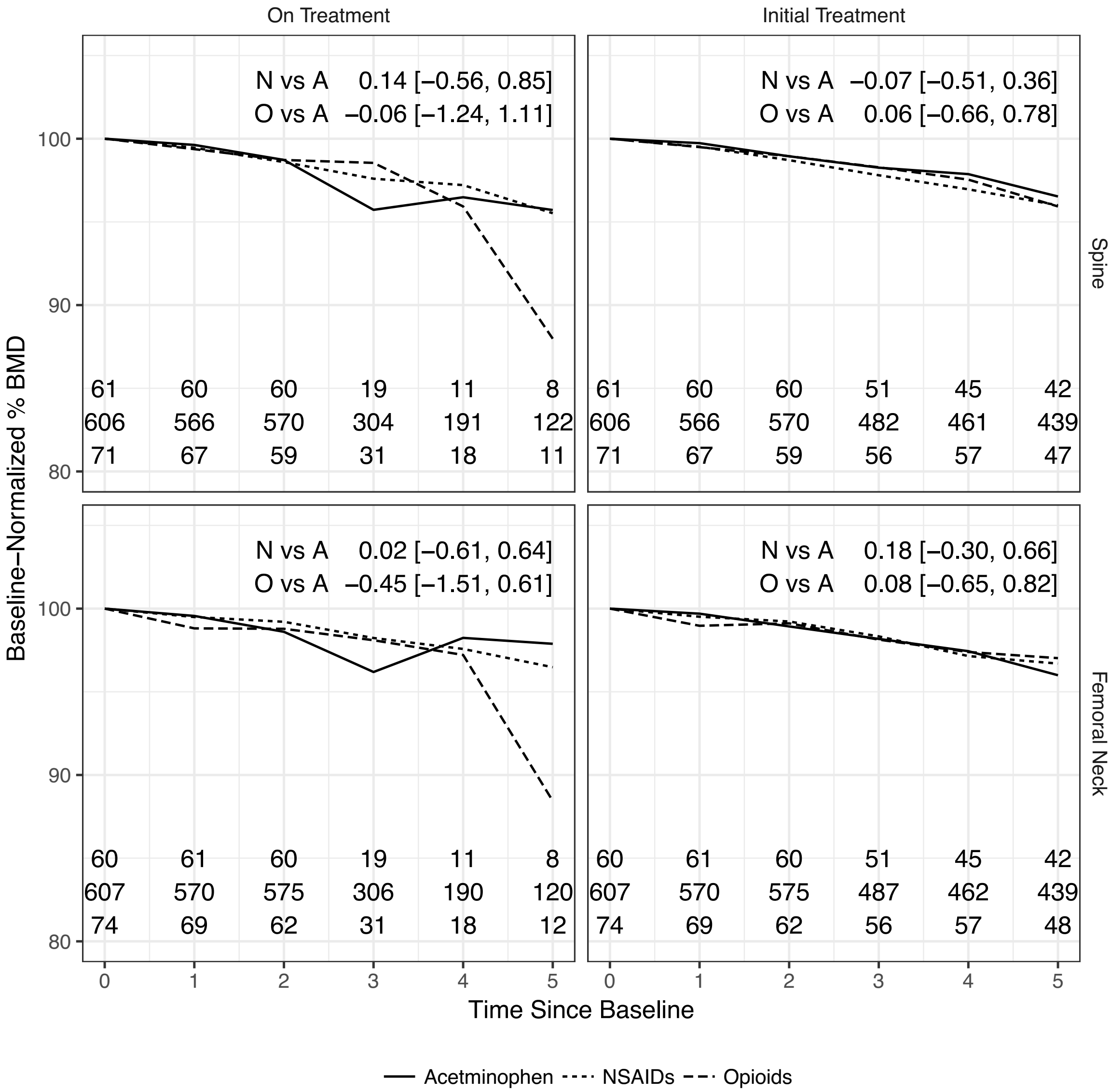
Abbreviations: APAP: acetaminophen; NSAID: non-steroidal anti-inflammatory drug; SMD: standardized mean difference; BMI: body mass index; Menopausal status: menopausal status define by menstrual cycles (See **eTable 2-2**); BMD: bone mineral density in g/cm^2; QoL: quality of life.

Table 2-3. Main bone mineral density analysis results from generalized estimating equations.

Analysis Type	Site	Group	Mean Annual Change (%)	Group Difference (%)
On Treatment	Spine	Acetaminophen	-0.90 [-1.58, -0.21]	Ref.
		NSAIDs	-0.76 [-0.92, -0.59]	0.14 [-0.56, 0.85]
		Opioids	-0.96 [-1.92, -0.00]	-0.06 [-1.24, 1.11]
	Femoral Neck	Acetaminophen	-0.61 [-1.21, -0.02]	Ref.
		NSAIDs	-0.60 [-0.81, -0.39]	0.02 [-0.61, 0.64]
		Opioids	-1.07 [-1.95, -0.19]	-0.45 [-1.51, 0.61]
Initial Treatment	Spine	Acetaminophen	-0.72 [-1.13, -0.30]	Ref.
		NSAIDs	-0.79 [-0.93, -0.66]	-0.07 [-0.51, 0.36]
		Opioids	-0.66 [-1.25, -0.07]	0.06 [-0.66, 0.78]
	Femoral Neck	Acetaminophen	-0.82 [-1.28, -0.36]	Ref.
		NSAIDs	-0.64 [-0.80, -0.49]	0.18 [-0.30, 0.66]
		Opioids	-0.74 [-1.31, -0.16]	0.08 [-0.65, 0.82]

On-treatment analysis censored patients at the time they changed analgesic categories, whereas initial treatment group analysis retained these patients in the initial treatment groups.
Abbreviations: NSAID: non-steroidal anti-inflammatory drug; Ref.: Reference.

Figure 2-1. Group mean trajectories of baseline-normalized % bone mineral density (BMD).



The numbers at the bottom of each panel are number of individuals contributing BMD measurements (Top: Acetaminophen, Middle: NSAIDs, and bottom: Opioids). On-treatment analysis censored patients at the time they changed analgesic categories, whereas initial treatment group analysis retained these patients in the initial treatment groups.

Abbreviations: NSAID: non-steroidal anti-inflammatory drug; N vs A: NSAID group vs Acetaminophen group; O vs A: Opioid group vs Acetaminophen group; Spine: lumbar spine BMD; Time Since Baseline: Time since the baseline visit in years.

eTable 2-1. Hierarchical definition of exposure category.

Acetaminophen use	NSAID use	Opioid use	Exposure Category
+	-	-	Acetaminophen user
±	+	-	NSAID user
±	±	+	Opioid user

Abbreviations: (+): use; (-): no use; (±): either use or no use.

eTable 2-2. Menopause transition stage definition assessed at the individual-specific baseline visit.

Status	Original Status	Definition
Pre/Early Peri	Pre-menopause	No change in menstrual cycles.
	Early peri-menopause	More irregular menstrual cycles than prior, but no gaps in cycles of > 3 months.
Late Peri	Late peri-menopause	No menstrual cycles for 3-11 months.
Post	Post-menopause	No menstrual cycles for 12 or more months.
	Post-BSO	Post-bilateral salpingo-oophorectomy (ovary removal).
Unknown	Post-hysterectomy	Hysterectomy without bilateral oophorectomy prior to the FMP.
(Excluded)	HT before FMP	HT use prior to FMP obscuring classification. Excluded from the current study.
	Pregnant or breastfeeding	Menstrual cycles are obscured by pregnancy or breastfeeding. Excluded from the current study.

We used the classification under "Status" in our analysis. The "Original Status" represents the coding in the SWAN database.

Abbreviations: BSO: bilateral salpingo-oophorectomy; FMP: final menstrual period; HT: hormone therapy for menopause.

eTable 2-3. Final menstrual period-based menopause transition stage definition assessed at the individual-specific baseline visit.

Status	Definition
Pretransmenopause	More than 1 year before FMP
Transmenopause	From 1 year before FMP to 2 years after FMP
Postmenopause	More than two years after FMP

Abbreviations: FMP: Final menstrual period.

eTable 2-4. Break down of NSAIDs and opioids by generic names used by the subjects in the main analysis.

	NSAIDs			Opioids
n	659		n	84
Generic name (%)			Generic name (%)	
Ibuprofen*	484 (73.4)		Codeine	30 (35.7)
Naproxen	113 (17.1)		Propoxyphene	17 (20.2)
Celecoxib†	16 (2.4)		Tramadol	15 (17.9)
Nabumetone	11 (1.7)		Oxycodone	12 (14.3)
Rofecoxib†	12 (1.8)		Meperidine	3 (3.6)
Etodolac	7 (1.1)		Fentanyl	2 (2.4)
Diclofenac	4 (0.6)		Methadone	2 (2.4)
Oxaprozin	3 (0.5)		Morphine	2 (2.4)
Indomethacin	2 (0.3)		Hydromorphone	1 (1.2)
Ketorolac	2 (0.3)			
Piroxicam	2 (0.3)			
Meclofenamate	1 (0.2)			
Mefenamic acid	1 (0.2)			
Sulindac	1 (0.2)			

* Includes one user who used both ibuprofen and naproxen; † COX-2 selective agents.
Abbreviations: NSAID: non-steroidal anti-inflammatory drug; COX-2: Prostaglandin-endoperoxide synthase 2

eTable 2-5. Mean follow up duration in years in each analysis.

Cohort	Analysis Type	Acetaminophen	NSAIDs	Opioids
Main analysis	On Treatment	2.6	3.0	2.8
	Initial Treatment	4.3	4.4	4.2
FMP subgroup	On Treatment	2.8	3.1	2.6
	Initial Treatment	4.4	4.7	4.0
Postmenopausal subgroup	On Treatment	3.0	2.8	2.9
	Initial Treatment	3.9	4.0	3.5

On-treatment analysis censored patients at the time they changed analgesic categories, whereas initial treatment group analysis retained these patients in the initial treatment groups.
Abbreviations: FMP: final menstrual period; NSAID: non-steroidal anti-inflammatory drug.

eTable 2-6. Baseline characteristics of the subgroup of analgesics new users with final menstrual period (FMP) data *before* propensity score weighting.

	APAP users	NSAID users	Opioid users	SMD
N	39	381	33	
Age (mean (SD))	50.27 (4.04)	49.78 (4.10)	51.56 (4.85)	0.268
Ethnicity (%)				0.605
Caucasian	15 (38.5)	210 (55.1)	15 (45.5)	
Black	15 (38.5)	120 (31.5)	18 (54.5)	
Asian	9 (23.1)	51 (13.4)	0 (0.0)	
Income (%)				0.467
Low (-19k)	7 (23.3)	25 (7.5)	5 (22.7)	
Middle (20k-49k)	7 (23.3)	121 (36.3)	9 (40.9)	
High (50k-)	16 (53.3)	187 (56.2)	8 (36.4)	
College education (%)	11 (28.2)	174 (46.3)	7 (21.2)	0.364
BMI (mean (SD))	27.60 (8.34)	29.72 (7.58)	33.43 (8.53)	0.472
Physical activity [3-15] (mean (SD))	7.32 (1.45)	7.81 (1.68)	7.31 (2.23)	0.190
Vasomotor symptoms (%)	20 (51.3)	189 (49.9)	12 (36.4)	0.203
FMP category (%)				0.420
Pretransmenopause	21 (53.8)	242 (63.5)	16 (48.5)	
Transmenopause	10 (25.6)	81 (21.3)	4 (12.1)	
Postmenopause	8 (20.5)	58 (15.2)	13 (39.4)	
Lumbar spine BMD g/cm2 (mean (SD))	1.04 (0.14)	1.08 (0.16)	1.10 (0.13)	0.303
Femoral neck BMD g/cm2 (mean (SD))	0.81 (0.15)	0.86 (0.14)	0.88 (0.13)	0.326
Pain-related QoL [0-100] (mean (SD))	69.22 (21.14)	69.93 (19.30)	49.38 (24.32)	0.614
Overall perception of health (%)				0.333
Same	16 (43.2)	155 (43.2)	13 (48.1)	
Better	16 (43.2)	175 (48.7)	8 (29.6)	
Worse	5 (13.5)	29 (8.1)	6 (22.2)	
Alcohol (%)				0.239
None/Low	19 (61.3)	171 (49.9)	13 (56.5)	
Moderate	8 (25.8)	88 (25.7)	7 (30.4)	
High	4 (12.9)	84 (24.5)	3 (13.0)	
Current smoker (%)	8 (20.5)	46 (12.1)	8 (24.2)	0.212
Thyroid disease (%)	4 (10.5)	28 (7.4)	2 (6.1)	0.109
Diabetes (%)	3 (7.7)	22 (5.8)	11 (33.3)	0.496
Calcium supplement (%)	13 (33.3)	116 (30.4)	4 (12.1)	0.348
Vitamin D supplement (%)	9 (23.1)	48 (12.6)	4 (12.1)	0.194
Oral glucocorticoids (%)	2 (5.1)	8 (2.1)	0 (0.0)	0.233

Abbreviations: APAP: acetaminophen; NSAID: non-steroidal anti-inflammatory drug; SMD: standardized mean difference; BMI: body mass index; FMP category: final menstrual period category (See **eTable 2-3**); BMD: bone mineral density in g/cm^2; QoL: quality of life.

eTable 2-7. Baseline characteristics of the subgroup of analgesics new users with final menstrual period (FMP) data *after* propensity score weighting.

	APAP users	NSAID users	Opioid users	SMD
Age (mean (SD))	49.14 (3.32)	50.49 (4.64)	49.55 (4.23)	0.219
Ethnicity (%)				0.105
Caucasian	52.4	48.9	45.3	
Black	47.6	51.1	54.7	
Asian	0.0	0.0	0.0	
Income (%)				0.142
Low (-19k)	26.3	28.4	29.3	
Middle (20k-49k)	27.7	30.6	30.1	
High (50k-)	45.9	41.0	40.6	
College education (%)	29.2	25.1	27.7	0.091
BMI (mean (SD))	30.89 (8.71)	30.15 (7.92)	30.35 (6.42)	0.081
Physical activity [3-15] (mean (SD))	7.05 (1.72)	7.18 (1.79)	7.18 (2.10)	0.104
Vasomotor symptoms (%)	35.4	47.1	54.0	0.259
FMP category (%)				0.279
Pretransmenopause	69.5	56.3	59.4	
Transmenopause	20.4	19.0	21.2	
Postmenopause	10.1	24.7	19.4	
Lumbar spine BMD g/cm2 (mean (SD))	1.09 (0.16)	1.07 (0.18)	1.11 (0.12)	0.156
Femoral neck BMD g/cm2 (mean (SD))	0.86 (0.17)	0.84 (0.15)	0.87 (0.14)	0.139
Pain-related QoL [0-100] (mean (SD))	66.18 (15.50)	60.02 (20.73)	60.08 (18.80)	0.254
Overall perception of health (%)				0.285
Same	52.9	44.5	35.1	
Better	31.7	32.9	37.1	
Worse	15.4	22.5	27.8	
Alcohol (%)				0.388
None/Low	64.9	53.6	63.7	
Moderate	33.3	30.0	27.5	
High	1.8	16.4	8.8	
Current smoker (%)	18.5	31.5	31.3	0.235
Thyroid disease (%)	5.9	8.7	7.2	0.094
Diabetes (%)	8.2	14.9	16.7	0.182
Calcium supplement (%)	19.9	19.0	24.9	0.106
Vitamin D supplement (%)	12.0	15.0	20.4	0.153
Oral glucocorticoids (%)	0.0	0.0	0.0	<0.001

Abbreviations: APAP: acetaminophen; NSAID: non-steroidal anti-inflammatory drug; SMD: standardized mean difference; BMI: body mass index; FMP category: final menstrual period category (See **eTable 2-3**); BMD: bone mineral density in g/cm^2; QoL: quality of life.

eTable 2-8. FMP-based bone mineral density analysis results from generalized estimating equations.

Analysis Type	Site	Group	Mean Annual Change (%)	Group Difference (%)
On Treatment	Spine	Acetaminophen	-0.97 [-1.95, -0.00]	Ref.
		NSAIDs	-0.99 [-1.28, -0.69]	-0.01 [-1.03, 1.01]
		Opioids	-0.67 [-1.46, 0.12]	0.30 [-0.95, 1.56]
	Femoral Neck	Acetaminophen	-0.95 [-1.88, -0.02]	Ref.
		NSAIDs	-0.62 [-0.97, -0.26]	0.33 [-0.66, 1.33]
		Opioids	-0.80 [-1.65, 0.06]	0.15 [-1.11, 1.42]
Initial Treatment	Spine	Acetaminophen	-0.58 [-1.39, 0.22]	Ref.
		NSAIDs	-1.01 [-1.23, -0.79]	-0.43 [-1.26, 0.41]
		Opioids	-0.59 [-1.46, 0.27]	-0.01 [-1.19, 1.17]
	Femoral Neck	Acetaminophen	-0.70 [-1.51, 0.12]	Ref.
		NSAIDs	-0.72 [-0.99, -0.45]	-0.02 [-0.88, 0.84]
		Opioids	-0.61 [-1.53, 0.30]	0.08 [-1.15, 1.31]

On-treatment analysis censored patients at the time they changed analgesic categories, whereas initial treatment group analysis retained these patients in the initial treatment groups.
Abbreviations: NSAID: non-steroidal anti-inflammatory drug; Ref.: Reference.

eTable 2-9. Bone mineral density sensitivity analysis results further adjusting for menopausal transition stage, income category, vasomotor symptoms, calcium supplement use, ethnicity, and smoking status.

Analysis Type	Site	Group	Group Difference (%)
On Treatment	Spine	Acetaminophen	Ref.
		NSAIDs	0.15 [-0.55, 0.86]
		Opioids	0.08 [-1.01, 1.17]
	Femoral Neck	Acetaminophen	Ref.
		NSAIDs	-0.02 [-0.66, 0.62]
		Opioids	-0.40 [-1.44, 0.64]
Initial Treatment	Spine	Acetaminophen	Ref.
		NSAIDs	-0.07 [-0.51, 0.36]
		Opioids	0.08 [-0.62, 0.79]
	Femoral Neck	Acetaminophen	Ref.
		NSAIDs	0.15 [-0.30, 0.61]
		Opioids	0.04 [-0.62, 0.70]

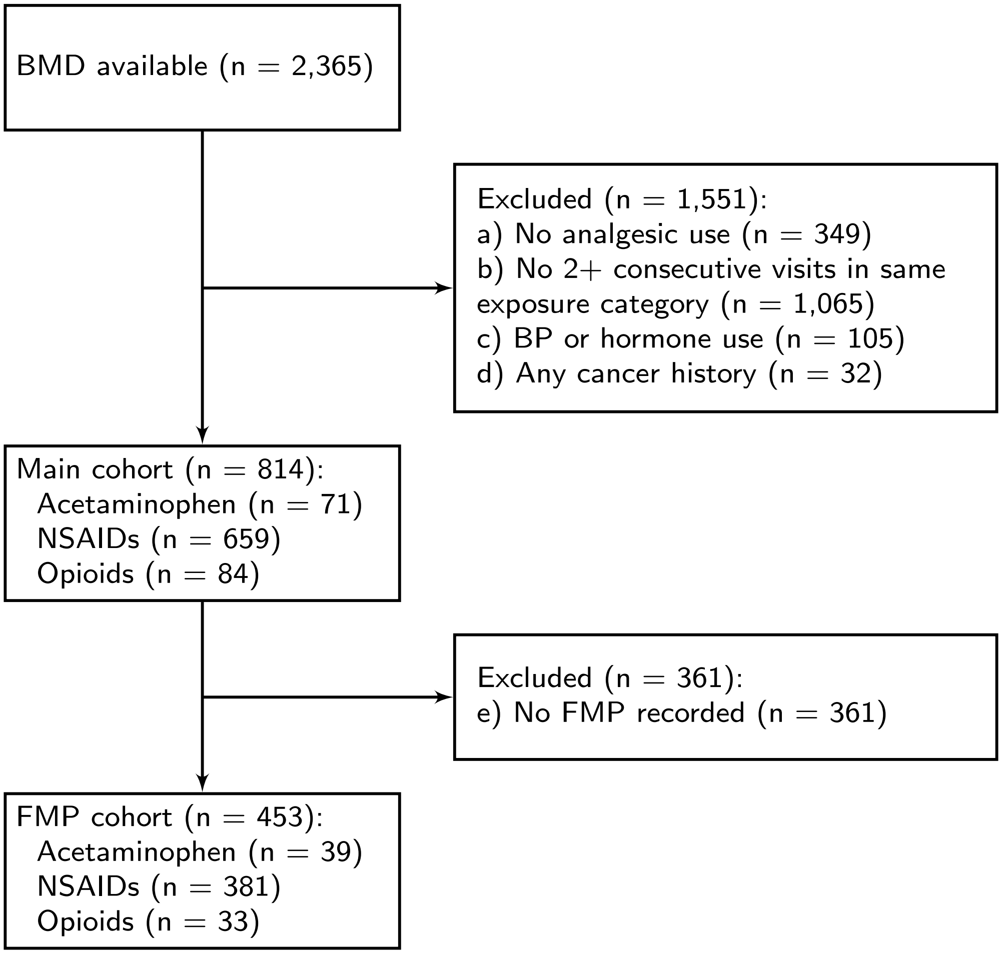
On-treatment analysis censored patients at the time they changed analgesic categories, whereas initial treatment group analysis retained these patients in the initial treatment groups. Adjustment was conducted in the outcome model for variables that had standardized mean differences > 0.1 after balancing by matching weights.
Abbreviations: NSAID: non-steroidal anti-inflammatory drug; Ref.: Reference.

eTable 2-10. Bone mineral density annual slope differences observed in key clinical trials.

Trial Name	FIT[27]	HORIZON[28]	FREEDOM[29]	FPT[30]
Intervention	Alendronate	IV Zolendronate	IV Denosumab	SC Teriparatide
Comparator	Placebo	Placebo	Placebo	Placebo
Follow up	3 years	3 years	3 years	1.5 years
Spine	+2.5% / year	+2.2% / year	+3.1% / year	+5.7% / year
Femoral Neck	+1.0% / year	+1.7% / year	N/A	+2.3% / year
Fracture Prevention	Spine, Hip	Spine, Hip	Spine, Hip	Spine

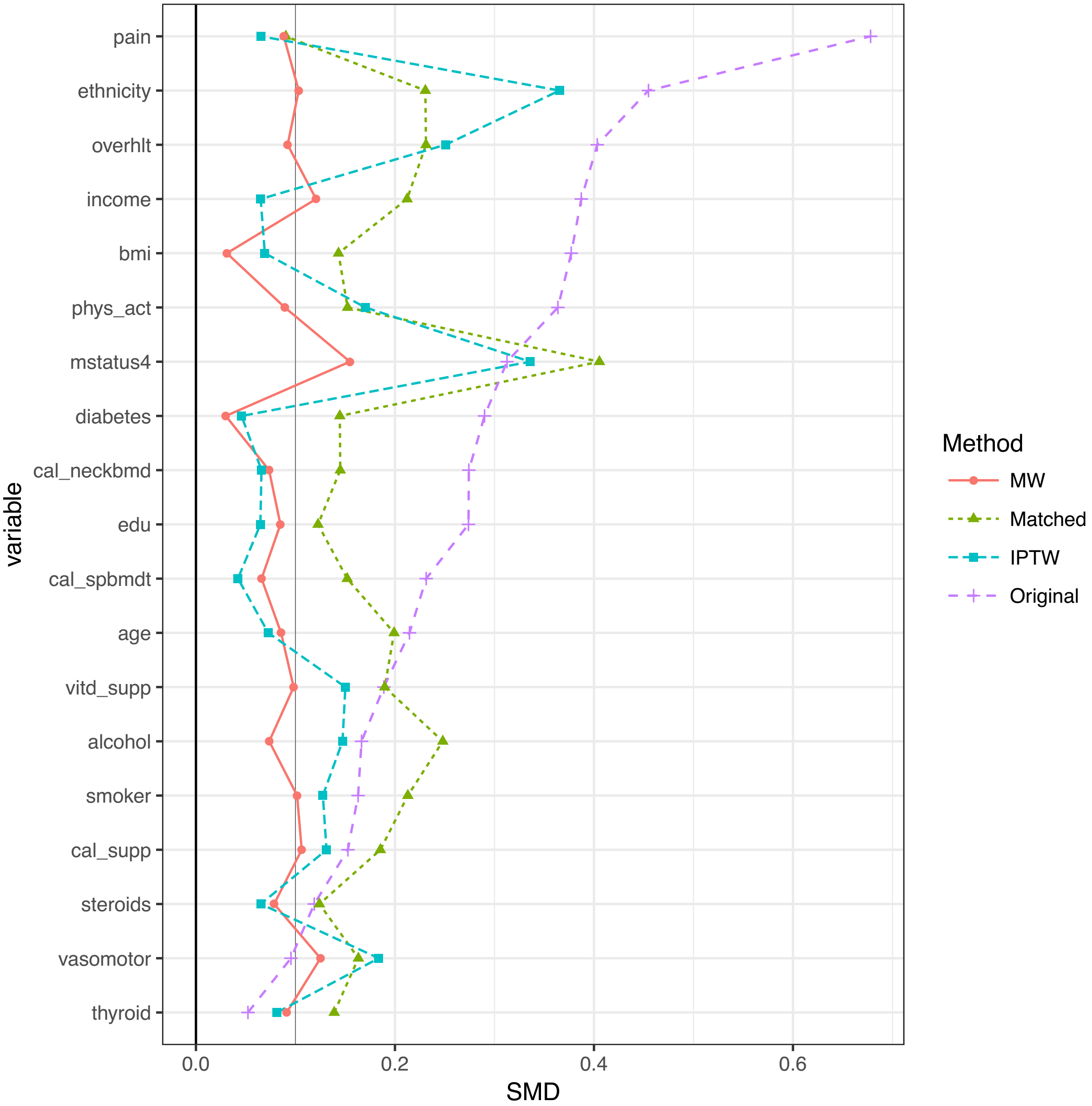
Abbreviations: N/A: not available.

eFigure 2-1. Derivation of the main study cohort and FMP cohort.


BMD available (n = 2,365)
Excluded (n = 1,551):
a) No analgesic use (n = 349)
b) No 2+ consecutive visits in same exposure category (n = 1,065)
c) BP or hormone use (n = 105)
d) Any cancer history (n = 32)
Main cohort (n = 814):
Acetaminophen (n = 71)
NSAIDs (n = 659)
Opioids (n = 84)
Excluded (n = 361):
e) No FMP recorded (n = 361)
FMP cohort (n = 453):
Acetaminophen (n = 39)
NSAIDs (n = 381)
Opioids (n = 33)

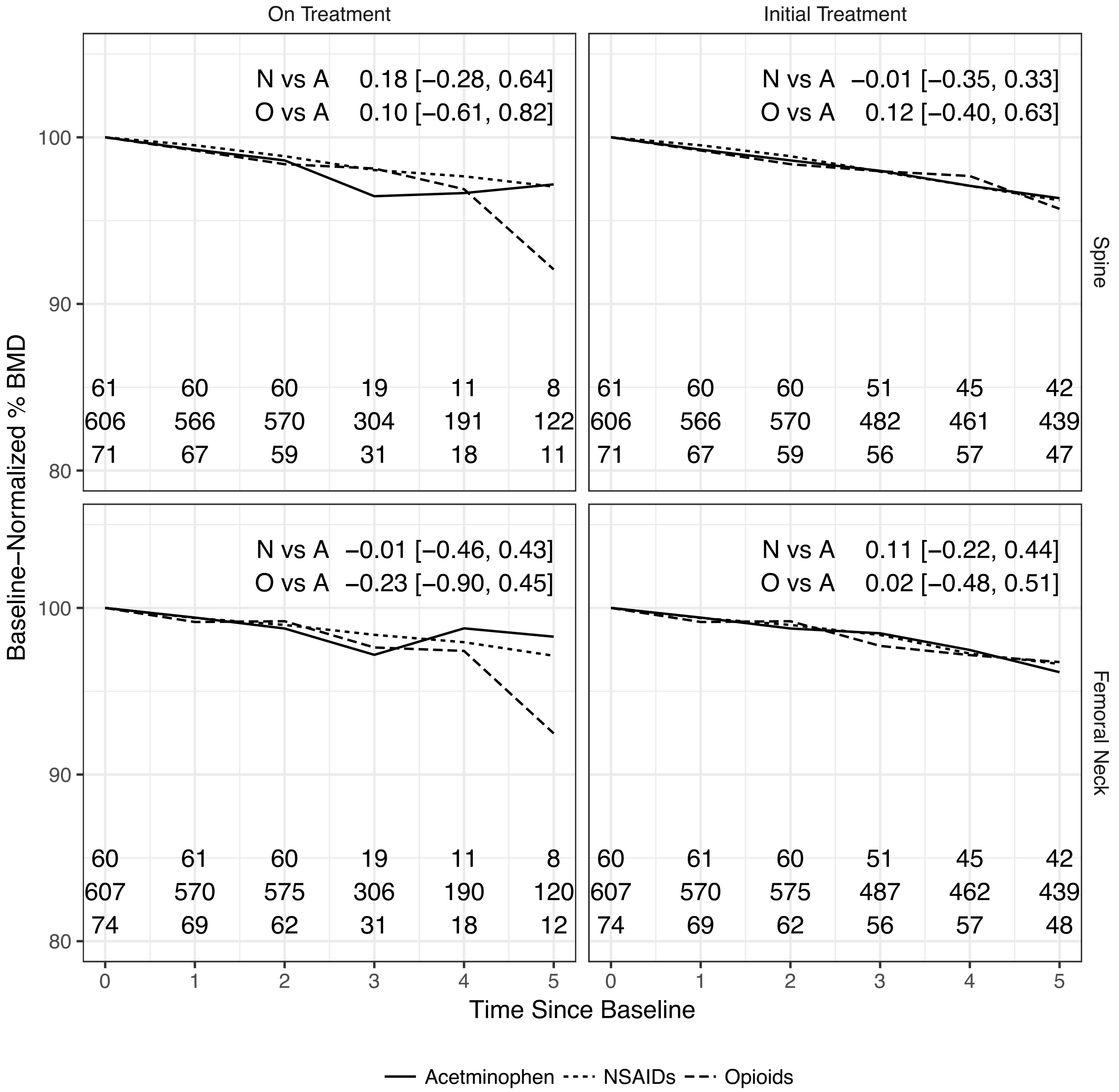
Abbreviations: BMD: bone mineral density; BP: bisphosphonate; NSAID: non-steroidal anti-inflammatory drug; FMP: final menstrual period.

eFigure 2-2. Standardized mean difference (SMD) before and after matching weights (MW).



Abbreviations: SMD: standardized mean difference; MW: covariate balance after matching weights; Matched: covariate balance after three-way matching; IPTW: covariate balance after inverse probability of treatment weights; Original: covariate balance before any adjustment.

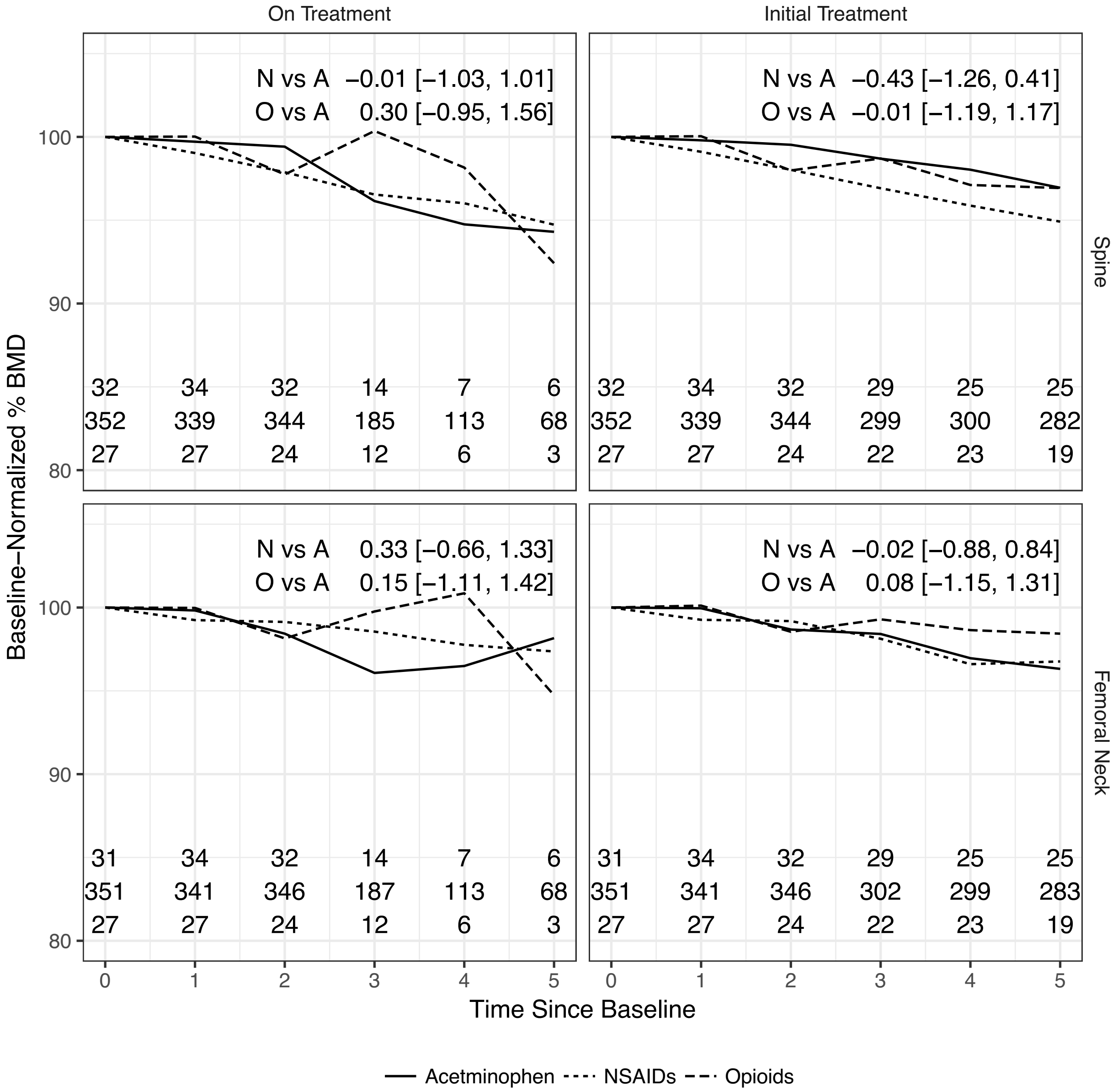
eFigure 2-3. Group mean trajectories of baseline-normalized % bone mineral density (BMD) in the main unadjusted analysis.



The numbers at the bottom of each panel are number of individuals contributing BMD measurements (Top: Acetaminophen, Middle: NSAIDs, and bottom: Opioids). On-treatment analysis censored patients at the time they changed analgesic categories, whereas initial treatment group analysis retained these patients in the initial treatment groups.

Abbreviations: NSAID: non-steroidal anti-inflammatory drug; N vs A: NSAID group vs Acetaminophen group; O vs A: Opioid group vs Acetaminophen group; Spine: lumbar spine BMD; Time Since Baseline: Time since the baseline visit in years.

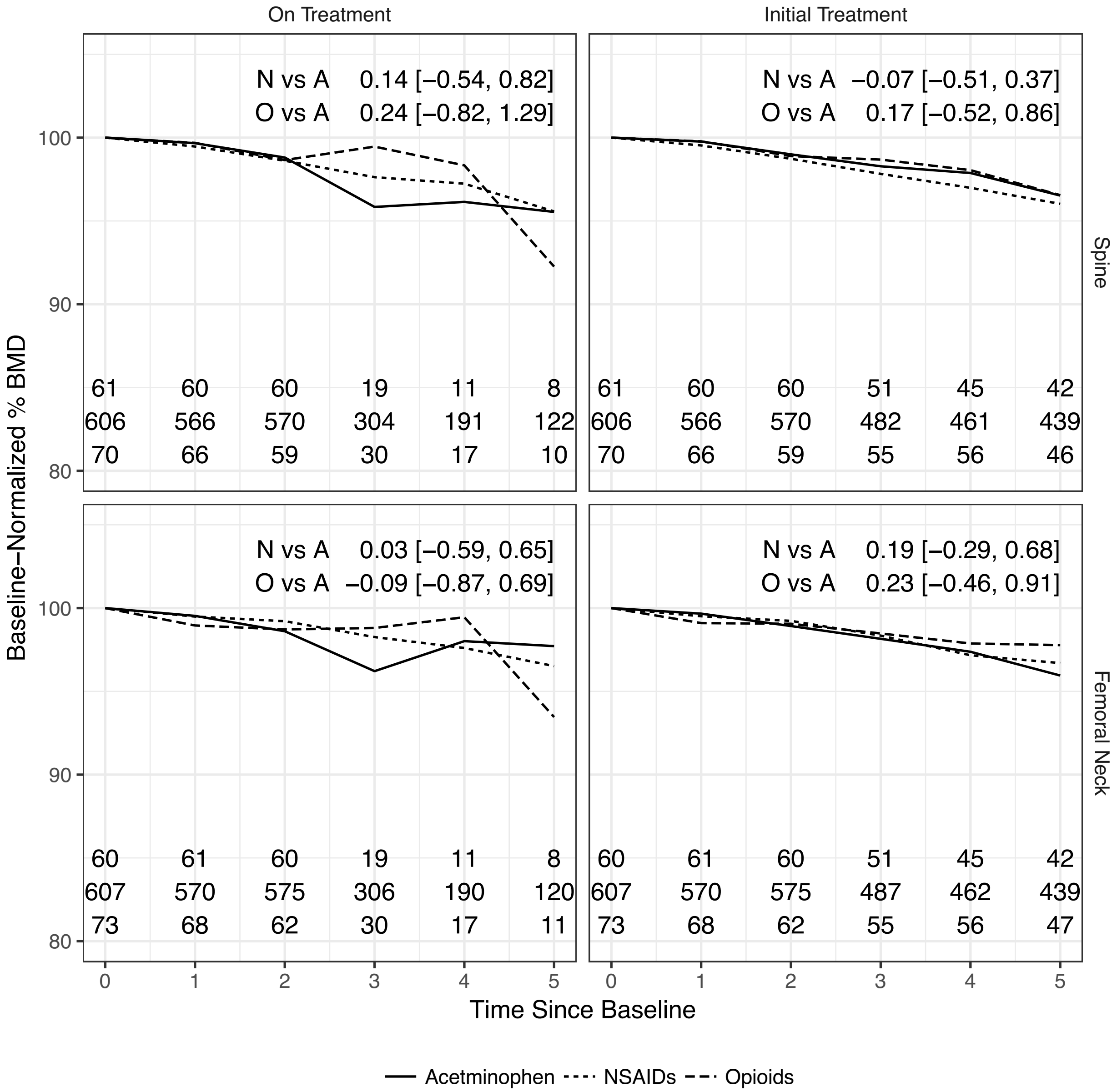
eFigure 2-4. Group mean trajectories of baseline-normalized % bone mineral density (BMD) in the final menstrual period (FMP) subgroup.



The numbers at the bottom of each panel are number of individuals contributing BMD measurements (Top: Acetaminophen, Middle: NSAIDs, and bottom: Opioids). On-treatment analysis censored patients at the time they changed analgesic categories, whereas initial treatment group analysis retained these patients in the initial treatment groups.

Abbreviations: NSAID: non-steroidal anti-inflammatory drug; N vs A: NSAID group vs Acetaminophen group; O vs A: Opioid group vs Acetaminophen group; Spine: lumbar spine BMD; Time Since Baseline: Time since the baseline visit in years.

eFigure 2-5. Group mean trajectories of baseline-normalized % bone mineral density (BMD) after excluding an outlying opioid user.



The numbers at the bottom of each panel are number of individuals contributing BMD measurements (Top: Acetaminophen, Middle: NSAIDs, and bottom: Opioids). On-treatment analysis censored patients at the time they changed analgesic categories, whereas initial treatment group analysis retained these patients in the initial treatment groups.

Abbreviations: NSAID: non-steroidal anti-inflammatory drug; N vs A: NSAID group vs Acetaminophen group; O vs A: Opioid group vs Acetaminophen group; Spine: lumbar spine BMD; Time Since Baseline: Time since the baseline visit in years.

1 Details of matching weights

When there are K treatment categories, K propensity scores are defined. That is, individual i has a predicted probability of receiving treatment for each one of the K treatment categories. These scores are often called the generalized propensity scores. The scores need to add up to one, thus, estimation is usually conducted with multinomial logistic regression. Using these generalized propensity scores and treatment categories, the weights are defined as follows.

$$\text{Matching Weight}_i = \frac{\min(e_{1i}, ..., e_{Ki})}{\sum_{k=1}^{K} I(Z_i = k) e_{ki}}$$

where e_{ki} is the generalized propensity score for the k-th treatment (*i.e.*, probability of receiving the k-th treatment), $Z_i \in \{1, ..., K\}$ is a categorical treatment variable, and $I(\cdot)$ is an inductor variable (1 if true and 0 if false).

2 Details of outcome modeling

2.1 Procedure

Generalized estimating equation was used via `geepack R` package (`geeglm()` function) with the Gaussian error structure and identity link (`family = gaussian(link = "identity")`).

2.2 Model

The outcome variable used was the baseline-normalized %BMD. This variable was created within each individual by dividing the BMD values by the year 0 BMD value and multiplying by 100. Therefore, the outcome started at 100% at year 0 for all individuals regardless of the treatment group. This is the outcome depicted in figures such as Figure 1.

The mean model formula for the outcome for individual i at time point j (Y_{ij}) was the following.

$$E[Y_{ij} - 100|\text{COVARIATES}_{ij}] = \beta_1 \text{YEAR}_{ij} + \beta_2 \text{NSAID}_i \times \text{YEAR}_{ij} + \beta_3 \text{OPIOID}_i \times \text{YEAR}_{ij}$$

Here the outcome was further modified by subtracting 100 to ease the modeling process because this "intercept" was common to all individuals (thus, no need to estimate). The explanatory variables were the year term (time effect), NSAID group indicator-year interaction term, and opioid group indicator-year interaction term. We only need the interaction terms for the NSAID and opioid group indicator variables, but no the main effect terms for the group indicator variables, because all individuals in all treatment groups have the same outcome value at year 0 ($Y_{ij} - 100 = 0$ for all individuals).

As a result, the estimates shown in **Table 3** are estimates for the following coefficients or sum of coefficients. β_1 represents the mean annual change (slope on the %BMD) in the acetaminophen group. To obtain the mean annual changes (%) for the NSAIDs group and opioids group, the respective group difference coefficients (interaction coefficients; slope differences) were added to the acetaminophen (reference) group mean annual change (%).

Group	Mean Annual Change (%)	Group Difference
Acetaminophen	β_1	Reference
NSAIDs	$\beta_1 + \beta_2$	β_2
Opioids	$\beta_1 + \beta_3$	β_3

Chapter 3: A Tool for Empirical Equipoise Assessment in Multi-group Comparative Effectiveness Research

AUTHORS: Kazuki Yoshida(1,2), Daniel H. Solomon(3,4), Sebastien Haneuse(2), Seoyoung C. Kim(3,4), Elisabetta Patorno(4), Sara K. Tedeshi(3), Houchen Lyu(3), Sonia Hernandez-Diaz(1), Robert J. Glynn(2, 4)

AFFILIATIONS

1. Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, United States.

2. Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, United States.

3. Division of Rheumatology, Immunology and Allergy, Brigham and Women's Hospital, Boston, Massachusetts, United States.

4. Division of Pharmacoepidemiology and Pharmacoeconomics, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts, United States.

DISCLOSURES:
KY received financial support for his doctoral study from the Pharmacoepidemiology Program at the Harvard T.H. Chan School of Public Health (partially supported by training grants from Pfizer, Takeda, Bayer and ASISA) and Honjo International Scholarship Foundation. DHS receives salary support from institutional research grants from Amgen, Abbvie, Pfizer, Genentech, Bristol Myers Squibb, and Corrona. He also receives royalties from UpToDate. SCK has received research grants to the Brigham and Women's Hospital from Pfizer, Roche and Bristol-Myers Squibb for unrelated studies. EP is investigator of investigator-initiated grants to the Brigham and Women's Hospital from GSK and Boehringer Ingelheim, not directly related to the topic of the submitted work. SHD received research support in the form of grants to her institution for pregnancy safety studies from Eli Lilly and GSK. She also consulted for Roche. RJG received research support in the form of grants to his institution for clinical trial design, monitoring, and analysis from AstraZeneca, Kowa, Novartis, and Pfizer.

FUNDING INFORMATION:
EP is supported by a career development grant K08AG055670 from the National Institute on Aging. DHS receives salary support from NIH-K24AR055989 and NIH-P30AR072577.

ABSTRACT

PURPOSE: In observational research, equipoise concerns whether groups being compared are *similar enough* for valid inference. *Empirical equipoise* was previously proposed as a tool to assess patient similarity based on propensity scores (PS). We extended this work for multi-group observational studies.

METHODS: We modified the tool to allow for multinomial exposures such that the proposed definition reduces to the original when there are only two groups. We illustrated how the tool can be used as method to assess study design within three-group clinical examples. We then conducted three-group simulations to assess how the tool performed in a setting with residual confounding after PS weighting.

RESULTS: In a clinical example based on rheumatoid arthritis, 44.5% of the sample fell within the region of empirical equipoise when considering first-line biologics, whereas 57.7% did so for second-line biologics, consistent with the expectation that a second-line design results in better equipoise. In a simulation where the unmeasured confounder had the same magnitude of association with the treatment as the measured confounders and a 25% greater association with the outcome, the tool crossed the proposed threshold for empirical equipoise at a residual confounding of 20% on the ratio scale. When the unmeasured variable had a twice larger association with treatment, the tool became less sensitive and crossed the threshold at a residual confounding of 30%.

CONCLUSION: Our proposed tool may be useful in guiding cohort identification in multi-group observational studies, particularly with similar effects of unmeasured and measured covariates on treatment and outcome.

INTRODUCTION

Pharmacoepidemiologists are often concerned with whether the exposure groups in an observational study are *similar enough* for unbiased causal inference. Lack of similarity can imply dangers of positivity violation[1] and residual confounding from imperfectly measured and unmeasured variables. Statistical analyses alone cannot fully address these issues and design stage efforts[2], such as the active comparator design[3,4], are necessary. However, no well-accepted measure exists for deciding whether groups are *similar enough*, particularly in comparisons among three or more treatments.

Walker *et al*. introduced the concept of *empirical equipoise*[5] in the setting of two-group comparative effectiveness research (CER). Empirical equipoise is a manifestation of underlying *clinical equipoise*[6]: a state of collective uncertainty among medical providers regarding the best treatment option for a specific patient population. In this circumstance, prescriber opinions, rather than patient characteristics, largely determine treatment choices[5]. A treatment assignment mechanism that is mostly independent of patient characteristics results in treatment groups that are similar and overlapping in covariates.

Since clinical equipoise pertains to prescriber opinions, it is not directly measurable in typical CER datasets such as administrative claims. Empirical equipoise is a measure of similarity of the distributions of potential confounders available in CER datasets and can be useful as a study design assessment tool[7]. To our knowledge, no such tool exists for studies with three or more groups even though multi-group CER is increasingly relevant due to the development of many treatment options for rheumatoid arthritis (RA)[8], diabetes mellitus[9], and atrial fibrillation[10] to name a few. In this paper, we provide a detailed explanation of Walker *et* *al.*'s empirical equipoise tool, propose an extension to the multi-group CER setting, illustrate its face validity in empirical data, and examine its performance in simulations.

METHODS

Empirical equipoise assessment tool

Consider a two-group CER study. Let A_i be an indicator of the binary treatment for the i-th study participant, $\mathbf{X}_i$ a vector of potential confounders, and consider the following logistic model for the propensity score (PS), denoted e_i:

$$\log\left(\frac{e_i}{1-e_i}\right) = \text{logit}(E[A_i|\mathbf{X}_i]) = \alpha_0 + \mathbf{X}_i^T\boldsymbol{\alpha}_{\boldsymbol{X}}$$

Walker *et al.* proposed a prevalence-adjusted version of PS, the *preference score,* denoted π_i defined by:

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \log\left(\frac{e_i}{1-e_i}\right) - \log\left(\frac{p}{1-p}\right)$$

where p is the marginal prevalence of treatment. The second term has the same form as the intercept adjustment for risk prediction from case-control data.[11–13] Given this, the model for the preference score can re-written as:

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \left[\alpha_0 - \log\left(\frac{p}{1-p}\right)\right] + \mathbf{X}_i^T\boldsymbol{\alpha}_{\boldsymbol{X}}$$

Thus, the preference score considers treatment assignment in a hypothetical population with a treatment prevalence of 50% but for which the covariate effect on assignment remains the same as in the study population (**eAppendix** 1.2). If the covariates have no effect on the treatment assignment (*i.e.*, $\boldsymbol{\alpha}_{\mathrm{X}} = \mathbf{0}$), the right-hand side reduces to zero, giving a preference score of 0.5 for every individual[5]. Solving the defining equation for the preference score gives:

$$\pi_i = \frac{\frac{e_i}{p}}{\frac{1-e_i}{1-p} + \frac{e_i}{p}}$$

for which the numerator can be considered as an inverse prevalence scaled PS and the denominator seen as a normalizer to constrain π_i within [0,1]. This transformation eliminates the influence of the treatment prevalence. For example, if the treatment is rare (small p), e_i is generally small whereas π_i is not because of the e_i/p operation (small value/small value).

Walker and colleagues proposed an assessment tool based on the proportion of each exposure group that falls within the central region of the preference score distribution [0.3, 0.7] (i.e., 0.5 ± 0.2). Specifically, they proposed that having 50% or more of the subjects in this region indicates that the two drugs are in *empirical equipoise*.[5] That is, the measured prognostic factors do not distinguish the users of one drug from the other, suggesting less danger of confounding by indication.

Extension to the multi-group setting

Here we propose an extension of the tool to settings where interest lies in comparing more than two treatments. Specifically, suppose there are $J + 1$ treatment groups so that A_i is a categorical variable taking on a value in $\{0, 1, \ldots, J\}$. The generalized PS[14] is defined as $e_{ji} = \mathrm{P}[A_i = j \mid \mathbf{X}_i]$ for $j \in \{0, 1, \ldots, J\}$ where $\sum_j e_{ji} = 1$ for all i. One option for modeling the generalized PS is to adopt a baseline-category logit PS model[15], defined by the following J linear predictors:

$$\log\left(\frac{e_{ji}}{e_{0i}}\right) = \log\left(\frac{P[A_i = j|\mathbf{X}_i]}{P[A_i = 0|\mathbf{X}_i]}\right) = \alpha_{0j} + \mathbf{X}_i^T \boldsymbol{\alpha}_{Xj} \text{ for } j \in \{1, \ldots, J\}$$

Let p_j (j = 0, …, J) describe the marginal prevalence of j-th treatment ($\sum_j p_j = 1$) and π_{ji} denote the multinomial preference score defined for the treatment group j for the i-th subject. We propose the *generalized preference score*, defined by the following J equations:

$$\log\left(\frac{\pi_{ji}}{\pi_{0i}}\right) = \log\left(\frac{e_{ji}}{e_{0i}}\right) - \log\left(\frac{p_j}{p_0}\right) \text{ for } j \in \{1, \dots, J\}$$

Solving these equations for π_{ji} using a constraint $\sum_j \pi_{ji} = 1$ (eAppendix 2.1) gives:

$$\pi_{ji} = \frac{\frac{e_{ji}}{p_j}}{\sum_{k=0}^{J} \frac{e_{ki}}{p_k}} \text{ for } j \in \{0,1, \dots, J\}$$

which can be interpreted as the generalized PS scaled by the corresponding group's marginal prevalence.

In extending the definition of the region of empirical equipoise, the threshold value needs to account for the number of groups. Thus, we propose the generalized threshold as:

$$\pi_{ji} \geq \left(\frac{3}{5}\right)\left(\frac{1}{J+1}\right) \text{ for all } j \in \{0,1, \dots, J\}$$

The threshold is 0.3 in the two-group setting and becomes more lenient with the number of groups, for example, 0.2 in the three-group setting. This is necessary because once there are four groups, no individual can have $\pi_{ji} \geq 0.3$ for all four treatments (**eAppendix** 2.2). We note that an appealing feature of the proposed region is that it reduces to [0.3, 0.7] in the two-group case (**eAppendix** 2.3).

Data examples in the three-group setting

We use two observational datasets to demonstrate the face validity of the tool. We used *ternary plots* (**eAppendix** 3.1).[16] The Partners Healthcare Institutional Review Board approved these analyses.

Non-steroidal anti-inflammatory drugs example

This example was an observational study of non-steroidal anti-inflammatory drugs (NSAIDs) taken from an original study of cardiovascular and gastrointestinal safety of analgesics among Medicare beneficiaries with osteoarthritis or rheumatoid arthritis (**eAppendix** 3.2).[17] The dataset included 23,532 naproxen, 21,880 ibuprofen, and 5,261 diclofenac users. As they belong to the same pharmacological class, we expected clinical equipoise. In **Figure 3-1** (left panel), closeness to each corner indicates a high propensity for the corresponding group. The prevalence imbalance drove the center of the distribution away from the smallest diclofenac corner (right lower). Preference scores (**Figure 3-1**, right panel) re-centered the distribution. Of the entire cohort, 86.6 percent fell within the proposed region of empirical equipoise. The individual covariates mostly gave absolute standardized mean distance (SMD) less than 0.1 (**eFigure 3-1** and **eTable 3-1**).[18,19]

Biological disease-modifying anti-rheumatic drugs example

This example was an observational dataset of new users of biological disease-modifying anti-rheumatic drugs (bDMARDs) taken from original studies of cardiovascular safety among rheumatoid arthritis patients (**eAppendix** 3.3)[20,21]. We constructed a first-line bDMARDs cohort and a second-line (switch) bDMARDs cohort after prior use of one of the five tumor necrosis factor inhibitors (TNFi). We expected that the second-line design would result in better equipoise based on clinical reasoning (**eAppendix** 3.3) and a previous study[22]. We used this example to assess if the tool correctly identified the second-line design as superior. In the first-line cohort, there were 2,260 abatacept, 645 tocilizumab, and 27,939 TNFi users. The second-line cohort had 475 abatacept, 187 tocilizumab, and 1,277 *second* TNFi users (switch within TNFi). Only 44.5% of the first-line cohort fell in the proposed region of empirical equipoise (**Figure 3-2**, right upper panel). Using the second-line design (**Figure 3-2**, right lower panel) resulted in improvement with a higher proportion of the cohort (57.7%) falling in the proposed region of empirical equipoise. Absolute SMDs generally decreased, particularly for relevant risk factors such as oral glucocorticoids (**eFigure 3-2**, **eTable 3-2**, and **eTable 3-3**).

Simulation setup

We conducted a simulation study to examine the settings under which the proposed tool reflected the risk of residual confounding.

Data generating mechanism. Details regarding the data generating models are provided in the **eAppendix** (Section 4.1). Briefly, we used the multivariate normal distribution to generate seven correlated normal covariates at correlation values $\rho = 0, 0.1, 0.3, 0.5, 0.7$, and 0.9. These initial covariates were all standard normal marginally. We kept X_1 and X_7 as standard normal. X_2 was transformed to a uniform(0,1) random variable and then to a Poisson variable with mean 1. X_3 through X_6 were similarly transformed to Bernoulli variables with prevalence 20%. Treatment A_i was assigned via a three-group multinomial logistic regression model including all seven covariates. The coefficient for X_7 took on values zero, half, same, or twice as large as the coefficients for X_1- X_6. The coefficients were then simultaneously increased (less equipoise) or decreased (more equipoise). The outcome Y_i was generated as a count outcome using a log-linear model including all covariates and treatment to avoid the issue of non-collapsibility[23]. The rate ratio for X_7 was 1.2 (same as other covariates), 1.5, or 2.0. We handled X_7 as an unmeasured continuous variable in the subsequent analysis.

Methods to be evaluated. The region of empirical equipoise was defined at the threshold of 0.2 as stated above. We examined two assessment rules of three-group empirical equipoise: (1)

whether the proportion of those who were in the region of empirical equipoise in the entire sample was greater than 50% (overall proportion); (2) whether the minimum of three group-specific proportions was greater than 50% (group-specific proportion).

Estimand of interest. The estimands were the rate ratios (RR) for groups 1 vs. 0, groups 2 vs. 0, and groups 2 vs. 1. We conducted unadjusted analysis as well as three PS-weighted analyses with inverse probability of treatment weights (IPTW)[24], matching weights (MW)[25,26], and overlap weights (OW)[27–29]. See **eAppendix (**Section 4.2) for weight definitions.

Performance measures. We examined the relationship between the residual confounding after PS weighting in the RRs and the proportions in the region of empirical equipoise. The desired result was a decreasing trend in the proportions in the region with increasing residual confounding. We also examined the approximate value of residual confounding at which the 50% threshold was crossed.

RESULTS

Figures 3-5 summarize the results from scenarios with no correlation among covariates ($\rho = 0$) and approximately equal group sizes (33:33:33). The columns of panels correspond to PS weighting methods. The rows of panels correspond to the RR for the unmeasured X_7. Focusing on the panel in the MW column and RR 1.5 row (third column, second row) in **Figure 3-3**, the X-axis represents the multiplicative bias in RR estimates, whereas the Y-axis represents the average proportion of the simulated cohorts within the region of empirical equipoise (overall proportion).

The relationship between the residual confounding after PS weighting and the overall proportion varied with the relative strength of association of X_7 with the treatment (denoted by line types). Given an unmeasured confounder with a similar association with treatment (*Same* line type), having an overall proportion of 50% in the region of empirical equipoise (crossing of the horizontal 50% line) corresponded to residual confounding of roughly 1.2 (20% upward bias in RR estimates). This indicates in a setting where the unmeasured factor's treatment association is similar to those of measured factors and the outcome association is only modestly stronger (+25%), the empirical equipoise tool would give an alert (overall proportion would drop below 50%) once the residual confounding is greater than 20%. A proportion above 50% means less bias.

Still focusing on the same panel in **Figure 3-3**, the level of residual confounding at which the empirical equipoise tool gave an alert depended on the associations of X_7 with the treatment and outcome. On the other hand, the type of PS weights (IPTW, MW, and OW) made little difference. When the relative treatment association of X_7 was decreased to the lower extreme end (no unmeasured confounding; solid line), the tool became overly sensitive. That is, the 50% threshold was crossed without a corresponding increase in residual confounding. On the other hand, as we increased the association of the unmeasured variable X_7 and the treatment to twice as large as the measured ones, the slopes became shallower. This means the tool became less sensitive to residual confounding, only crossing the 50% overall proportion threshold at a residual confounding level of about 1.37. That is, the unmeasured variable increasingly had a stronger effect on treatment not represented by the association between measured variables and treatment.

We also varied the level of unmeasured confounding by changing the RR between the unmeasured variable X_7 and the outcome (rows of panels; RR 1.2, 1.5, and 2.0). For example, decreasing the unmeasured variable-outcome association to the same level as the other variables (third column, top row in **Figure 3-3**) resulted in the tool giving an alert at a residual confounding of roughly 1.1 (more sensitive) when X_7 had the same treatment association. When increasing the RR between the unmeasured variable X_7 and the outcome to 2.0 (67% increase over measured variables), the tool gave an alert at a residual confounding of around 1.3 (less sensitive). When both associations were strong for the unmeasured variable X_7, the 50% overall proportion threshold was crossed at a residual confounding of around 1.6. This means having barely 50% of the cohort in this region does not assure a small level of unmeasured confounding in this setting.

For the group 2 vs. 0 contrast (**Figure 3-4**), which was designed to have more different covariate distributions, greater levels of residual bias were observed. The group 2 vs. 1 contrast (**Figure 3-5**) gave similar results to the group 1 vs. 0 contrast. In all contrasts (**Figures 3-5**), using a threshold of 75% instead of 50% would lead to a smaller range of biases although this comes at the cost of disregarding study design where the unmeasured variable indeed had weaker associations than measured ones. The results were similar when we varied treatment prevalence (**eAppendix** 5.1) and when we switched the assessment metric to the group-specific proportion (**eAppendix** 5.2). Also, the results were invariant with increasing correlation among covariates except in the very extreme setting with $\rho = 0.9$, in which the residual confounding was reduced by surrogacy via highly correlated measured variables. (**eAppendix** 5.3).

DISCUSSION

We extended Walker *et al.*'s tool[5] for assessing simultaneous empirical equipoise among multiple treatment groups in CER. We demonstrated its face validity in empirical data and examined its performance in simulations with three groups. Our simulations showed that having at least 50% of the overall cohort in the region of empirical equipoise can give a reasonable assurance of relatively small magnitude of residual bias. However, in settings with a strong unmeasured variable (outcome association RR of 2.0) and a strong influence of the unmeasured variable on treatment choice (twice more on the logit scale), a relatively large residual bias went undetected by the 50% threshold. As a result, the tool was most useful when we could assume the unmeasured confounder had covariate-treatment associations similar in magnitude to the measured confounders.

There are several ways this empirical equipoise assessment tool could be useful in the implementation of multi-group CER. First, when several datasets are available for a specific multi-group CER question, the tool could indicate which dataset may suffer less from residual confounding as well as positivity issue. Second, when dealing with one dataset, the tool may help in choosing eligibility criteria although sample size issues may need to be taken into consideration. Thirdly, another potential change in the study design is to refrain from conducting all comparison if the groups do not achieve reasonable simultaneous empirical equipoise (*e.g.*, if key covariates are highly imbalanced in one group but not in the others). In this case, dropping one or more groups from the comparison may identify a subset of groups in better equipoise.

Our tool is useful in providing a feasibility assessment[7] for *simultaneous* multi-group comparison in a single outcome analysis dataset. However, when there are three or more groups, pairwise PS-matched or PS-weighted cohort construction is more common in practice. A potential drawback of the pairwise approach is that it produces multiple outcome analysis datasets, one for each pairwise comparison, with potentially different target populations. Thus, *non-transitivity* can arise. That is, in the three-group setting, the first two comparisons do not sufficiently inform the result of the third comparison. For example in the study by Rassen et al.[30] (see their supplement), COX2 selective inhibitors (coxibs) in comparison to NSAIDs had a hazard ratio (HR) of 1.86 [95% confidence interval (CI) 1.14, 3.03] for myocardial infarction (MI), whereas opioids in contrast to NSAIDs had an HR of 1.40 [0.81, 2.40]. One would expect an HR around 0.75 (= 1.40/1.86) for the remaining opioids vs. coxibs comparison. However, their pairwise analysis gave an HR of 1.02 [0.74, 1.41] for this third contrast. Simultaneous empirical equipoise assessment followed by construction of a single PS weighted cohort eliminates this issue by focusing on individuals who are reasonable candidates for all treatments.

There are differences between the context in which Walker *et al.* developed the original empirical equipoise tool[5] and the context for our proposed tool. We considered the drugs of interest that we want to compare in the proposed multi-group CER as given. Walker *et al.* proposed the tool as a prioritization tool given a source dataset that contains information on the use of many drugs. They developed their tool to assess the empirical equipoise of all possible pairwise contrasts of groups for prioritization. On the other hand, we framed our problem in a setting where we already had several drugs of interest *a priori*, with several alternative data sources or alternative designs to choose from.

In conclusion, to examine the roles that equipoise assessment may play in the setting of multi-group CER, we extended Walker *et al.*'s empirical equipoise tool. Our tool gave reasonable guidance for unmeasured confounding when the associations of the unmeasured variables to the treatment and outcome were similar to associations of measured covariates. With this assumption, when the proportion in the region of empirical equipoise is very high, for example, > 75%, we can reasonably assume that the level of residual confounding is small. A lower value, particularly < 50%, should prompt reconsideration of the study design or data source.

References

1. Petersen ML, Porter KE, Gruber S, Wang Y, van der Laan MJ. Diagnosing and responding to violations in the positivity assumption. *Stat Methods Med Res* 2012; **21**: 31–54. doi:10.1177/0962280210386207.

2. Rubin DB. For objective causal inference, design trumps analysis. *Ann Appl Stat* 2008; **2**: 808–840. doi:10.1214/08-AOAS187.

3. Yoshida K, Solomon DH, Kim SC. Active-comparator design and new-user design in observational studies. *Nat Rev Rheumatol* 2015; **11**: 437–441. doi:10.1038/nrrheum.2015.30.

4. Lund JL, Richardson DB, Stürmer T. The active comparator, new user study design in pharmacoepidemiology: historical foundations and contemporary application. *Curr Epidemiol Rep* 2015; **2**: 221–228. doi:10.1007/s40471-015-0053-5.

5. Walker AM, Patrick AR, Lauer MS, *et al.* A tool for assessing the feasibility of comparative effectiveness research. *Comp Eff Res* 2013: 11. doi:10.2147/CER.S40357.

6. Freedman B. Equipoise and the ethics of clinical research. *N Engl J Med* 1987; **317**: 141–145. doi:10.1056/NEJM198707163170304.

7. Girman CJ, Faries D, Ryan P, *et al.* Pre-study feasibility and identifying sensitivity analyses for protocol pre-specification in comparative effectiveness research. *J Comp Eff Res* 2014; **3**: 259–270. doi:10.2217/cer.14.16.

8. Singh JA, Saag KG, Bridges SL, *et al.* 2015 American College of Rheumatology Guideline for the Treatment of Rheumatoid Arthritis. *Arthritis Care Res (Hoboken)* 2016; **68**: 1–25. doi:10.1002/acr.22783.

9. American Diabetes Association. 8. Pharmacologic Approaches to Glycemic Treatment: Standards of Medical Care in Diabetes-2018. *Diabetes Care* 2018; **41**: S73–S85. doi:10.2337/dc18-S008.

10. January CT, Wann LS, Alpert JS, *et al.* 2014 AHA/ACC/HRS guideline for the management of patients with atrial fibrillation: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines and the Heart Rhythm Society. *J Am Coll Cardiol* 2014; **64**: e1-76. doi:10.1016/j.jacc.2014.03.022.

11. Hosmer DW, Lemeshow S, Sturdivant RX. *Applied logistic regression*. Third edition. Hoboken, New Jersey: Wiley, 2013.

12. Scott AJ, Wild CJ. Fitting Logistic Models Under Case-Control or Choice Based Sampling. *J Royal Stat Soc* 1986; **48**: 170–182.

13. Prentice RL, Pyke R. Logistic Disease Incidence Models and Case-Control Studies. *Biometrika* 1979; **66**: 403–411. doi:10.2307/2335158.

14. Imbens GW. The role of the propensity score in estimating dose-response functions. *Biometrika* 2000; **87**: 706–710. doi:10.1093/biomet/87.3.706.

15. Agresti A. *Categorical Data Analysis*. 3 edition. Hoboken, NJ: Wiley, 2012.

16. Hamilton N. *ggtern: An Extension to "ggplot2", for the Creation of Ternary Diagrams*., 2017. Available at: https://cran.r-project.org/web/packages/ggtern/index.html. Accessed January 25, 2018.

17. Solomon DH, Rassen JA, Glynn RJ, Lee J, Levin R, Schneeweiss S. The comparative safety of analgesics

in older adults with arthritis. *Arch Intern Med* 2010; **170**: 1968–1976. doi:10.1001/archinternmed.2010.391.

18. Austin PC. An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. *Multivariate Behav Res* 2011; **46**: 399–424. doi:10.1080/00273171.2011.568786.

19. Austin PC. Using the Standardized Difference to Compare the Prevalence of a Binary Variable Between Two Groups in Observational Research. *Communications in Statistics - Simulation and Computation* 2009; **38**: 1228–1234. doi:10.1080/03610910902859574.

20. Kim SC, Solomon DH, Rogers JR, *et al.* Cardiovascular Safety of Tocilizumab Versus Tumor Necrosis Factor Inhibitors in Patients With Rheumatoid Arthritis: A Multi-Database Cohort Study. *Arthritis & Rheumatology (Hoboken, NJ)* 2017; **69**: 1154–1164. doi:10.1002/art.40084.

21. Kang EH, Jin Y, Brill G, *et al.* Comparative Cardiovascular Risk of Abatacept and Tumor Necrosis Factor Inhibitors in Patients With Rheumatoid Arthritis With and Without Diabetes Mellitus: A Multidatabase Cohort Study. *J Am Heart Assoc* 2018; **7**. doi:10.1161/JAHA.117.007393.

22. Frisell T, Baecklund E, Bengtsson K, *et al.* Patient characteristics influence the choice of biological drug in RA, and will make non-TNFi biologics appear more harmful than TNFi biologics. *Ann Rheum Dis* 2017. doi:10.1136/annrheumdis-2017-212395.

23. Greenland S, Robins JM, Pearl J. Confounding and Collapsibility in Causal Inference. *Statistical Science* 1999; **14**: 29–46.

24. Robins JM, Hernán MA, Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology* 2000; **11**: 550–560.

25. Li L, Greene T. A weighting analogue to pair matching in propensity score analysis. *Int J Biostat* 2013; **9**: 215–234. doi:10.1515/ijb-2012-0030.

26. Yoshida K, Hernandez-Diaz S, Solomon DH, *et al.* Matching Weights to Simultaneously Compare Three Treatment Groups: Comparison to Three-way Matching. *Epidemiology* 2017; **28**: 387–395. doi:10.1097/EDE.0000000000000627.

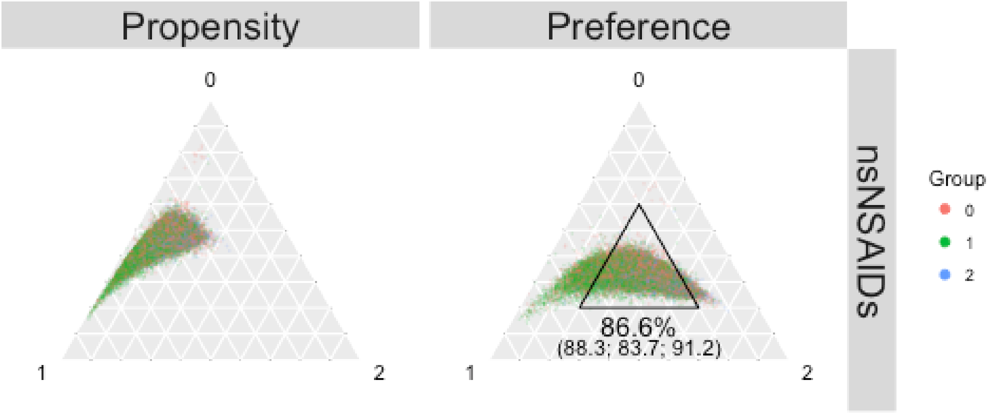
27. Li F, Morgan KL, Zaslavsky AM. Balancing Covariates via Propensity Score Weighting. *Journal of the American Statistical Association* 2016; **0**: 1–11. doi:10.1080/01621459.2016.1260466.

28. Li F, Thomas LE, Li F. Addressing Extreme Propensity Scores via the Overlap Weights. *Am J Epidemiol* 2018. doi:10.1093/aje/kwy201.

29. Li F, Li F. Propensity Score Weighting for Causal Inference with Multi-valued Treatments. *arXiv:180805339 [stat]* 2018. Available at: http://arxiv.org/abs/1808.05339. Accessed August 23, 2018.

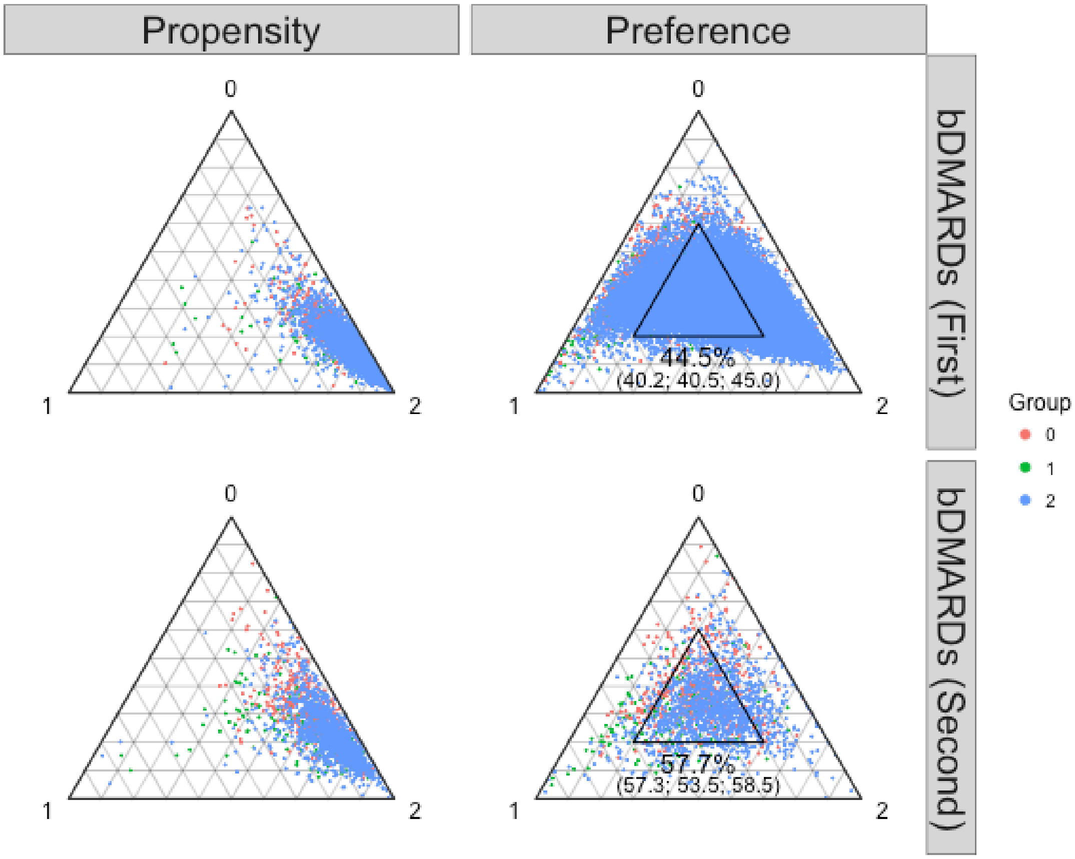
30. Rassen JA, Shelat AA, Franklin JM, Glynn RJ, Solomon DH, Schneeweiss S. Matching by propensity score in cohort studies with three treatment groups. *Epidemiology* 2013; **24**: 401–409. doi:10.1097/EDE.0b013e318289dedf.

Figure 3-1. Propensity score (left) and preference score (right) distributions in the naproxen (0 red; n = 23,532), ibuprofen (green 1; n = 21,880), and diclofenac (2 blue; n = 5,261) example.



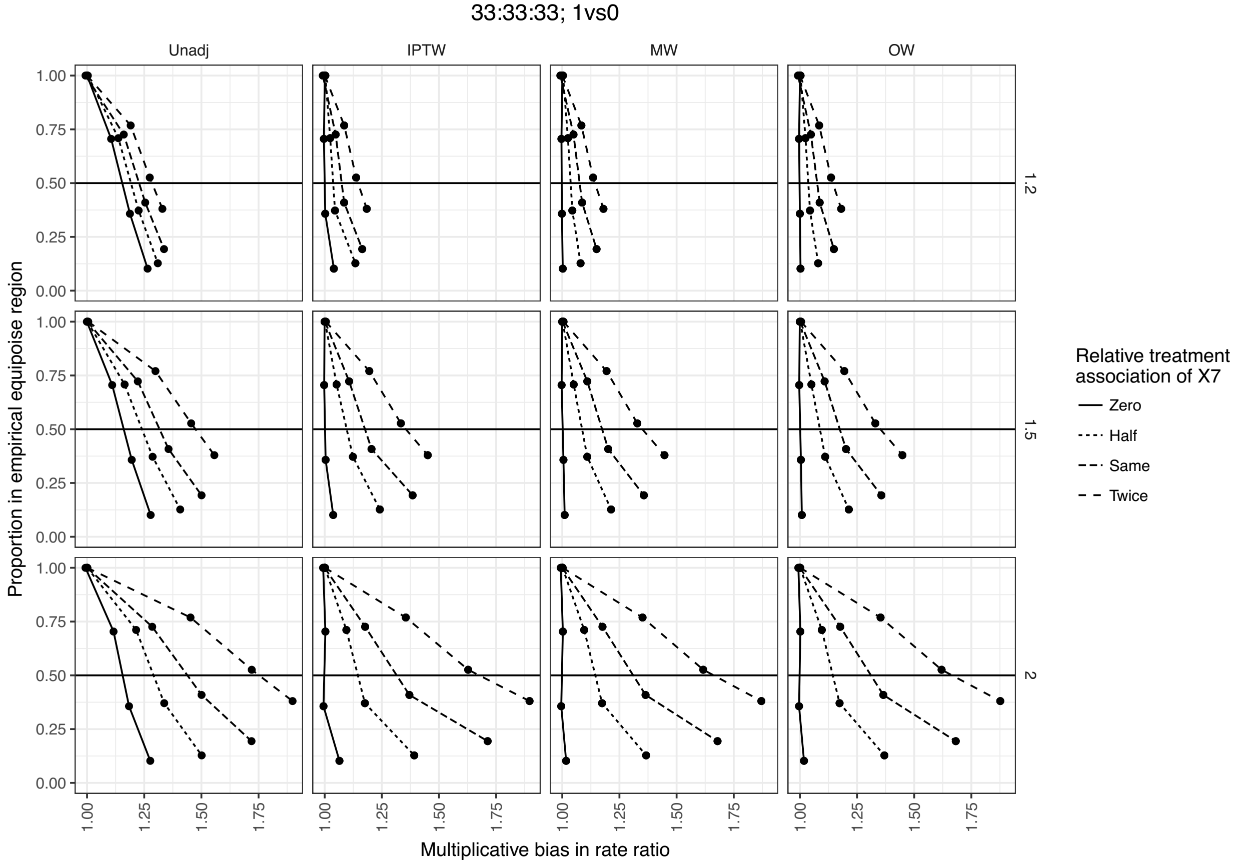
The inner triangular area in the right panel indicates the region of empirical equipoise proposed in the text. Overall 86.6% of the cohort fell into this region (88.3% of naproxen users, 83.7% of ibuprofen users, and 91.2% of diclofenac users).

Figure 3-2. Propensity score (left) and preference score (right) distributions in the abatacept (0 red), tocilizumab (1 green), and TNFi (2 blue) examples.



The inner triangular area in the right panel indicates the region of empirical equipoise proposed in the text. Among the first-line bDMARD users, 44.5% of the cohort fell into this region (40.2% of abatacept users, 40.5% of tocilizumab users, and 45.0% of TNFi users). Among the second-line bDMARD users, 57.7% of the cohort fell into this region (57.3% of abatacept users, 53.5% of tocilizumab users, and 58.5% of TNFi users).
Abbreviations: TNFi (tumor necrosis factor inhibitor); bDMARD: biological disease-modifying antirheumatic drug.

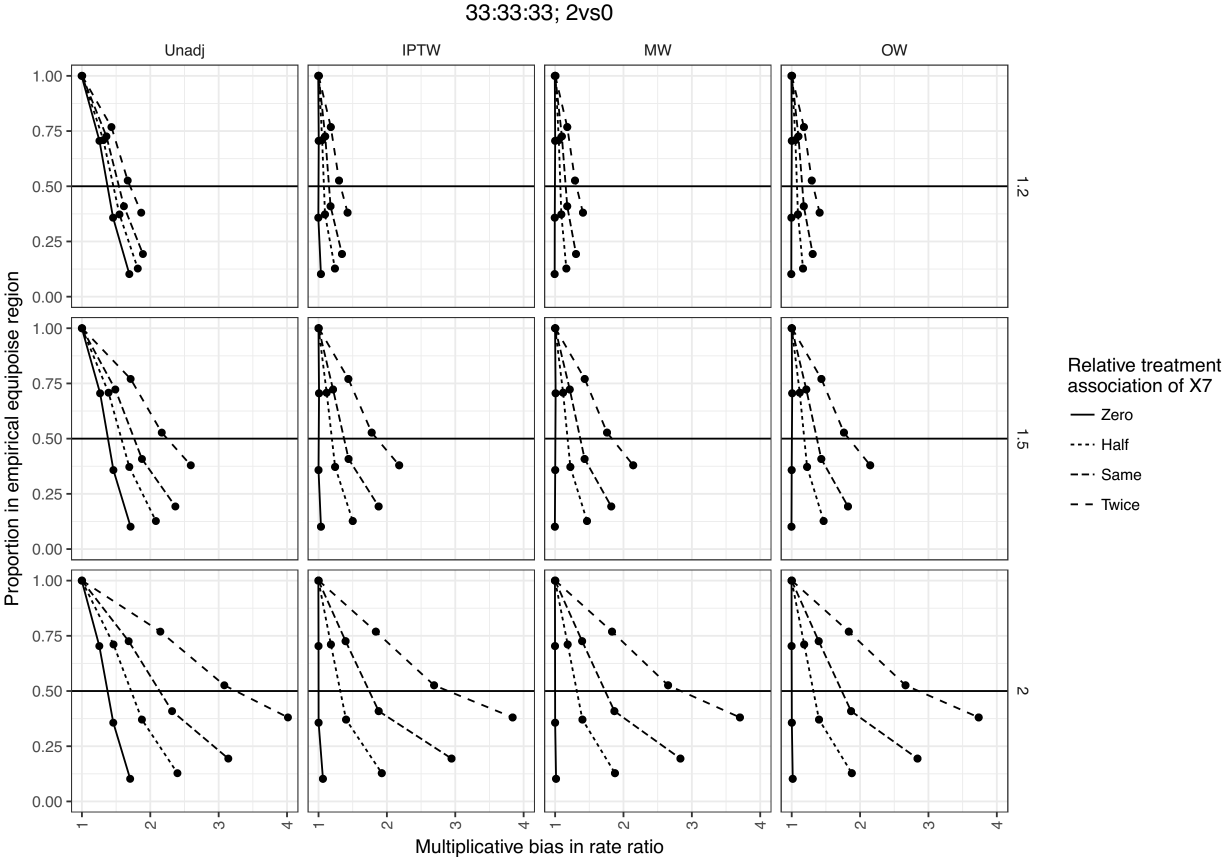
Figure 3-3. Simulation results from scenarios with equal group sizes (1 vs 0 contrast).



The *columns* of panels denote different confounding adjustment methods. The *rows* of panels denote different levels of associations between X_7 (unmeasured covariate) and outcome. A rate ratio of 1.2 was the same strength of association as the measured covariates, whereas only X_7 had a stronger outcome association at a rate ratio of 1.5 and 2.0. In each panel, the X-axis represents the multiplicative bias in RR estimates, whereas the Y-axis represents the average proportion of the simulated cohorts within the region of empirical equipoise (overall proportion). The *line types* denote different levels of associations between X_7 and treatment relative to the associations between measured variables and treatment.

Abbreviations: Unadj.: unadjusted; IPTW: inverse probability of treatment weights; MW: matching weights; OW: overlap weights.

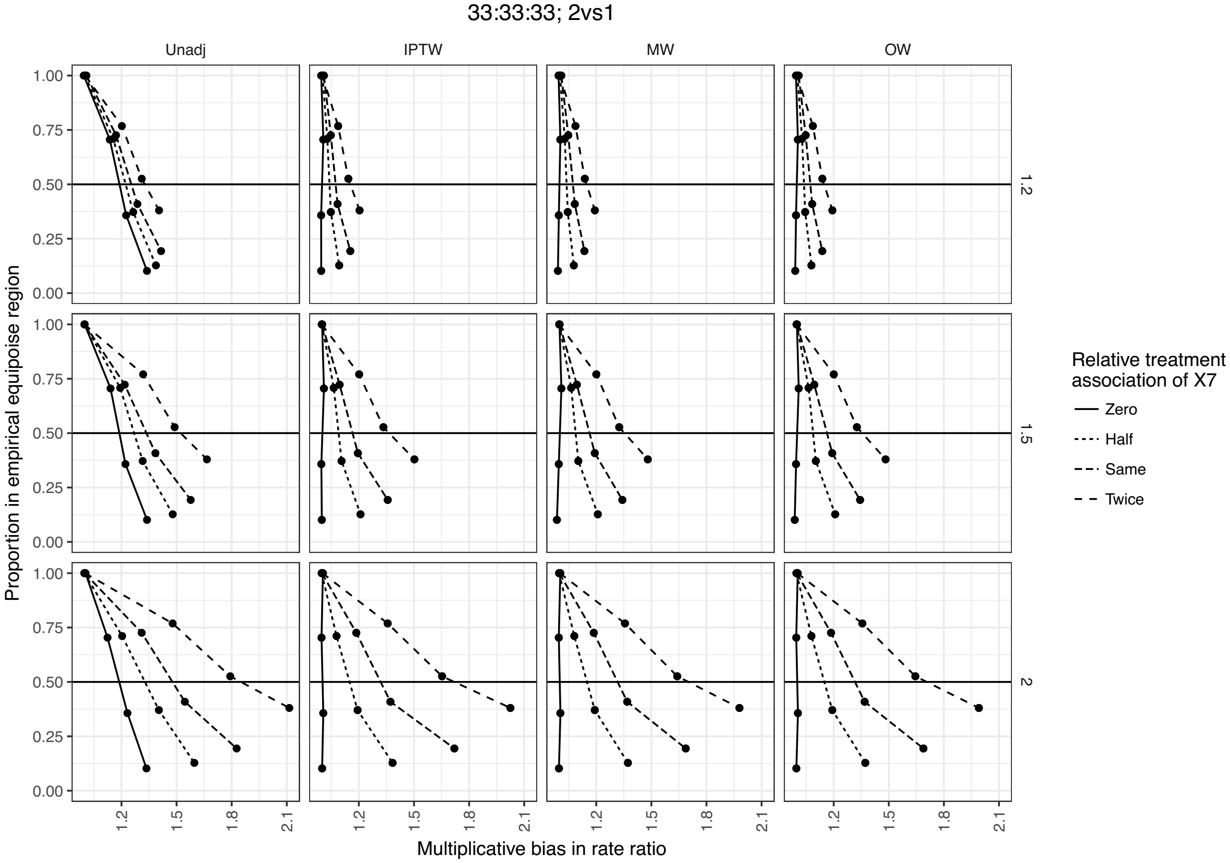
Figure 3-4. Simulation results from scenarios with equal group sizes (2 vs 0 contrast).



The *columns* of panels denote different confounding adjustment methods. The *rows* of panels denote different levels of associations between X_7 (unmeasured covariate) and outcome. A rate ratio of 1.2 was the same strength of association as the measured covariates, whereas only X_7 had a stronger outcome association at a rate ratio of 1.5 and 2.0. In each panel, the X-axis represents the multiplicative bias in RR estimates, whereas the Y-axis represents the average proportion of the simulated cohorts within the region of empirical equipoise (overall proportion). The *line types* denote different levels of associations between X_7 and treatment relative to the associations between measured variables and treatment.

Abbreviations: Unadj.: unadjusted; IPTW: inverse probability of treatment weights; MW: matching weights; OW: overlap weights.

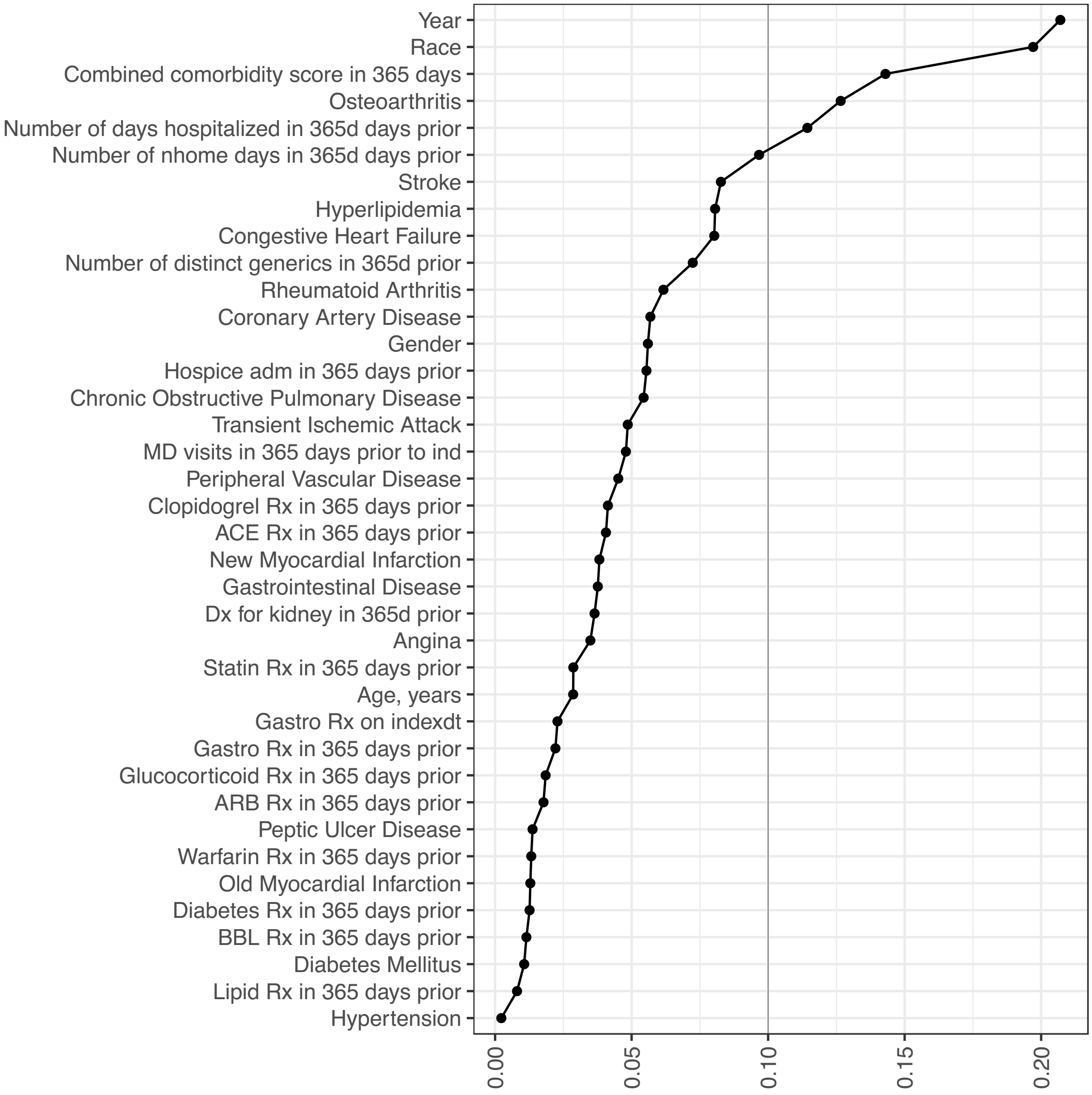
Figure 3-5. Simulation results from scenarios with equal group sizes (2 vs 1 contrast).



The *columns* of panels denote different confounding adjustment methods. The *rows* of panels denote different levels of associations between X_7 (unmeasured covariate) and outcome. A rate ratio of 1.2 was the same strength of association as the measured covariates, whereas only X_7 had a stronger outcome association at a rate ratio of 1.5 and 2.0. In each panel, the X-axis represents the multiplicative bias in RR estimates, whereas the Y-axis represents the average proportion of the simulated cohorts within the region of empirical equipoise (overall proportion). The *line types* denote different levels of associations between X_7 and treatment relative to the associations between measured variables and treatment.

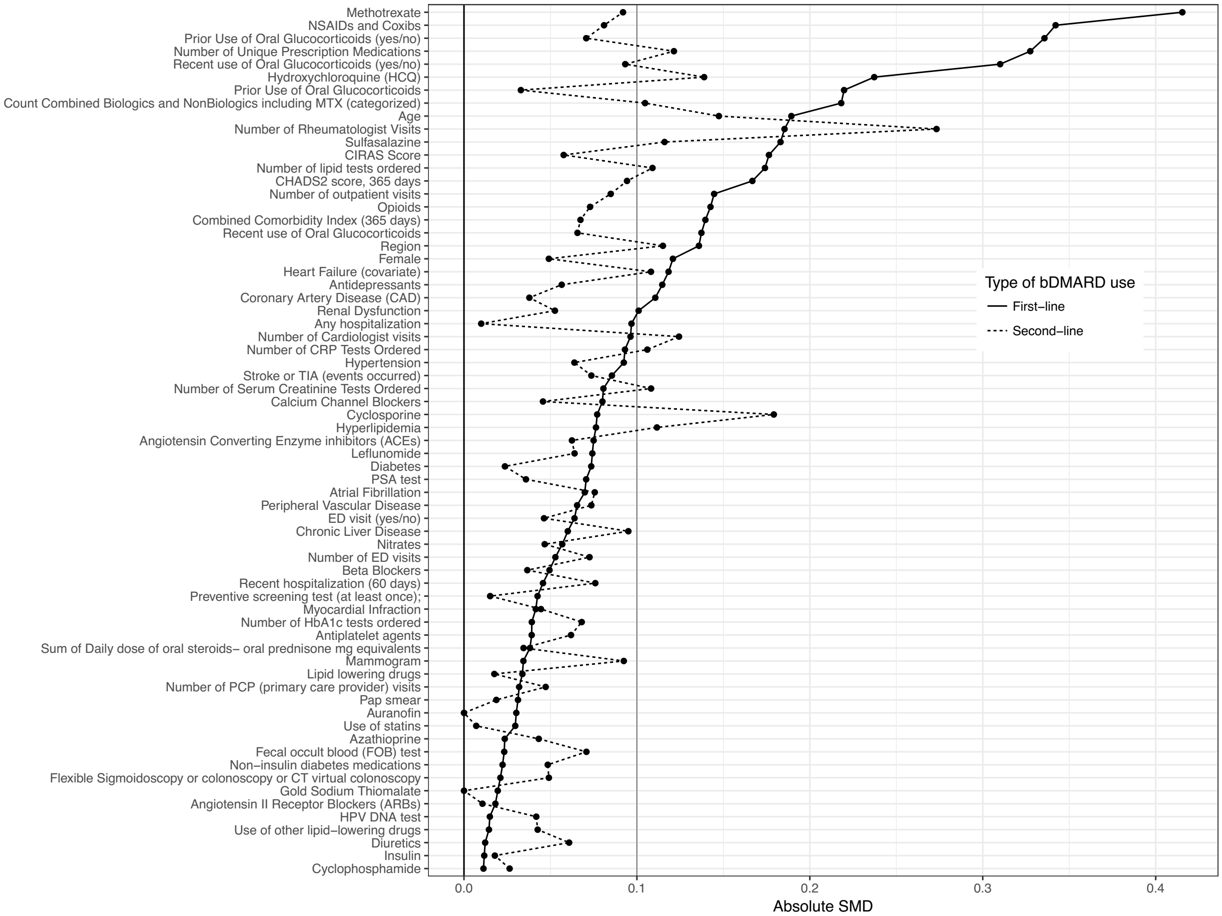
Abbreviations: Unadj.: unadjusted; IPTW: inverse probability of treatment weights; MW: matching weights; OW: overlap weights.

eFigure 3-1. Average absolute standardized mean differences between groups in the three non-selective NSAIDs example.



Abbreviations: adm: admission; MD: physician; ind: index date; Rx: prescription; ACE: angiotensin converting enzyme; ARB: angiotensin receptor blocker; BBL: beta blocker.

eFigure 3-2. Average absolute standardized mean differences between groups in the three biological DMARDs example.



The solid line indicates the cohort of first-line bDMARD users, whereas the dotted line indicates the cohort of second-line bDMARD users.

Abbreviations: SMD: absolute standardized mean difference; MTX: methotrexate; CRP: C-reactive protein; NSAID: non-steroidal anti-inflammatory drug; PSA: prostate-specific antigen; ED: emergency department; HbA1c: hemoglobin A1c; TIA: transient ischemic attack; CT: computed tomography; MI: myocardial infarction; HPV: human papilloma virus.

eTable 3-1. Patient group characteristics in the three non-selective non-steroidal anti-inflammatory drugs example.

Variable	Naproxen	Ibuprofen	Diclofenac	SMD
n	23532	21880	5261	
White (%)	20126 (85.5)	17823 (81.5)	4777 (90.8)	0.182
Combined comorbidity score in 365 days (mean (sd))	1.40 (2.38)	1.72 (2.60)	1.20 (2.19)	0.143
Osteoarthritis (%)	9936 (42.2)	8919 (40.8)	2640 (50.2)	0.127
Number of days hospitalized in 365d days prior (mean (sd))	3.25 (9.25)	4.23 (10.82)	2.63 (7.61)	0.114
Number of nhome days in 365d days prior (mean (sd))	1.62 (9.24)	2.75 (12.71)	1.24 (7.64)	0.097
Stroke (%)	2153 (9.1)	2305 (10.5)	370 (7.0)	0.083
Hyperlipidemia (%)	14102 (59.9)	12242 (56.0)	3256 (61.9)	0.081
Congestive Heart Failure (%)	4208 (17.9)	4579 (20.9)	855 (16.3)	0.080
Number of distinct generics in 365d prior (mean (sd))	10.98 (5.82)	11.53 (6.18)	10.88 (5.65)	0.072
Rheumatoid Arthritis (%)	862 (3.7)	721 (3.3)	271 (5.2)	0.062
Coronary Artery Disease (%)	9042 (38.4)	8859 (40.5)	1912 (36.3)	0.057
Male (%)	4538 (19.3)	4348 (19.9)	875 (16.6)	0.056
Hospice adm in 365 days prior (%)	39 (0.2)	102 (0.5)	3 (0.1)	0.055
Chronic Obstructive Pulmonary Disease (%)	5043 (21.4)	5084 (23.2)	1046 (19.9)	0.054
Transient Ischemic Attack (%)	1216 (5.2)	1372 (6.3)	243 (4.6)	0.049
MD visits in 365 days prior to indexdt (mean (sd))	10.30 (7.49)	9.87 (7.37)	10.39 (7.04)	0.048
Peripheral Vascular Disease (%)	3800 (16.1)	4057 (18.5)	841 (16.0)	0.045
Clopidogrel Rx in 365 days prior (%)	1852 (7.9)	1899 (8.7)	369 (7.0)	0.041
ACE Rx in 365 days prior (%)	6540 (27.8)	6219 (28.4)	1353 (25.7)	0.041
New Myocardial Infarction (%)	619 (2.6)	743 (3.4)	128 (2.4)	0.038
Gastrointestinal Disease (%)	1129 (4.8)	1183 (5.4)	221 (4.2)	0.038
Dx for kidney in 365d prior (%)	1017 (4.3)	1071 (4.9)	199 (3.8)	0.036
Angina (%)	1435 (6.1)	1458 (6.7)	285 (5.4)	0.035
Statin Rx in 365 days prior (%)	8519 (36.2)	7472 (34.1)	1900 (36.1)	0.029
Age, years (mean (sd))	77.77 (7.07)	78.07 (7.32)	77.93 (6.96)	0.029
Gastro Rx on indexdt (%)	4863 (20.7)	4829 (22.1)	1121 (21.3)	0.023
Gastro Rx in 365 days prior (%)	7871 (33.4)	7651 (35.0)	1757 (33.4)	0.022
Glucocorticoid Rx in 365 days prior (%)	2445 (10.4)	2353 (10.8)	592 (11.3)	0.019
ARB Rx in 365 days prior (%)	2733 (11.6)	2422 (11.1)	627 (11.9)	0.018
Peptic Ulcer Disease (%)	4998 (21.2)	4810 (22.0)	1112 (21.1)	0.014
Warfarin Rx in 365 days prior (%)	1445 (6.1)	1428 (6.5)	318 (6.0)	0.013
Old Myocardial Infarction (%)	1172 (5.0)	1175 (5.4)	260 (4.9)	0.013
Diabetes Rx in 365 days prior (%)	4802 (20.4)	4597 (21.0)	1065 (20.2)	0.013
BBL Rx in 365 days prior (%)	8309 (35.3)	7906 (36.1)	1863 (35.4)	0.011
Diabetes Mellitus (%)	7745 (32.9)	7366 (33.7)	1745 (33.2)	0.011
Lipid Rx in 365 days prior (%)	683 (2.9)	647 (3.0)	145 (2.8)	0.008
Hypertension (%)	19048 (80.9)	17700 (80.9)	4263 (81.0)	0.002

Abbreviations: nhome: nursing home; adm: admission; MD: physician; indexdt: index date; Rx: prescription; ACE: angiotensin converting enzyme; ARB: angiotensin receptor blocker; BBL: beta blocker.

eTable 3-2. Patient group characteristics in the three biological DMARDs example (first-line bDMARDs).

Variable	ABA	TCZ	TNF	SMD
n	2260	645	27939	
Methotrexate (%)	1140 (50.4)	257 (39.8)	19487 (69.7)	0.415
NSAIDs and Coxibs (%)	716 (31.7)	147 (22.8)	13018 (46.6)	0.342
Prior Use of Oral Glucocorticoids (yes/no) (%)	1207 (53.4)	295 (45.7)	19556 (70.0)	0.336
Number of Unique Prescription Medications (mean (sd))	9.44 (8.78)	7.51 (8.48)	11.49 (7.26)	0.327
Recent use of Oral Glucocorticoids (yes/no) (%)	1122 (49.6)	280 (43.4)	18460 (66.1)	0.310
Hydroxychloroquine (HCQ) (%)	672 (29.7)	109 (16.9)	8932 (32.0)	0.237
Prior Use of Oral Glucocorticoids (mean (sd))	149.52 (161.35)	128.12 (159.25)	180.10 (153.86)	0.220
Count Combined Biologics and NonBiologics including MTX (categorized) (mean (sd))	0.48 (0.71)	0.29 (0.60)	0.51 (0.70)	0.218
Age (mean (sd))	54.91 (13.08)	54.24 (13.24)	51.26 (12.49)	0.189
Number of Rheumatologist Visits (mean (sd))	12.62 (18.17)	15.48 (22.19)	10.33 (14.41)	0.185
Sulfasalazine (%)	176 (7.8)	26 (4.0)	3129 (11.2)	0.183
CIRAS Score (mean (sd))	6.47 (2.06)	6.51 (1.91)	6.99 (1.94)	0.176
Number of lipid tests ordered (mean (sd))	0.74 (1.47)	1.09 (1.31)	0.76 (1.26)	0.174
CHADS2 score, 365 days (mean (sd))	0.83 (1.01)	0.85 (1.05)	0.61 (0.85)	0.167
Number of outpatient visits (mean (sd))	14.69 (9.42)	14.54 (8.44)	12.85 (7.93)	0.145
Opioids (%)	1133 (50.1)	300 (46.5)	15967 (57.1)	0.143
Combined Comorbidity Index (365 days) (mean (sd))	0.56 (1.46)	0.58 (1.44)	0.31 (1.12)	0.140
Recent use of Oral Glucocorticoids (mean (sd))	96.24 (127.77)	87.91 (128.51)	113.63 (120.09)	0.137
Region (%)				0.136
Northeast	406 (18.0)	116 (18.0)	4589 (16.4)	
North Central	495 (21.9)	118 (18.3)	5860 (21.0)	
South	885 (39.2)	233 (36.1)	11668 (41.8)	
West	394 (17.4)	145 (22.5)	4753 (17.0)	
Unknown	80 (3.5)	33 (5.1)	1069 (3.8)	
Female (%)	1864 (82.5)	529 (82.0)	20982 (75.1)	0.121
Heart Failure (covariate) (%)	103 (4.6)	24 (3.7)	430 (1.5)	0.118
Antidepressants (%)	592 (26.2)	150 (23.3)	8626 (30.9)	0.115
Coronary Artery Disease (CAD) (%)	250 (11.1)	62 (9.6)	1788 (6.4)	0.111
Renal Dysfunction (%)	135 (6.0)	43 (6.7)	943 (3.4)	0.101
Any hospitalization (%)	346 (15.3)	97 (15.0)	2921 (10.5)	0.097
Number of Cardiologist visits (mean (sd))	1.22 (3.87)	1.09 (3.52)	0.73 (2.97)	0.096
Number of CRP Tests Ordered (mean (sd))	2.03 (2.35)	2.37 (2.74)	2.05 (2.10)	0.093
Hypertension (%)	1217 (53.8)	347 (53.8)	13113 (46.9)	0.092
Stroke or TIA (events occurred) (%)	52 (2.3)	21 (3.3)	377 (1.3)	0.086
Number of Serum Creatinine Tests Ordered (mean (sd))	3.60 (3.16)	3.96 (3.56)	3.56 (2.87)	0.081
Calcium Channel Blockers (%)	267 (11.8)	53 (8.2)	2966 (10.6)	0.080
Cyclosporine (%)	75 (3.3)	10 (1.6)	594 (2.1)	0.077

eTable 3-2 (continued)

Variable	ABA	TCZ	TNF	SMD
Hyperlipidemia (%)	824 (36.5)	238 (36.9)	8794 (31.5)	0.076
Angiotensin Converting Enzyme inhibitors (ACEs) (%)	295 (13.1)	70 (10.9)	4077 (14.6)	0.075
Leflunomide (%)	297 (13.1)	62 (9.6)	3255 (11.7)	0.074
Diabetes (%)	429 (19.0)	115 (17.8)	4149 (14.9)	0.074
PSA test (%)	103 (4.6)	30 (4.7)	1964 (7.0)	0.071
Atrial Fibrillation (%)	81 (3.6)	14 (2.2)	526 (1.9)	0.070
Peripheral Vascular Disease (%)	79 (3.5)	18 (2.8)	534 (1.9)	0.065
ED visit (yes/no) (%)	122 (5.4)	33 (5.1)	959 (3.4)	0.064
Chronic Liver Disease (%)	147 (6.5)	48 (7.4)	1467 (5.3)	0.060
Nitrates (%)	48 (2.1)	18 (2.8)	434 (1.6)	0.057
Number of ED visits (mean (sd))	0.24 (1.25)	0.28 (1.64)	0.17 (1.40)	0.053
Beta Blockers (%)	353 (15.6)	84 (13.0)	3885 (13.9)	0.049
Recent hospitalization (60 days) (%)	47 (2.1)	17 (2.6)	460 (1.6)	0.046
Preventive screening test (at least once); (%)	1149 (50.8)	319 (49.5)	14709 (52.6)	0.043
Myocardial Infraction (%)	9 (0.4)	3 (0.5)	36 (0.1)	0.042
Number of HbA1c tests ordered (mean (sd))	0.38 (0.88)	0.41 (0.94)	0.36 (0.85)	0.039
Antiplatelet agents (%)	74 (3.3)	23 (3.6)	714 (2.6)	0.039
Sum of Daily dose of oral steroids- oral prednisone mg equivalents (mean (sd))	880.02 (3609.78)	1595.18 (20763.81)	1022.15 (5922.25)	0.038
Mammogram (%)	723 (32.0)	191 (29.6)	8319 (29.8)	0.034
Lipid lowering drugs (%)	463 (20.5)	127 (19.7)	6075 (21.7)	0.034
Number of PCP (primary care provider) visits (mean (sd))	6.58 (14.91)	6.00 (13.90)	5.92 (11.68)	0.032
Pap smear (%)	536 (23.7)	146 (22.6)	6880 (24.6)	0.031
Auranofin (%)	2 (0.1)	0 (0.0)	8 (0.0)	0.030
Use of statins (%)	416 (18.4)	115 (17.8)	5466 (19.6)	0.030
Azathioprine (%)	50 (2.2)	13 (2.0)	481 (1.7)	0.024
Fecal occult blood (FOB) test (%)	163 (7.2)	46 (7.1)	2251 (8.1)	0.023
Non-insulin diabetes medications (%)	162 (7.2)	46 (7.1)	2240 (8.0)	0.022
Flexible Sigmoidoscopy or colonoscopy or CT virtual colonoscopy (%)	200 (8.8)	63 (9.8)	2701 (9.7)	0.021
Gold Sodium Thiomalate (%)	0 (0.0)	0 (0.0)	12 (0.0)	0.020
Angiotensin II Receptor Blockers (ARBs) (%)	243 (10.8)	64 (9.9)	2848 (10.2)	0.018
HPV DNA test (%)	170 (7.5)	45 (7.0)	2112 (7.6)	0.015
Use of other lipid-lowering drugs (%)	88 (3.9)	26 (4.0)	1208 (4.3)	0.014
Diuretics (%)	287 (12.7)	78 (12.1)	3500 (12.5)	0.012
Insulin (%)	81 (3.6)	23 (3.6)	912 (3.3)	0.012
Cyclophosphamide (%)	0 (0.0)	0 (0.0)	4 (0.0)	0.011

Abbreviations: ABA: abatacept; TCZ: tocilizumab; TNF: tumor necrosis factor; SMD: absolute standardized mean difference; MTX: methotrexate; CRP: C-reactive protein; NSAID: non-steroidal anti-inflammatory drug; PSA: prostate-specific antigen; ED: emergency department; HbA1c: hemoglobin A1c; TIA: transient ischemic attack; CT: computed tomography; MI: myocardial infarction; HPV: human papilloma virus.

eTable 3-3. Patient group characteristics in the three biological DMARDs example (second-line bDMARDs).

Variable	ABA	TCZ	TNF	SMD
n	475	187	1277	
Methotrexate (%)	325 (68.4)	127 (67.9)	947 (74.2)	0.092
NSAIDs and Coxibs (%)	178 (37.5)	66 (35.3)	526 (41.2)	0.081
Prior Use of Oral Glucocorticoids (yes/no) (%)	335 (70.5)	125 (66.8)	916 (71.7)	0.071
Number of Unique Prescription Medications (mean (sd))	11.66 (8.22)	11.01 (8.35)	12.48 (7.67)	0.121
Recent use of Oral Glucocorticoids (yes/no) (%)	306 (64.4)	109 (58.3)	831 (65.1)	0.093
Hydroxychloroquine (HCQ) (%)	200 (42.1)	60 (32.1)	450 (35.2)	0.139
Prior Use of Oral Glucocorticoids (mean (sd))	207.21 (161.90)	199.22 (162.96)	201.55 (160.56)	0.033
Count Combined Biologics and NonBiologics including MTX (categorized) (mean (sd))	1.37 (0.73)	1.26 (0.79)	1.37 (0.70)	0.105
Age (mean (sd))	54.55 (12.52)	55.67 (12.93)	52.82 (13.09)	0.147
Number of Rheumatologist Visits (mean (sd))	10.57 (14.88)	16.28 (23.57)	8.57 (14.49)	0.273
Sulfasalazine (%)	41 (8.6)	10 (5.3)	127 (9.9)	0.116
CIRAS Score (mean (sd))	6.05 (1.93)	6.12 (2.02)	6.22 (1.88)	0.058
Number of lipid tests ordered (mean (sd))	0.72 (1.18)	1.00 (2.03)	0.73 (1.45)	0.109
CHADS2 score, 365 days (mean (sd))	0.78 (0.97)	0.83 (1.03)	0.69 (0.92)	0.094
Number of outpatient visits (mean (sd))	13.71 (8.33)	13.84 (8.26)	12.77 (8.57)	0.085
Opioids (%)	301 (63.4)	113 (60.4)	839 (65.7)	0.073
Combined Comorbidity Index (365 days) (mean (sd))	0.45 (1.33)	0.48 (1.60)	0.35 (1.15)	0.067
Recent use of Oral Glucocorticoids (mean (sd))	136.42 (145.06)	122.73 (145.09)	122.43 (133.52)	0.066
Region (%)				0.115
Northeast	91 (19.2)	39 (20.9)	215 (16.8)	
North Central	92 (19.4)	41 (21.9)	278 (21.8)	
South	206 (43.4)	74 (39.6)	562 (44.0)	
West	79 (16.6)	28 (15.0)	198 (15.5)	
Unknown	7 (1.5)	5 (2.7)	24 (1.9)	
Female (%)	386 (81.3)	151 (80.7)	1000 (78.3)	0.049
Heart Failure (covariate) (%)	18 (3.8)	8 (4.3)	20 (1.6)	0.108
Antidepressants (%)	166 (34.9)	73 (39.0)	455 (35.6)	0.057
Coronary Artery Disease (CAD) (%)	50 (10.5)	20 (10.7)	115 (9.0)	0.038
Renal Dysfunction (%)	27 (5.7)	10 (5.3)	51 (4.0)	0.053
Any hospitalization (%)	75 (15.8)	29 (15.5)	205 (16.1)	0.010
Number of Cardiologist visits (mean (sd))	1.27 (3.11)	0.71 (3.17)	0.81 (2.63)	0.124
Number of CRP Tests Ordered (mean (sd))	1.70 (2.10)	2.05 (3.23)	1.63 (1.90)	0.106
Hypertension (%)	268 (56.4)	106 (56.7)	663 (51.9)	0.064
Stroke or TIA (events occurred) (%)	12 (2.5)	2 (1.1)	18 (1.4)	0.074
Number of Serum Creatinine Tests Ordered (mean (sd))	3.33 (3.05)	3.63 (3.84)	3.08 (3.02)	0.108
Calcium Channel Blockers (%)	55 (11.6)	22 (11.8)	177 (13.9)	0.046
Cyclosporine (%)	11 (2.3)	0 (0.0)	42 (3.3)	0.179

eTable 3-3 (Continued)

Variable	ABA	TCZ	TNF	SMD
Hyperlipidemia (%)	158 (33.3)	69 (36.9)	371 (29.1)	0.112
Angiotensin Converting Enzyme inhibitors (ACEs) (%)	76 (16.0)	35 (18.7)	250 (19.6)	0.062
Leflunomide (%)	73 (15.4)	24 (12.8)	207 (16.2)	0.064
Diabetes (%)	86 (18.1)	33 (17.6)	214 (16.8)	0.024
PSA test (%)	23 (4.8)	9 (4.8)	77 (6.0)	0.036
Atrial Fibrillation (%)	12 (2.5)	7 (3.7)	24 (1.9)	0.076
Peripheral Vascular Disease (%)	8 (1.7)	1 (0.5)	19 (1.5)	0.074
ED visit (yes/no) (%)	27 (5.7)	8 (4.3)	74 (5.8)	0.046
Chronic Liver Disease (%)	35 (7.4)	12 (6.4)	52 (4.1)	0.095
Nitrates (%)	12 (2.5)	7 (3.7)	41 (3.2)	0.047
Number of ED visits (mean (sd))	0.20 (0.97)	0.16 (0.81)	0.29 (1.47)	0.073
Beta Blockers (%)	94 (19.8)	33 (17.6)	233 (18.2)	0.037
Recent hospitalization (60 days) (%)	10 (2.1)	2 (1.1)	33 (2.6)	0.076
Preventive screening test (at least once); (%)	234 (49.3)	90 (48.1)	615 (48.2)	0.015
Myocardial Infraction (%)	1 (0.2)	0 (0.0)	2 (0.2)	0.045
Number of HbA1c tests ordered (mean (sd))	0.34 (0.83)	0.27 (0.66)	0.35 (0.87)	0.068
Antiplatelet agents (%)	24 (5.1)	6 (3.2)	44 (3.4)	0.062
Sum of Daily dose of oral steroids- oral prednisone mg equivalents (mean (sd))	905.53 (1760.44)	911.18 (1691.07)	1057.57 (3848.50)	0.034
Mammogram (%)	145 (30.5)	47 (25.1)	311 (24.4)	0.092
Lipid lowering drugs (%)	109 (22.9)	45 (24.1)	302 (23.6)	0.018
Number of PCP (primary care provider) visits (mean (sd))	6.11 (11.83)	5.31 (9.94)	6.00 (13.05)	0.047
Pap smear (%)	107 (22.5)	40 (21.4)	288 (22.6)	0.019
Auranofin = FALSE (%)	475 (100.0)	187 (100.0)	1277 (100.0)	<0.001
Use of statins (%)	101 (21.3)	39 (20.9)	266 (20.8)	0.007
Azathioprine (%)	12 (2.5)	3 (1.6)	32 (2.5)	0.043
Fecal occult blood (FOB) test (%)	41 (8.6)	11 (5.9)	85 (6.7)	0.071
Non-insulin diabetes medications (%)	48 (10.1)	15 (8.0)	118 (9.2)	0.048
Flexible Sigmoidoscopy or colonoscopy or CT virtual colonoscopy (%)	43 (9.1)	17 (9.1)	144 (11.3)	0.049
Gold Sodium Thiomalate = FALSE (%)	475 (100.0)	187 (100.0)	1277 (100.0)	<0.001
Angiotensin II Receptor Blockers (ARBs) (%)	53 (11.2)	21 (11.2)	137 (10.7)	0.011
HPV DNA test (%)	41 (8.6)	13 (7.0)	90 (7.0)	0.042
Use of other lipid-lowering drugs (%)	19 (4.0)	10 (5.3)	63 (4.9)	0.043
Diuretics (%)	79 (16.6)	34 (18.2)	189 (14.8)	0.061
Insulin (%)	19 (4.0)	8 (4.3)	58 (4.5)	0.018
Cyclophosphamide (%)	0 (0.0)	0 (0.0)	1 (0.1)	0.026

Abbreviations: ABA: abatacept; TCZ: tocilizumab; TNF: tumor necrosis factor; SMD: absolute standardized mean difference; MTX: methotrexate; CRP: C-reactive protein; NSAID: non-steroidal anti-inflammatory drug; PSA: prostate-specific antigen; ED: emergency department; HbA1c: hemoglobin A1c; TIA: transient ischemic attack; CT: computed tomography; MI: myocardial infarction; HPV: human papilloma virus.

Contents

1 Two-group definitions

1.1 Preference score definition

Let $\mathbf{X}_i$ be a covariate vector and A_i be a binary treatment indicator. Then, $e_i = E[A_i|\mathbf{X}_i]$ is the propensity score. Its expectation is the treatment prevalence by iterative expectation $p = E[e_i] = E[E[A_i|\mathbf{X}_i]] = E[A_i]$. [Walker et al., 2013] defined the *preference score* as π_i that satisfied the following relationship.

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \log\left(\frac{e_i}{1-e_i}\right) - \log\left(\frac{p}{1-p}\right)$$

If we solve for π_i, we can obtain the following.

$$\begin{aligned}
\log\left(\frac{\pi_i}{1-\pi_i}\right) &= \log\left(\frac{e_i}{1-e_i}\right) - \log\left(\frac{p}{1-p}\right)\\
\log\left(\frac{\pi_i}{1-\pi_i}\right) &= \log\left(\frac{e_i}{1-e_i}\bigg/\frac{p}{1-p}\right)\\
\frac{\pi_i}{1-\pi_i} &= \frac{e_i}{1-e_i}\bigg/\frac{p}{1-p}\\
&= \frac{\frac{e_i}{p}}{\frac{1-e_i}{1-p}}\\
\pi_i &= \frac{\frac{\frac{e_i}{p}}{\frac{1-e_i}{1-p}}}{1+\frac{\frac{e_i}{p}}{\frac{1-e_i}{1-p}}}\\
&= \frac{\frac{e_i}{p}}{\frac{1-e_i}{1-p}+\frac{e_i}{p}}
\end{aligned}$$

This form gives insight into its re-centering property. When the treatment is rare, e_i is generally small. The numerator $\frac{e_i}{p}$ corrects this by dividing the generally small e_i with a small p. In particular, those individuals who happen to have the mean PS, *i.e.*, $e_i = p$, receive $pi_i = 0.5$. This transformation brings the "average individuals" to the center of the scale.

Also if we solve for e_i, we can obtain the following.

$$\begin{aligned}
\frac{\pi_i}{1-\pi_i} &= \frac{e_i}{1-e_i}\bigg/\frac{p}{1-p}\\
\frac{\pi_i p}{(1-\pi_i)(1-p)} &= \frac{e_i}{1-e_i}\\
e_i &= \frac{\frac{\pi_i p}{(1-\pi_i)(1-p)}}{1+\frac{\pi_i p}{(1-\pi_i)(1-p)}}\\
&= \frac{\pi_i p}{(1-\pi_i)(1-p)+\pi_i p}
\end{aligned}$$

1.2 Intercept-adjustment interpretation of the preference score

$$\text{Note}$$
$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \log\left(\frac{e_i}{1-e_i}\right) - \log\left(\frac{p}{1-p}\right)$$

$$= \log\left(\frac{P[A_i = 1|\mathbf{X}_i]}{1 - P[A_i = 1|\mathbf{X}_i]}\right) - \log\left(\frac{P[A_i = 1]}{1 - P[A_i = 1]}\right)$$

Assuming logistic models

$$\log\left(\frac{P[A_i = 1|\mathbf{X}_i]}{1 - P[A_i = 1|\mathbf{X}_i]}\right) = \alpha_0 + \mathbf{X}_i^T \boldsymbol{\alpha}_{\mathbf{x}}$$

$$\log\left(\frac{\pi_i}{1 - \pi_i}\right) = \left[\alpha_0 - \log\left(\frac{p}{1-p}\right)\right] + \mathbf{X}_i^T \boldsymbol{\alpha}_{\mathbf{x}}$$

The last expression has the same form as the intercept-adjusted logistic regression used for risk prediction from a logistic regression fit on a case-control dataset [Hosmer et al., 2013]. It is known that a case-control logistic regression and the corresponding cohort logistic regression give the same coefficients except for the intercepts [Prentice and Pyke, 1979, Scott and Wild, 1986]. The intercept terms have the following relationship.

$$\alpha_0^{cohort} = \alpha_0^{case-control} - \log\left(\frac{\tau_1}{\tau_0}\right)$$

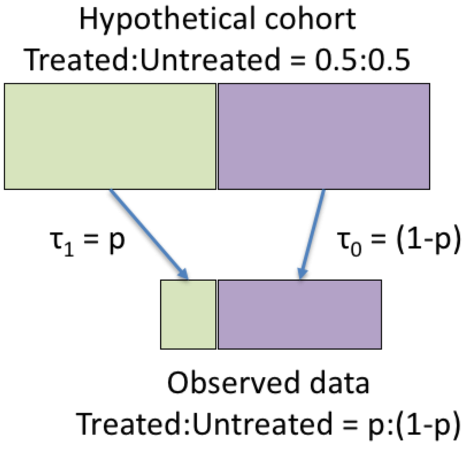
$$\text{where}$$

$$\tau_1 = \text{case sampling fraction}$$

$$\tau_0 = \text{control sampling fraction}$$

Intuitively, the case-control intercept is an overestimate because of the artificially high case prevalence in the case-control data. $\log\left(\frac{\tau_1}{\tau_0}\right) > 0$ if we oversample cases ($\tau_1 > \tau_0$).

We can consider the current study with a marginal treatment prevalence of p is a biased sample from a hypothetical population in which the covariate effects on the logit of treatment $\boldsymbol{\alpha}_{\mathbf{x}}$ are preserved but the marginal treatment prevalence is 0.5. The sampling fraction for the treated would be $\tau_1 = p$ and the sampling fraction for the untreated would be $\tau_1 = 1 - p$. We would obtain the desired ratio because $\frac{0.5p}{0.5(1-p)} = \frac{p}{1-p}$.



Under this framework, the initial PS model is the treatment assignment model for the biased sample with a treatment prevalence of p. The preference score model is the treatment assignment model for the super-population with a treatment prevalence of 0.5.

When the covariates have no role in determining treatment assignment (random treatment assignment), the right-hand side is always zero (preference score of 0.5) [Walker et al., 2013] because $P[A_i = 1|\mathbf{X}_i] = P[A_i = 1]$.

2 Multi-group definitions

2.1 Generalized preference score

Each generalized preference score is the following.

$$\pi_{ji} = \frac{\frac{e_{ji}}{p_j}}{\sum_{k=0}^{J} \frac{e_{ki}}{p_k}}$$

This expression came from the following proposed generalization of the defining equations (J simultaneous equations) using the baseline logit multinomial logistic regression in place of the binary logistic regression in the two-group definition.

$$\begin{gathered}\text{For } j \in \{1, ..., J\} \\ \log\left(\frac{\pi_{ji}}{\pi_{0i}}\right) = \log\left(\frac{e_{ji}}{e_{0i}}\right) - \log\left(\frac{p_j}{p_0}\right) \\ \text{where} \\ \sum_{k=0}^{J} \pi_{ki} = 1\end{gathered}$$

The sum constraint is necessary to maintain the interpretation as the prevalence-adjusted PS. For each $j \in \{1, ..., J\}$, we have the following.

$$\begin{aligned}\log\left(\frac{\pi_{ji}}{\pi_{0i}}\right) &= \log\left(\frac{e_{ji}}{e_{0i}}\right) - \log\left(\frac{p_j}{p_0}\right) \\ &= \log\left(\frac{e_{ji}}{e_{0i}}\bigg/\frac{p_j}{p_0}\right) \\ \frac{\pi_{ji}}{\pi_{0i}} &= \frac{e_{ji}}{e_{0i}}\bigg/\frac{p_j}{p_0} \\ &= \frac{e_{ji}}{p_j}\frac{p_0}{e_{0i}}\end{aligned}$$

First solve for π_{0i}.

$$\begin{gathered}\text{Sum } J \text{ equations} \\ \sum_{j=1}^{J} \frac{\pi_{ji}}{\pi_{0i}} = \sum_{j=1}^{J} \frac{e_{ji}}{p_j}\frac{p_0}{e_{0i}} \\ \frac{\sum_{j=1}^{J} \pi_{ji}}{\pi_{0i}} = \sum_{j=1}^{J} \frac{e_{ji}}{p_j}\frac{p_0}{e_{0i}} \\ \text{By } \sum_{j=0}^{J} \pi_{ji} = 1\end{gathered}$$

$$\begin{aligned}
\frac{1-\pi_{0i}}{\pi_{0i}} &= \sum_{j=1}^{J} \frac{e_{ji}}{p_j}\frac{p_0}{e_{0i}} \\
\frac{\pi_{0i}}{1-\pi_{0i}} &= \frac{1}{\sum\limits_{j=1}^{J} \frac{e_{ji}}{p_j}\frac{p_0}{e_{0i}}} \\
\pi_{0i} &= \frac{\frac{1}{\sum_{j=1}^{J} \frac{e_{ji}}{p_j}\frac{p_0}{e_{0i}}}}{1+\frac{1}{\sum_{j=1}^{J} \frac{e_{ji}}{p_j}\frac{p_0}{e_{0i}}}} \\
&= \frac{1}{1+\sum\limits_{j=1}^{J} \frac{e_{ji}}{p_j}\frac{p_0}{e_{0i}}} \\
&= \frac{\frac{e_{0i}}{p_0}}{\frac{e_{0i}}{p_0}+\sum\limits_{j=1}^{J} \frac{e_{ji}}{p_j}} \\
&= \frac{\frac{e_{0i}}{p_0}}{\sum\limits_{j=0}^{J} \frac{e_{ji}}{p_j}}
\end{aligned}$$

Now solve for an arbitrary $j \in \{1, ..., J\}$.

$$\begin{aligned}
\frac{\pi_{ji}}{\pi_{0i}} &= \frac{e_{ji}}{p_j}\frac{p_0}{e_{0i}} \\
\pi_{ji} &= \pi_{0i}\frac{e_{ji}}{p_j}\frac{p_0}{e_{0i}} \\
&= \pi_{0i}\frac{e_{ji}}{p_j}\frac{p_0}{e_{0i}} \\
&\text{Substitute } \pi_{0i} \\
&= \frac{\frac{e_{0i}}{p_0}}{\sum\limits_{k=0}^{J} \frac{e_{ki}}{p_k}}\frac{e_{ji}}{p_j}\frac{p_0}{e_{0i}} \\
&= \frac{1}{\sum\limits_{k=0}^{J} \frac{e_{ki}}{p_k}}\frac{e_{ji}}{p_j} \\
&= \frac{\frac{e_{ji}}{p_j}}{\sum\limits_{k=0}^{J} \frac{e_{ki}}{p_k}}
\end{aligned}$$

Taken together, for $j \in \{0, 1, ..., J\}$,

$$\pi_{ji} = \frac{\frac{e_{ji}}{p_j}}{\sum\limits_{k=0}^{J} \frac{e_{ki}}{p_k}}$$

2.2 Rationale for region of empirical equipoise

By the proposed generalization, each subject has a *preference score vector* $\boldsymbol{\pi}_i$ with $J+1$ elements π_{ji} where $j = 0, 1, ..., J$ and $\sum_{j=0}^{J+1} \pi_{ji} = 1$. Note the expectation of the corresponding *propensity score vector* $\mathbf{e}_i$ is the *treatment prevalence vector* $\mathbf{p}$ ($E[\mathbf{e}_i] = \mathbf{p}$).

# of Groups	Preference score space	Center of preference score space	Threshold
2	$[0,1]^2$	$\left(\frac{1}{2}, \frac{1}{2}\right)^T$	0.30
3	$[0,1]^3$	$\left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)^T$	0.20
4	$[0,1]^4$	$\left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right)^T$	0.15
5	$[0,1]^5$	$\left(\frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5}\right)^T$	0.12
6	$[0,1]^6$	$\left(\frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}\right)^T$	0.10
$\vdots$			
$J+1$	$[0,1]^{J+1}$	$\left(\frac{1}{J+1}, \dots, \frac{1}{J+1}\right)^T$	$\left(\frac{1}{J+1}\right)\left(\frac{3}{5}\right)$

An "average" individual with a PS vector agreeing with the treatment prevalence vector is given a preference score vector $\left(\frac{1}{J+1}, \dots, \frac{1}{J+1}\right)^T$. This is $\left(\frac{1}{2}, \frac{1}{2}\right)^T$ in the two-group setting, $\left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)^T$ in the three-group setting, and $\left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right)^T$ in the four-group setting.

Because of this change in the center of the preference score space, the threshold for defining the region for empirical equipoise assessment must adapt to the number of group. For example, the threshold of $\pi_{ji} > 0.3$ for all $j \in \{0, \dots, J\}$ is not possible once there are four groups.

2.3 Proof that the generalized definition reduces to the original two-group definition

We can check this definition reduces to the original definition in the two-group setting as follows.

Preference score is recovered as follows.

$$\begin{aligned}
\log\left(\frac{\pi_{1i}}{\pi_{0i}}\right) &= \log\left(\frac{e_{1i}}{e_{0i}}\right) - \log\left(\frac{p_1}{p_0}\right) \\
\log\left(\frac{\pi_{1i}}{1-\pi_{1i}}\right) &= \log\left(\frac{e_{1i}}{1-e_{1i}}\right) - \log\left(\frac{p_1}{1-p_1}\right) \\
\log\left(\frac{\pi_{i}}{1-\pi_{i}}\right) &= \log\left(\frac{e_{i}}{1-e_{i}}\right) - \log\left(\frac{p}{1-p}\right)
\end{aligned}$$

Let $I = \{1, ..., n\}$ be the set of indices for n individuals in the entire cohort and $\alpha_{J,w}$ be the threshold proposed above. The index set for the individuals in the region of empirical equipoise is the following for the $J+1$ group setting.

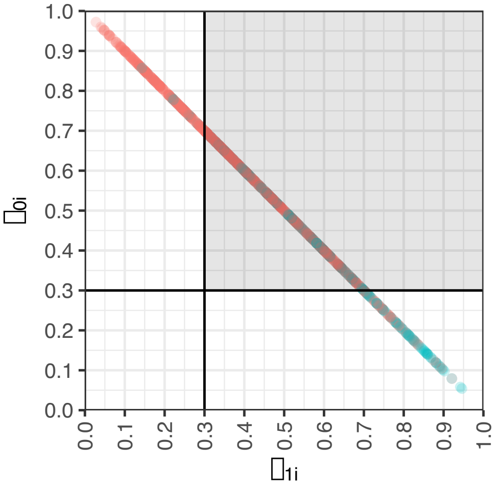
$$I_{J,w} = \{i \in I : \pi_{ji} \geq \alpha_{J,w} \ \forall \ j \in \{0, ..., J\}\}$$

We can show this expression reduces to the original two-group definition for $J = 1$ (two group setting).

$$\begin{aligned}
I_{1,w} &= \{i \in I : \pi_{ji} \geq \alpha_{J,w} \ \forall \ j \in \{0, 1\}\} \\
&= \{i \in I : \pi_{0i} \geq \alpha_{J,w}, \pi_{1i} \geq \alpha_{J,w}\} \\
&\quad \text{Since } \pi_{0i} = 1 - \pi_{1i} \\
&= \{i \in I : 1 - \pi_{1i} \geq \alpha_{J,w}, \pi_{1i} \geq \alpha_{J,w}\}
\end{aligned}$$

$$
\begin{aligned}
&= \{i \in I : \pi_{1i} \leq 1 - \alpha_{J,w}, \pi_{1i} \geq \alpha_{J,w}\} \\
&= \{i \in I : \alpha_{J,w} \leq \pi_{1i} \leq 1 - \alpha_{J,w}\} \\
&= \{i \in I : \pi_{1i} \in [\alpha_{1,w}, 1 - \alpha_{1,w}]\} \\
&\quad \text{Note } \pi_{1i} = \pi_i \text{ (two-group preference score).} \\
&\quad \alpha_{1,w} = 0.3 \\
&= \{i \in I : \pi_i \in [0.3, 0.7]\} \\
&= \text{ original two-group definition}
\end{aligned}
$$

If we visualize the two-group preference scores, we obtain a two-dimensional plot. However, because of the constraint that $\pi_{1i} + \pi_{0i} = 1$, all individuals (red group 0; blue group 1) appear on the diagonal line. That is, the information is one-dimensional, so we only need $\pi_i = \pi_{1i}$. With this visualization, we can see that individuals satisfy $\pi_{1i} \geq 0.3$ and $\pi_{0i} \geq 0.3$ (gray region) if and only if they satisfy $\pi_i = \pi_{1i} \in [0.3, 0.7]$.



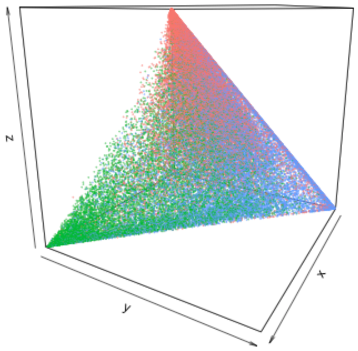
3 Empirical data demonstration

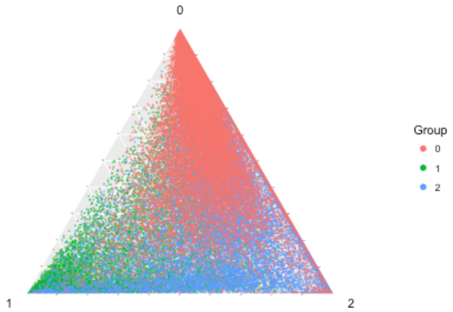
3.1 Visualization with a ternary plot

The generalized propensity score in the three-group setting is a vector of three elements $(e_{0i}, e_{1i}, e_{2i})^T$. The generalized preference score in the three-group setting is also a vector of three elements $(\pi_{0i}, \pi_{1i}, \pi_{2i})^T$. The following explanation is written in terms of the generalized propensity score, but the explanation is analogous for the generalized preference score.

As three dimensional data, individual subjects can be plotted in a three-dimensional cube $[0, 1]^3$ (left). The Z-axis represents e_{0i}, X-axis represents e_{1i}, and Y-axis represents e_{2i}. As seen in the three-dimensional plot (left), the points only occupy the diagonal triangular plane. This is because of the constraint $e_{0i}+e_{1i}+e_{2i} = 1$ for all i. In this case, we know what e_{2i} is as soon as we know e_{0i} and e_{1i}. That is, although the data are three-dimensional, the information carried is only two dimensional.

Therefore, we can take out this triangular plane in the left plot and represent as a two-dimensional plot (right). This two-dimensional representation is called a *ternary plot*. We used the `ggtern` R package for ternary plots [Hamilton, 2017].

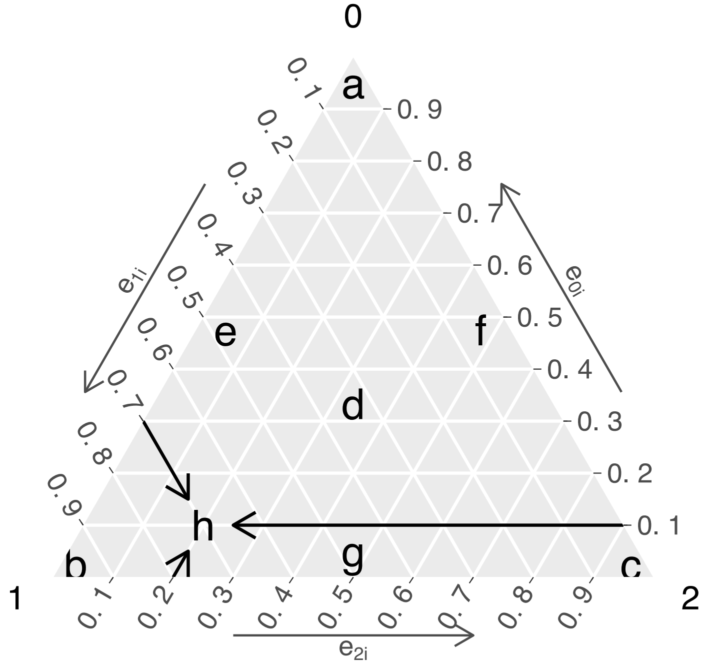




The coordinate systems is explained here. The top corner of the triangle (a) is $\mathbf{e}_i = (1, 0, 0)$, *i.e.*, 100% probability of being in Group 0. The left lower corner (b) is $\mathbf{e}_i = (0, 1, 0)$ and the right lower corner (c) is $\mathbf{e}_i = (0, 0, 1)$. The mid-point in the triangle (d) is $\mathbf{e}_i = (1/3, 1/3, 1/3)$. That is, equal probability of being in any of the three groups. The mid points on the edges are: (e) $\mathbf{e}_i = (1/2, 1/2, 0)$, (f) $\mathbf{e}_i = (1/2, 0, 1/2)$, and (g) $\mathbf{e}_i = (0, 1/2, 1/2)$.

To look up point (h), all three axes have to be looked up. The e_{0i} axis is on the right edge. Use the horizontal guide lines because the labels (0.1, etc) are horizontal. Point (h) is at $e_{0i} = 0.1$. The e_{1i} axis is on the left edge. Use the guide lines going into the lower right direction as the labels indicate. Point (h) is at $e_{1i} = 0.7$. The e_{2i} axis is on the bottom edge. Use the guide lines going into the upper right direction as the labels indicate. Point (h) is at $e_{2i} = 0.2$. As a result, Point (h) is at $\mathbf{e}_i = (0.1, 0.7, 0.2)$.

We omitted the axis labels in the empirical examples since we did not need precise value lookup. The general intuition is that being far from a given corner, for example, the top corner labeled 0, means having a low probability of being in that group.



	e_{0i}	e_{1i}	e_{2i}
(a)	1	0	0
(b)	0	1	0
(c)	0	0	1
(d)	$1/3$	$1/3$	$1/3$
(e)	$1/2$	$1/2$	0
(f)	$1/2$	0	$1/2$
(g)	0	$1/2$	$1/2$
(h)	0.1	0.7	0.2

3.2 Non-selective non-steroidal anti-inflammatory drugs (nsNSAIDs) example

This dataset contained demographic and clinical including dispensing information on Medicare beneficiaries from Pennsylvania and New Jersey who qualified for pharmaceutical assistance programs for low-income older adults (January 1, 1999, through December 31, 2005) [Solomon et al., 2010].

Individuals were required to have diagnoses for osteoarthritis or rheumatoid arthritis on two separate occasions and consistent use of health care services in the preceding 365 days. Those who had dispensing of analgesics within the preceding 180 days, those with malignancy, those using hospice services within the preceding 365 days, and those had simultaneous dispensing of multiple analgesics were secluded. The outcomes of interest of the original study included cardiovascular and gastrointestinal adverse events.

We chose three non-selective NSAIDs with different prevalence in the dataset for visual examination: naproxen, ibuprofen, and diclofenac. These non-selective NSAIDs were expected to have been used similarly in practice. This example was used to illustrate the centering property of the generalized preference score in the presence of groups of different sizes. Generalized PSs were estimated with 38 predictor variables thought to be risk factors for any of several potential adverse effects of nsNSAIDs (**eTable 1** and **eFigure 1**).

3.3 Biological disease-modifying anti-rheumatic drugs (bDMARDs) example

This example was taken from more recent MarketScan data (2011-June 2015) of new users of biological disease-modifying anti-rheumatic drugs (DMARDs) [Kim et al., 2017, Kang et al., 2018]. In the original studies, [Kim et al., 2017] examined the tocilizumab vs tumor necrosis factor (TNF) inhibitor comparison and [Kang et al., 2018] examined the abatacept TNF inhibitor comparison. Both studies used multiple data sources, but we focused on the MarketScan data for simplicity. Our three arms of interests were abatacept users, tocilizumab users, and TNF inhibitor users. Therefore, we re-extracted the datasets and combined such that we have three mutually exclusive groups.

Individuals were required to have two separate outpatient or one inpatient code for rheumatoid arthritis and initiation of the drugs of interest. The exclusion criteria were nursing home residents, patients with HIV/AIDS, patients with malignancy other than nonmelanoma skin cancer, and those with end-stage renal disease including use of dialysis or renal transplant. The outcome of interest of the original studies was composite cardiovascular events.

The most up-to-date recommendations list these three classes of bDMARDs as equally indicated [Singh et al., 2016, Smolen et al., 2017]. However, TNF inhibitors, by the virtue of being the first biological DMARDs to come on the market, were more often used first. On the other hand, tocilizumab and abatacept were market-approved more recently in the U.S. market, and thus, were more commonly used as subsequent biological DMARDs after failure of one or more biological DMARDs.

Therefore, first-line tocilizumab and abatacept users were expected to be somewhat special patients compared to first-line TNFi users, whereas users were expected to be more similar when using these agents as a second-line bDMARD. A second-line TNFi after one TNFi means that there was a switch from one specific agent to another within the five-member TNFi class (adalimumab, certolizumab pegol, etanercept, golimumab, infliximab).

4 Simulation: methodological details

4.1 Data generating mechanism

In all scenarios, our sample size was n = 6,000.

4.1.1 Covariate generation

Latent covariates Z_{1i} through Z_{7i} were generated from a multivariate normal distribution to induce a given level of correlation $\rho \in \{0, 0.1, 0.3, 0.5, 0.7, 0.9\}$.

$$\begin{bmatrix} Z_{1i} \\ Z_{2i} \\ Z_{3i} \\ Z_{4i} \\ Z_{5i} \\ Z_{6i} \\ Z_{7i} \end{bmatrix} \sim \left(\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho^1 & \rho^2 & \rho^3 & \rho^4 & \rho^5 & \rho^6 \\ \rho^1 & 1 & \rho^1 & \rho^2 & \rho^3 & \rho^4 & \rho^5 \\ \rho^2 & \rho^1 & 1 & \rho^1 & \rho^2 & \rho^3 & \rho^4 \\ \rho^3 & \rho^2 & \rho^1 & 1 & \rho^1 & \rho^2 & \rho^3 \\ \rho^4 & \rho^3 & \rho^2 & \rho^1 & 1 & \rho^1 & \rho^2 \\ \rho^5 & \rho^4 & \rho^3 & \rho^2 & \rho^1 & 1 & \rho^1 \\ \rho^6 & \rho^5 & \rho^4 & \rho^3 & \rho^2 & \rho^1 & 1 \end{bmatrix} \right)$$

This means each Z_{ji} was a standard normal marginally. The correlation of Z_{ji} and Z_{ki} for $j \neq k$ was $\rho^{|j-k|}$. These latent variables were then transformed as follows.

$$\begin{aligned} X_{1i} &:= Z_{1i} \\ X_{2i} &:= F^{-1}_{Pois,1}(\Phi(Z_{2i})) \\ X_{3i} &:= F^{-1}_{Bern,0.2}(\Phi(Z_{3i})) \\ X_{4i} &:= F^{-1}_{Bern,0.2}(\Phi(Z_{4i})) \\ X_{5i} &:= F^{-1}_{Bern,0.2}(\Phi(Z_{5i})) \\ X_{6i} &:= F^{-1}_{Bern,0.2}(\Phi(Z_{6i})) \\ X_{7i} &:= Z_{7i} \end{aligned}$$

$\Phi(\cdot)$ was the standard normal cumulative distribution function (`pnorm(x, mean = 0, sd = 1)` in `R`). $F^{-1}_{Pois,1}(\cdot)$ was the inverse distribution function for a Poisson distribution with a rate parameter of 1 (qpois(p, lambda = 1) in R). $F^{-1}_{Bern,0.2}(\cdot)$ was the inverse distribution function for a Bernoulli distribution with a success probability of 0.2 (qbinom(p, size = 1, prob = 0.2) in R). The first transformation gave a Uniform(0,1) variable, and the second transformation gave a random variable with the desired distribution. The correlation structure was preserved in X_{1i} through X_{7i} using this two-step covariate generation.

4.1.2 Treatment generation

Treatment A_i was assigned based on all covariates $\mathbf{X}_i = (X_{1i}, \dots, X_{7i})^T$.

Linear predictors

$$\begin{cases} \eta_{A1i} = \log\left(\dfrac{P[A_i = 1|\mathbf{X}_i]}{P[A_i = 0|\mathbf{X}_i]}\right) = \alpha_{01} + \mathbf{X}_i^T \boldsymbol{\alpha}_{X1} \\ \eta_{A2i} = \log\left(\dfrac{P[A_i = 2|\mathbf{X}_i]}{P[A_i = 0|\mathbf{X}_i]}\right) = \alpha_{02} + \mathbf{X}_i^T \boldsymbol{\alpha}_{X2} \end{cases}$$

True propensity scores

$$\begin{cases} e_{0i} = P(A_i = 0|\mathbf{X}_i) = \dfrac{1}{1 + \exp(\eta_{A1i}) + \exp(\eta_{A2i})} \\ e_{1i} = P(A_i = 1|\mathbf{X}_i) = \dfrac{\exp(\eta_{A1i})}{1 + \exp(\eta_{A1i}) + \exp(\eta_{A2i})} \\ e_{2i} = P(A_i = 2|\mathbf{X}_i) = \dfrac{\exp(\eta_{A2i})}{1 + \exp(\eta_{A1i}) + \exp(\eta_{A2i})} \end{cases}$$

Treatment assignment

$$A_i \in \{0, 1, 2\} \sim \text{Multinomial}\left((e_{0i}, e_{1i}, e_{2i})^T, 1\right)$$

The treatment model parameter values are in the following table.

- The Size column is the treatment prevalence setting.
- The "RelX7" column corresponds to the "Relative treatment association of X7" in the figures, the strength of the treatment association of X_7 relative to X_1 through X_6.
- The "Equipoise" column corresponds to the "Level of equipoise" in the figures. "Perfect" indicates no covariate effect on treatments (randomized treatment). Increasing levels of covariate effects were introduced for "Good", "Moderate", and "Poor" as seen in the magnitude of coefficients.
- The alternating rows correspond to the first and second linear predictors (See Contrast column).
- Column 0 corresponds to the intercept coefficient. Columns 1 through 7 correspond to the coefficients for X_1 through X_7.

Number	Size	RelX7	Equipoise	Contrast	0	1	2	3	4	5	6	7
1	33:33:33	Zero	Perfect	1vs0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1	33:33:33	Zero	Perfect	2vs0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
3	33:33:33	Zero	Good	1vs0	-0.34	0.25	0.25	0.25	0.25	0.25	0.25	0.00
3	33:33:33	Zero	Good	2vs0	-1.00	0.50	0.50	0.50	0.50	0.50	0.50	0.00
4	33:33:33	Zero	Moderate	1vs0	-0.40	0.50	0.50	0.50	0.50	0.50	0.50	0.00
4	33:33:33	Zero	Moderate	2vs0	-1.70	1.00	1.00	1.00	1.00	1.00	1.00	0.00
5	33:33:33	Zero	Poor	1vs0	-0.40	1.00	1.00	1.00	1.00	1.00	1.00	0.00

Continued

Number	Size	RelX7	Equipoise	Contrast	0	1	2	3	4	5	6	7
5	33:33:33	Zero	Poor	2vs0	-3.10	2.00	2.00	2.00	2.00	2.00	2.00	0.00
6	33:33:33	Half	Perfect	1vs0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
6	33:33:33	Half	Perfect	2vs0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
8	33:33:33	Half	Good	1vs0	-0.34	0.25	0.25	0.25	0.25	0.25	0.25	0.12
8	33:33:33	Half	Good	2vs0	-1.00	0.50	0.50	0.50	0.50	0.50	0.50	0.25
9	33:33:33	Half	Moderate	1vs0	-0.40	0.50	0.50	0.50	0.50	0.50	0.50	0.25
9	33:33:33	Half	Moderate	2vs0	-1.70	1.00	1.00	1.00	1.00	1.00	1.00	0.50
10	33:33:33	Half	Poor	1vs0	-0.40	1.00	1.00	1.00	1.00	1.00	1.00	0.50
10	33:33:33	Half	Poor	2vs0	-3.10	2.00	2.00	2.00	2.00	2.00	2.00	1.00
11	33:33:33	Same	Perfect	1vs0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
11	33:33:33	Same	Perfect	2vs0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
13	33:33:33	Same	Good	1vs0	-0.34	0.25	0.25	0.25	0.25	0.25	0.25	0.25
13	33:33:33	Same	Good	2vs0	-1.00	0.50	0.50	0.50	0.50	0.50	0.50	0.50
14	33:33:33	Same	Moderate	1vs0	-0.40	0.50	0.50	0.50	0.50	0.50	0.50	0.50
14	33:33:33	Same	Moderate	2vs0	-1.70	1.00	1.00	1.00	1.00	1.00	1.00	1.00
15	33:33:33	Same	Poor	1vs0	-0.40	1.00	1.00	1.00	1.00	1.00	1.00	1.00
15	33:33:33	Same	Poor	2vs0	-3.10	2.00	2.00	2.00	2.00	2.00	2.00	2.00
16	33:33:33	Twice	Perfect	1vs0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
16	33:33:33	Twice	Perfect	2vs0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
18	33:33:33	Twice	Good	1vs0	-0.34	0.25	0.25	0.25	0.25	0.25	0.25	0.50
18	33:33:33	Twice	Good	2vs0	-1.00	0.50	0.50	0.50	0.50	0.50	0.50	1.00
19	33:33:33	Twice	Moderate	1vs0	-0.40	0.50	0.50	0.50	0.50	0.50	0.50	1.00
19	33:33:33	Twice	Moderate	2vs0	-1.70	1.00	1.00	1.00	1.00	1.00	1.00	2.00
20	33:33:33	Twice	Poor	1vs0	-0.40	1.00	1.00	1.00	1.00	1.00	1.00	2.00
20	33:33:33	Twice	Poor	2vs0	-3.10	2.00	2.00	2.00	2.00	2.00	2.00	4.00
21	10:45:45	Zero	Perfect	1vs0	1.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00
21	10:45:45	Zero	Perfect	2vs0	1.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00
23	10:45:45	Zero	Good	1vs0	0.90	0.25	0.25	0.25	0.25	0.25	0.25	0.00
23	10:45:45	Zero	Good	2vs0	0.40	0.50	0.50	0.50	0.50	0.50	0.50	0.00
24	10:45:45	Zero	Moderate	1vs0	1.50	0.50	0.50	0.50	0.50	0.50	0.50	0.00
24	10:45:45	Zero	Moderate	2vs0	0.50	1.00	1.00	1.00	1.00	1.00	1.00	0.00
25	10:45:45	Zero	Poor	1vs0	1.70	1.00	1.00	1.00	1.00	1.00	1.00	0.00
25	10:45:45	Zero	Poor	2vs0	-0.30	2.00	2.00	2.00	2.00	2.00	2.00	0.00
26	10:45:45	Half	Perfect	1vs0	1.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00
26	10:45:45	Half	Perfect	2vs0	1.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00
28	10:45:45	Half	Good	1vs0	0.90	0.25	0.25	0.25	0.25	0.25	0.25	0.12
28	10:45:45	Half	Good	2vs0	0.40	0.50	0.50	0.50	0.50	0.50	0.50	0.25
29	10:45:45	Half	Moderate	1vs0	1.50	0.50	0.50	0.50	0.50	0.50	0.50	0.25
29	10:45:45	Half	Moderate	2vs0	0.50	1.00	1.00	1.00	1.00	1.00	1.00	0.50
30	10:45:45	Half	Poor	1vs0	1.70	1.00	1.00	1.00	1.00	1.00	1.00	0.50
30	10:45:45	Half	Poor	2vs0	-0.30	2.00	2.00	2.00	2.00	2.00	2.00	1.00
31	10:45:45	Same	Perfect	1vs0	1.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00
31	10:45:45	Same	Perfect	2vs0	1.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00
33	10:45:45	Same	Good	1vs0	0.90	0.25	0.25	0.25	0.25	0.25	0.25	0.25
33	10:45:45	Same	Good	2vs0	0.40	0.50	0.50	0.50	0.50	0.50	0.50	0.50
34	10:45:45	Same	Moderate	1vs0	1.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50
34	10:45:45	Same	Moderate	2vs0	0.50	1.00	1.00	1.00	1.00	1.00	1.00	1.00
35	10:45:45	Same	Poor	1vs0	1.70	1.00	1.00	1.00	1.00	1.00	1.00	1.00
35	10:45:45	Same	Poor	2vs0	-0.30	2.00	2.00	2.00	2.00	2.00	2.00	2.00
36	10:45:45	Twice	Perfect	1vs0	1.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00
36	10:45:45	Twice	Perfect	2vs0	1.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00
38	10:45:45	Twice	Good	1vs0	0.90	0.25	0.25	0.25	0.25	0.25	0.25	0.50
38	10:45:45	Twice	Good	2vs0	0.40	0.50	0.50	0.50	0.50	0.50	0.50	1.00
39	10:45:45	Twice	Moderate	1vs0	1.50	0.50	0.50	0.50	0.50	0.50	0.50	1.00
39	10:45:45	Twice	Moderate	2vs0	0.50	1.00	1.00	1.00	1.00	1.00	1.00	2.00
40	10:45:45	Twice	Poor	1vs0	1.70	1.00	1.00	1.00	1.00	1.00	1.00	2.00
40	10:45:45	Twice	Poor	2vs0	-0.30	2.00	2.00	2.00	2.00	2.00	2.00	4.00

Continued

Number	Size	RelX7	Equipoise	Contrast	0	1	2	3	4	5	6	7
41	10:10:80	Zero	Perfect	1vs0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
41	10:10:80	Zero	Perfect	2vs0	2.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
43	10:10:80	Zero	Good	1vs0	0.00	0.25	0.25	0.25	0.25	0.25	0.25	0.00
43	10:10:80	Zero	Good	2vs0	1.70	0.50	0.50	0.50	0.50	0.50	0.50	0.00
44	10:10:80	Zero	Moderate	1vs0	0.10	0.50	0.50	0.50	0.50	0.50	0.50	0.00
44	10:10:80	Zero	Moderate	2vs0	1.60	1.00	1.00	1.00	1.00	1.00	1.00	0.00
45	10:10:80	Zero	Poor	1vs0	0.80	1.00	1.00	1.00	1.00	1.00	1.00	0.00
45	10:10:80	Zero	Poor	2vs0	2.00	2.00	2.00	2.00	2.00	2.00	2.00	0.00
46	10:10:80	Half	Perfect	1vs0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
46	10:10:80	Half	Perfect	2vs0	2.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
48	10:10:80	Half	Good	1vs0	0.00	0.25	0.25	0.25	0.25	0.25	0.25	0.12
48	10:10:80	Half	Good	2vs0	1.70	0.50	0.50	0.50	0.50	0.50	0.50	0.25
49	10:10:80	Half	Moderate	1vs0	0.10	0.50	0.50	0.50	0.50	0.50	0.50	0.25
49	10:10:80	Half	Moderate	2vs0	1.60	1.00	1.00	1.00	1.00	1.00	1.00	0.50
50	10:10:80	Half	Poor	1vs0	0.80	1.00	1.00	1.00	1.00	1.00	1.00	0.50
50	10:10:80	Half	Poor	2vs0	2.00	2.00	2.00	2.00	2.00	2.00	2.00	1.00
51	10:10:80	Same	Perfect	1vs0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
51	10:10:80	Same	Perfect	2vs0	2.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
53	10:10:80	Same	Good	1vs0	0.00	0.25	0.25	0.25	0.25	0.25	0.25	0.25
53	10:10:80	Same	Good	2vs0	1.70	0.50	0.50	0.50	0.50	0.50	0.50	0.50
54	10:10:80	Same	Moderate	1vs0	0.10	0.50	0.50	0.50	0.50	0.50	0.50	0.50
54	10:10:80	Same	Moderate	2vs0	1.60	1.00	1.00	1.00	1.00	1.00	1.00	1.00
55	10:10:80	Same	Poor	1vs0	0.80	1.00	1.00	1.00	1.00	1.00	1.00	1.00
55	10:10:80	Same	Poor	2vs0	2.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00
56	10:10:80	Twice	Perfect	1vs0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
56	10:10:80	Twice	Perfect	2vs0	2.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
58	10:10:80	Twice	Good	1vs0	0.00	0.25	0.25	0.25	0.25	0.25	0.25	0.50
58	10:10:80	Twice	Good	2vs0	1.70	0.50	0.50	0.50	0.50	0.50	0.50	1.00
59	10:10:80	Twice	Moderate	1vs0	0.10	0.50	0.50	0.50	0.50	0.50	0.50	1.00
59	10:10:80	Twice	Moderate	2vs0	1.60	1.00	1.00	1.00	1.00	1.00	1.00	2.00
60	10:10:80	Twice	Poor	1vs0	0.80	1.00	1.00	1.00	1.00	1.00	1.00	2.00
60	10:10:80	Twice	Poor	2vs0	2.00	2.00	2.00	2.00	2.00	2.00	2.00	4.00

4.1.3 Outcome generation

The linear predictor (log rate) for the Poisson count outcome was assigned based on all covariates and treatment. The log link was used to avoid the issue of non-collapsibility of the logit link [Greenland et al., 1999].

$$\begin{aligned}\eta_{Yi} &= \beta_0 + \beta_{A1} I(A_i = 1) + \beta_{A2} I(A_i = 2)\\ &\quad + \mathbf{X}_i^T \boldsymbol{\beta}_X + I(A_i = 1)\mathbf{X}_i^T \boldsymbol{\beta}_{XA1} + I(A_i = 2)\mathbf{X}_i^T \boldsymbol{\beta}_{XA2}\end{aligned}$$

$$Y_i \sim \text{Poisson}\left(\exp(\eta_{Yi})\right)$$

Additionally, the following counterfactual log rates were kept for use in calculating the marginal causal effects.

$$\begin{aligned}\eta_{Y_i^0} &= \beta_0 + \mathbf{X}_i^T \boldsymbol{\beta}_X\\ \eta_{Y_i^1} &= \beta_0 + \beta_{A1} + \mathbf{X}_i^T \boldsymbol{\beta}_X + \mathbf{X}_i^T \boldsymbol{\beta}_{XA1}\\ \eta_{Y_i^2} &= \beta_0 + \beta_{A2} + \mathbf{X}_i^T \boldsymbol{\beta}_X + \mathbf{X}_i^T \boldsymbol{\beta}_{XA2}\end{aligned}$$

The outcome model parameter values were the following (RR: rate ratio).

$$\beta_0 = \log(0.20) \quad \text{Baseline rate}$$

$$(\beta_{A1}, \beta_{A2}) = (\log(1.0), \log(1.0)) \quad \text{Null main effects}$$

$$\boldsymbol{\beta}_X^T = \begin{cases} (\log(1.2), \log(1.2), \log(1.2), \log(1.2), \log(1.2), \log(1.2), \log(1.2)) & X_7 - Y \text{ RR } 1.2 \\ (\log(1.2), \log(1.2), \log(1.2), \log(1.2), \log(1.2), \log(1.2), \log(1.5)) & X_7 - Y \text{ RR } 1.5 \\ (\log(1.2), \log(1.2), \log(1.2), \log(1.2), \log(1.2), \log(1.2), \log(2.0)) & X_7 - Y \text{ RR } 2.0 \end{cases}$$

$$\begin{bmatrix} \boldsymbol{\beta}_{XA1}^T \\ \boldsymbol{\beta}_{XA2}^T \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \quad \text{No effect modification}$$

4.2 Estimands of interest

Four outcome analyses were conducted. The first was the unadjusted analysis. The other three were weighted analyses with inverse probability of treatment weights (IPTW) [Robins et al., 2000], matching weights (MW) [Li and Greene, 2013, Yoshida et al., 2017], and overlap weights [Li et al., 2016, Li et al., 2018, Li and Li, 2018].

$$\begin{aligned} IPTW_i &= \frac{1}{\sum_{j=0}^{2} I(A_i = j) e_{ji}} \\ MW_i &= \frac{\min(e_{0i}, e_{1i}, e_{2i})}{\sum_{j=0}^{2} I(A_i = j) e_{ji}} \\ OW_i &= \frac{\frac{1}{\frac{1}{e_{0i}} + \frac{1}{e_{1i}} + \frac{1}{e_{2i}}}}{\sum_{j=0}^{2} I(A_i = j) e_{ji}} \end{aligned}$$

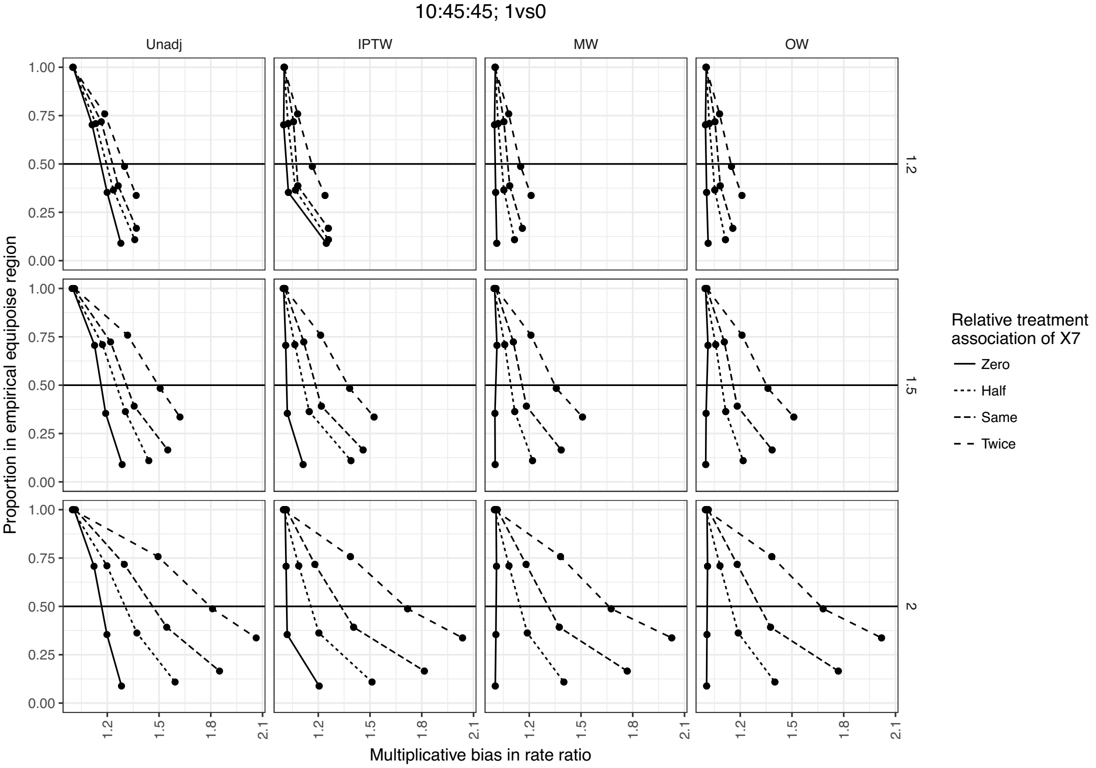
where $I(\cdot)$ is an indicator function that is 1 if the expression inside holds and 0 if not.

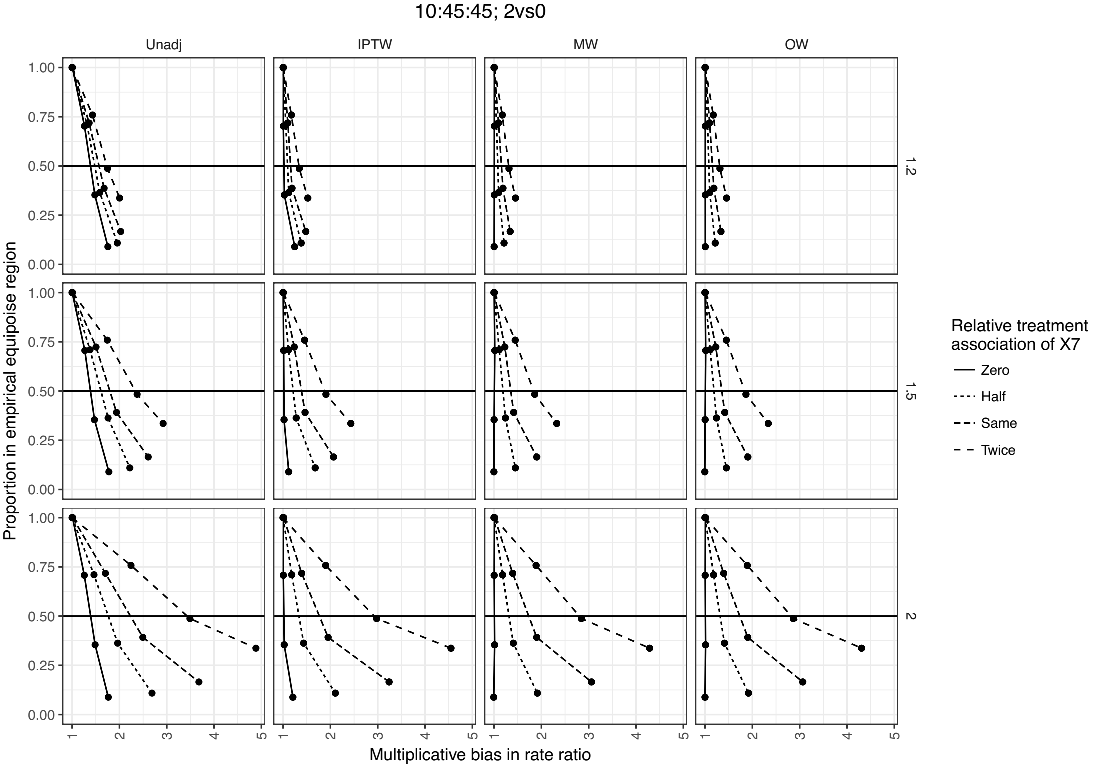
5 Simulation: additional results

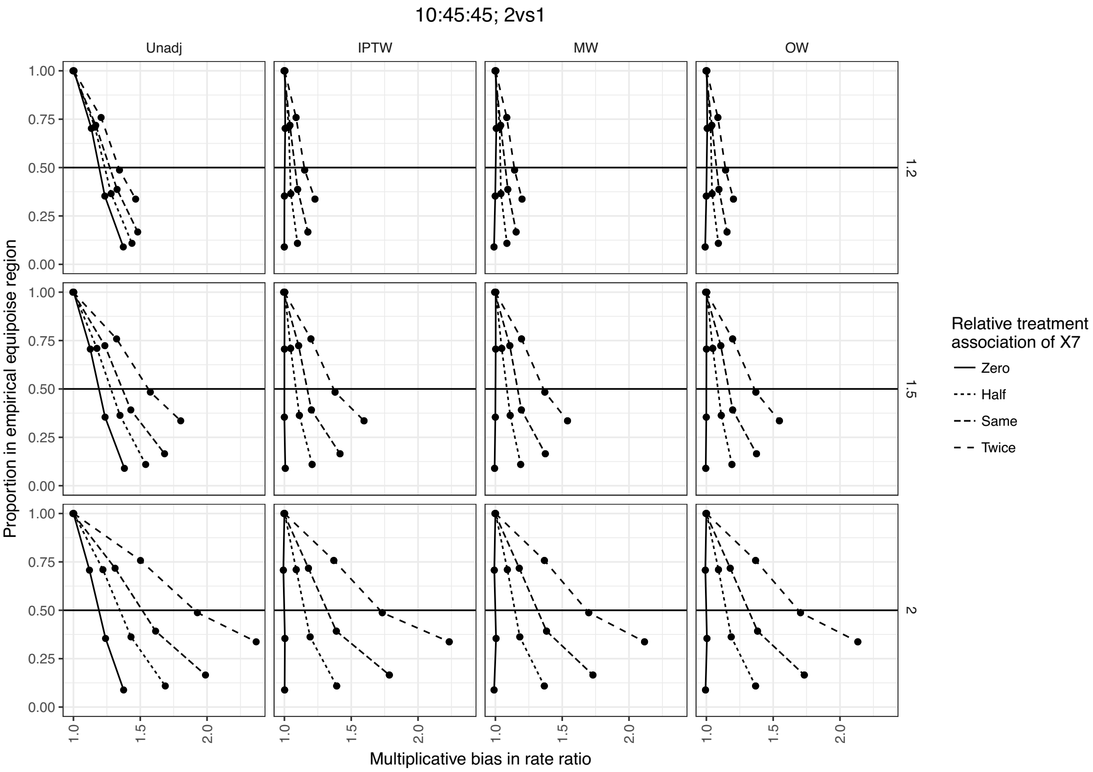
5.1 Overall proportion as the summary measure of empirical equipoise

In the following additional results, the index was the proportion of the overall cohort that fell into the proposed empirical equipoise region (same as Figures 1-3).

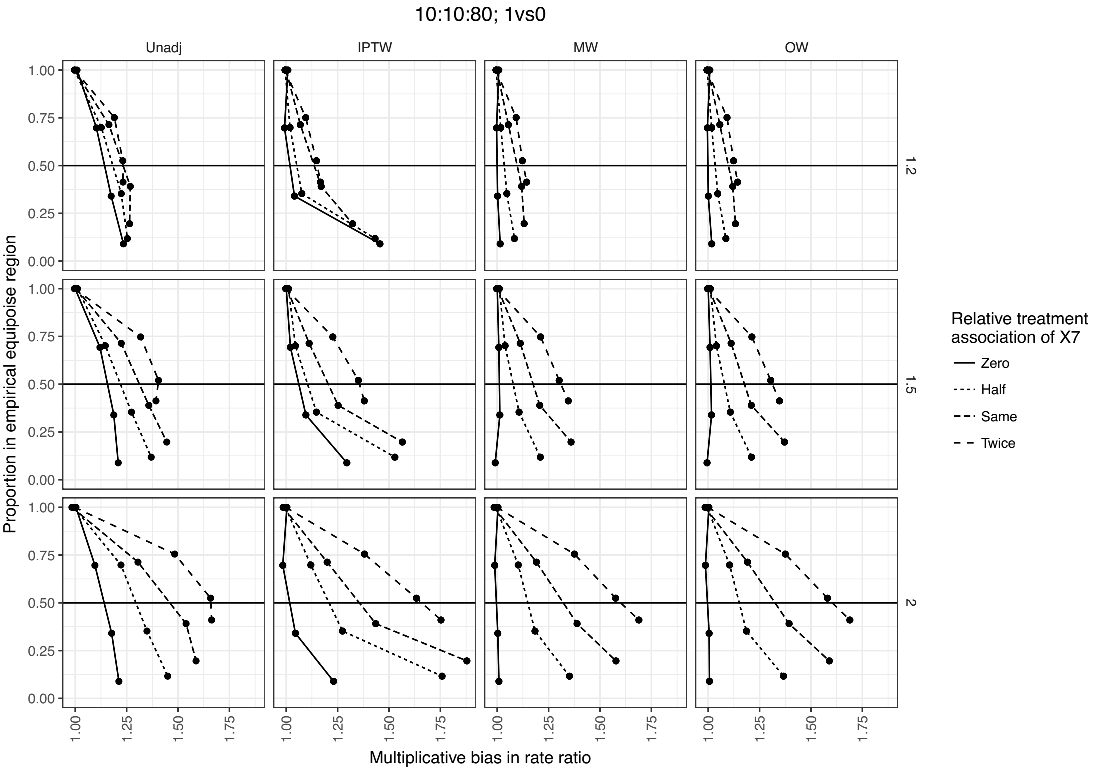
5.1.1 Unequal group sizes 10:45:45

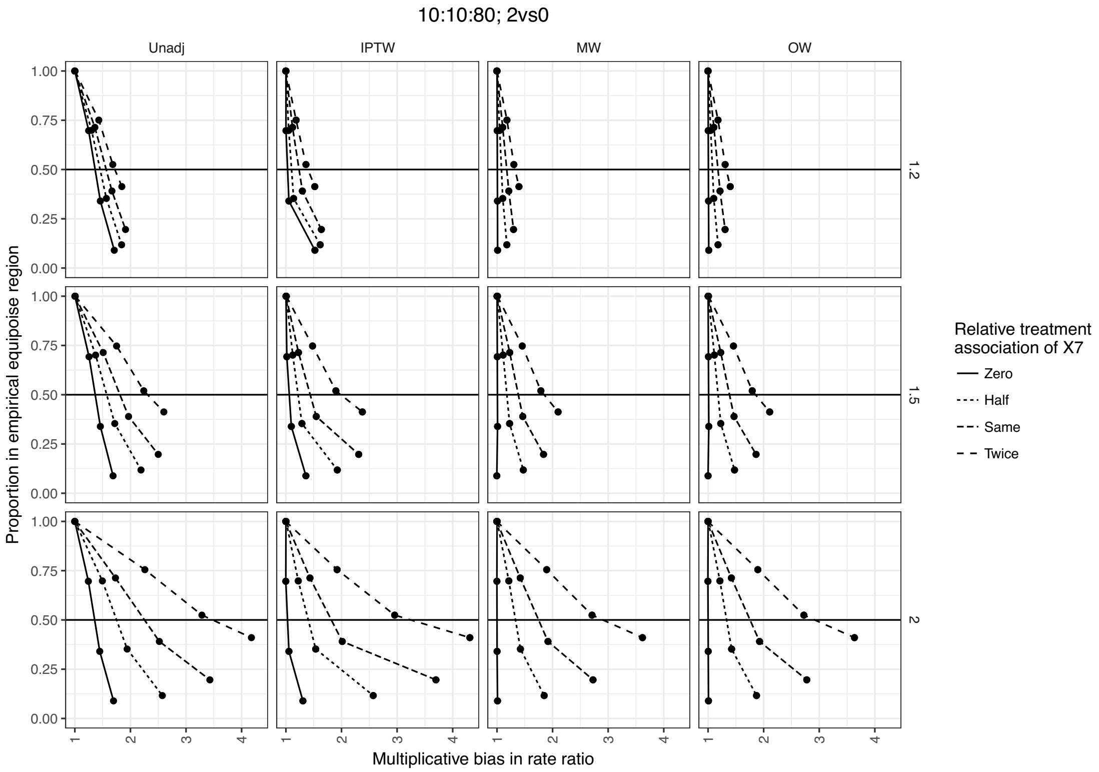


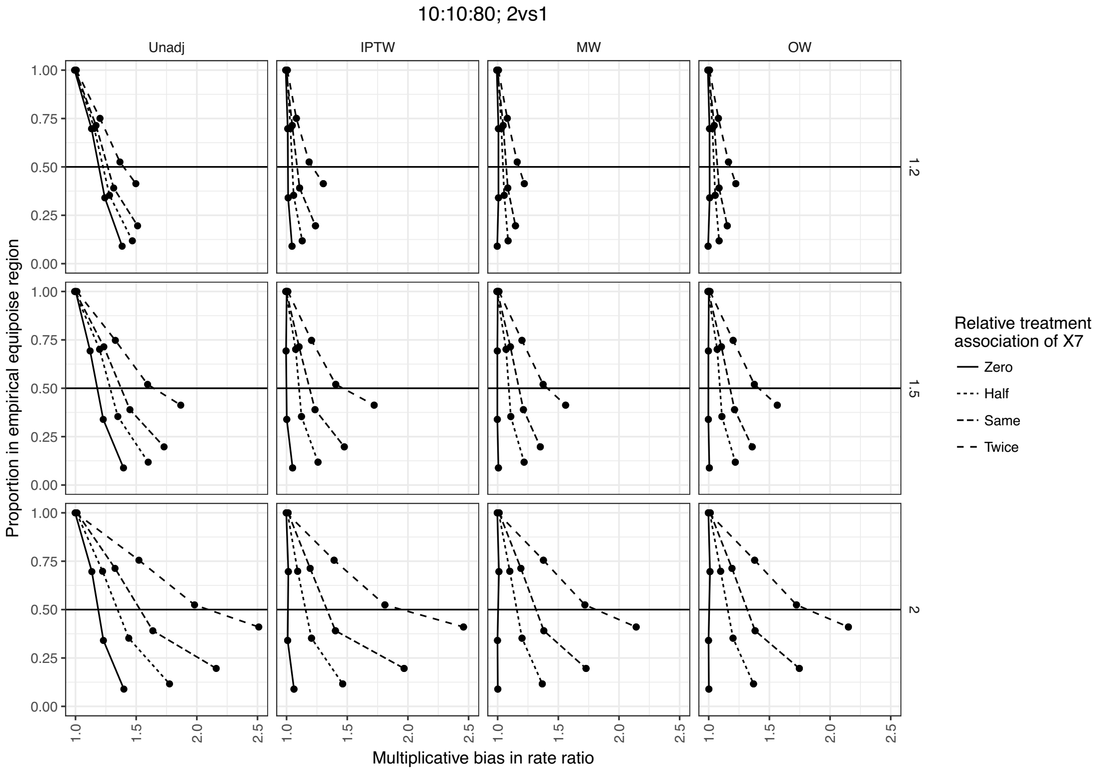




5.1.2 Unequal group sizes 10:10:80



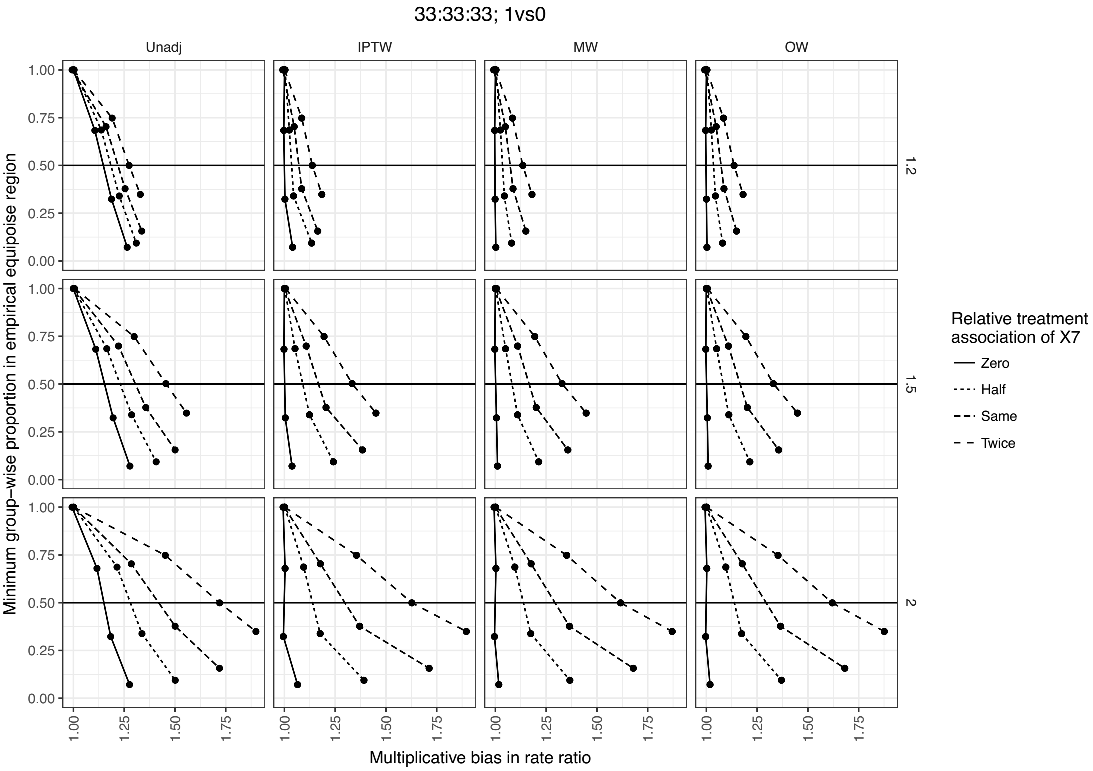


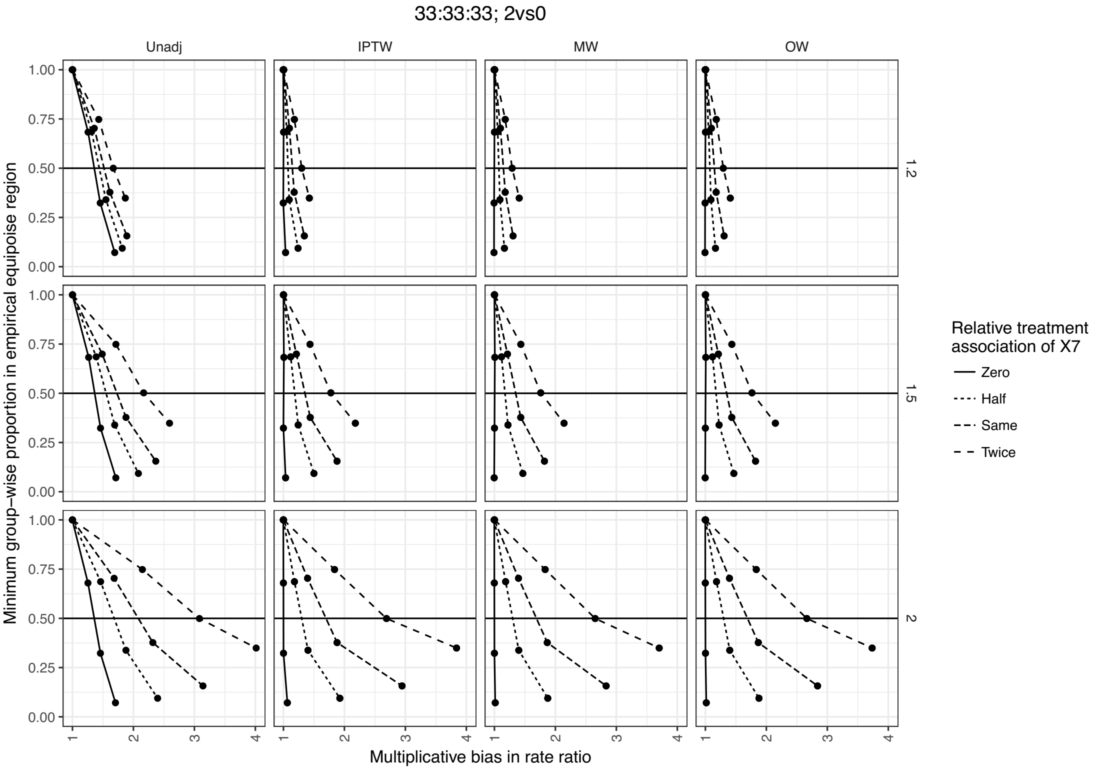


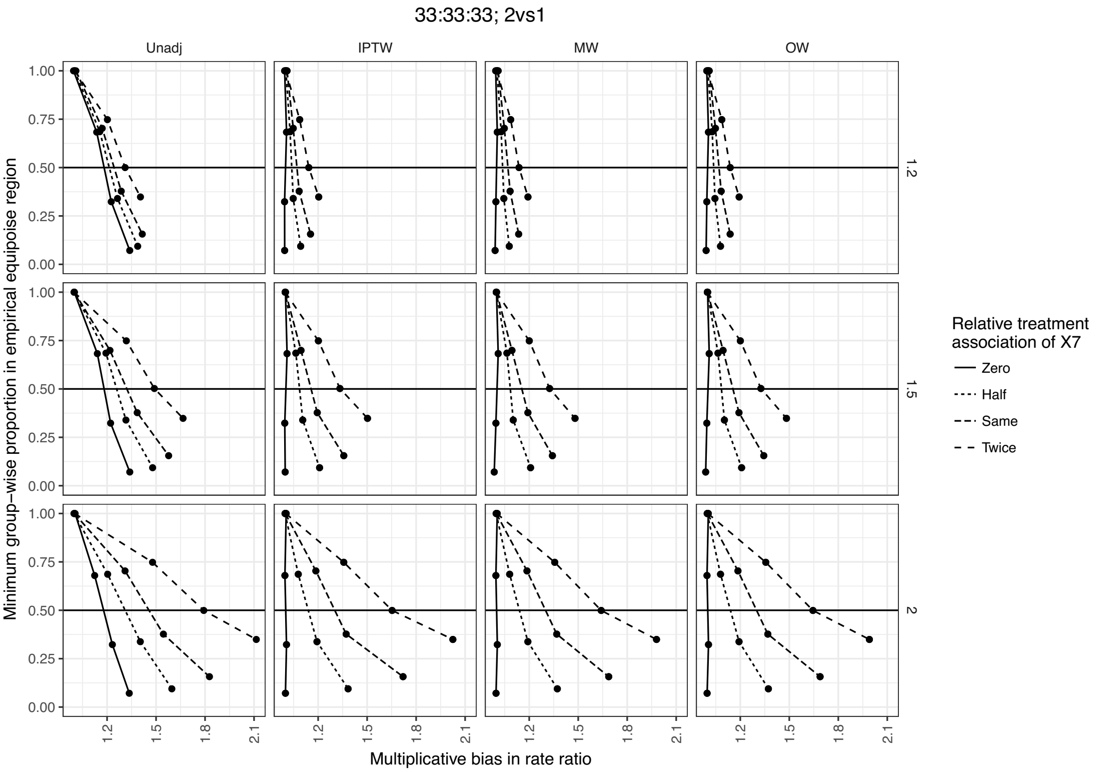
5.2 Minimum group-wise proportion as the summary measure of empirical equipoise

In the following additional results, the index was the minimum of the group-wise proportions of the treatment groups that fell into the proposed empirical equipoise region. The results were essentially the same as the

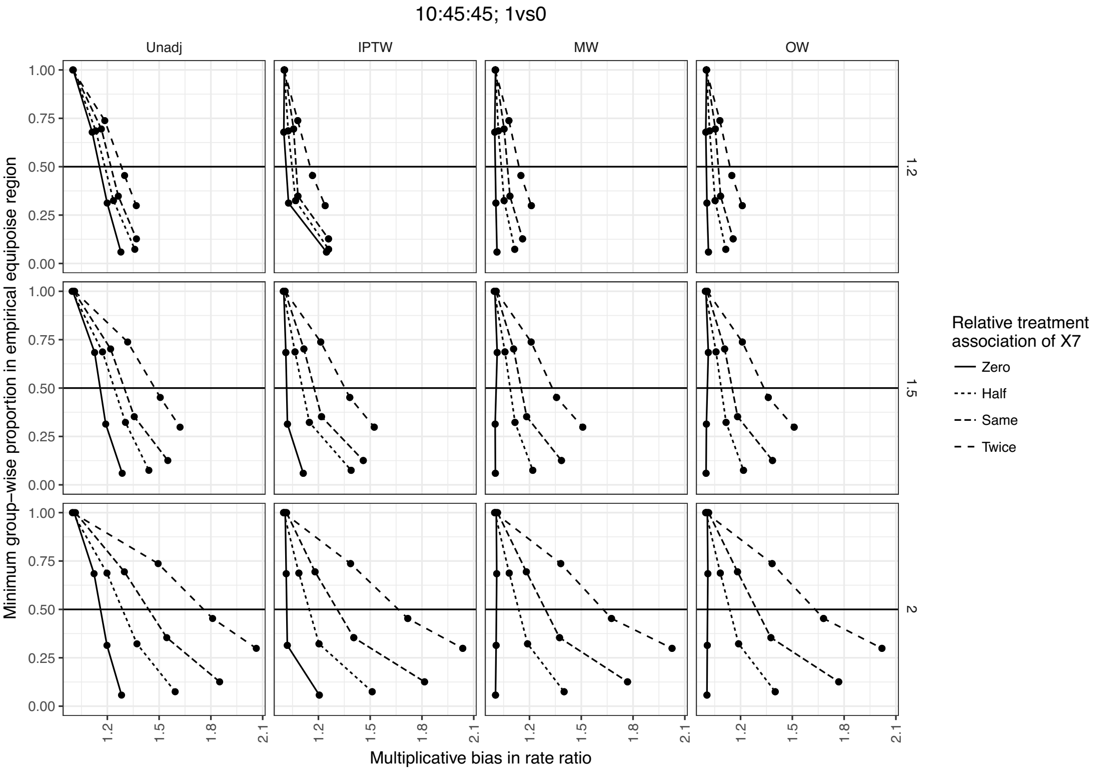
5.2.1 Equal group sizes 33:33:33

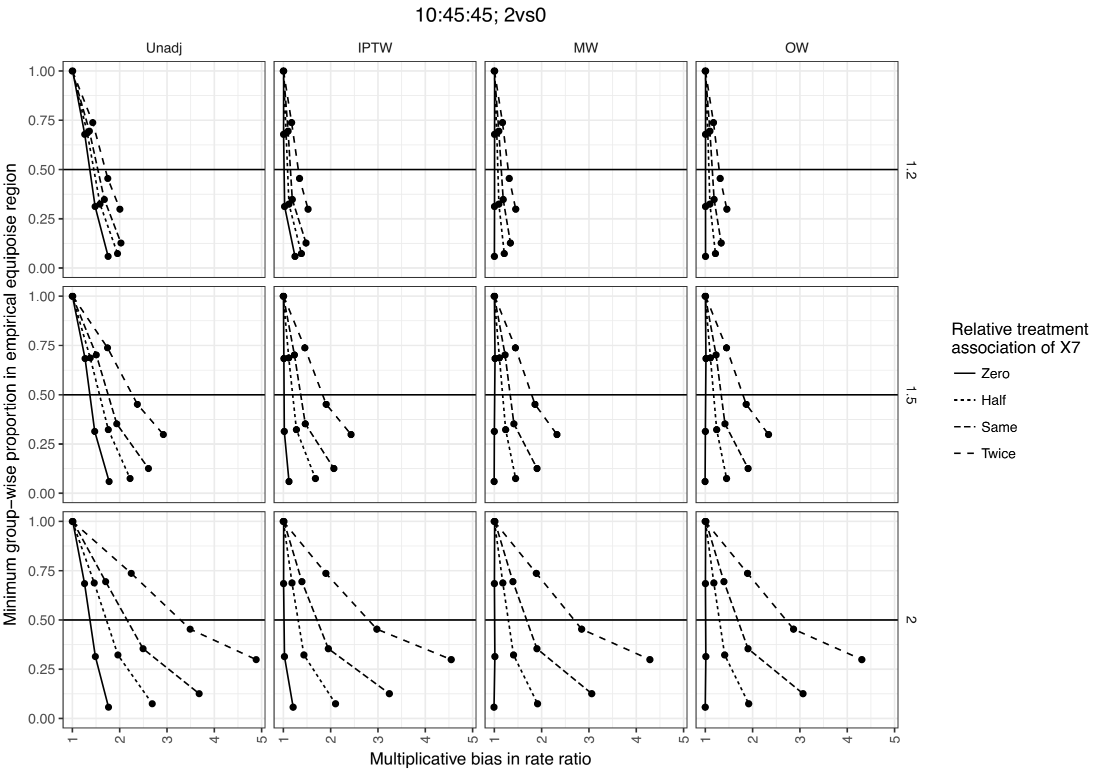




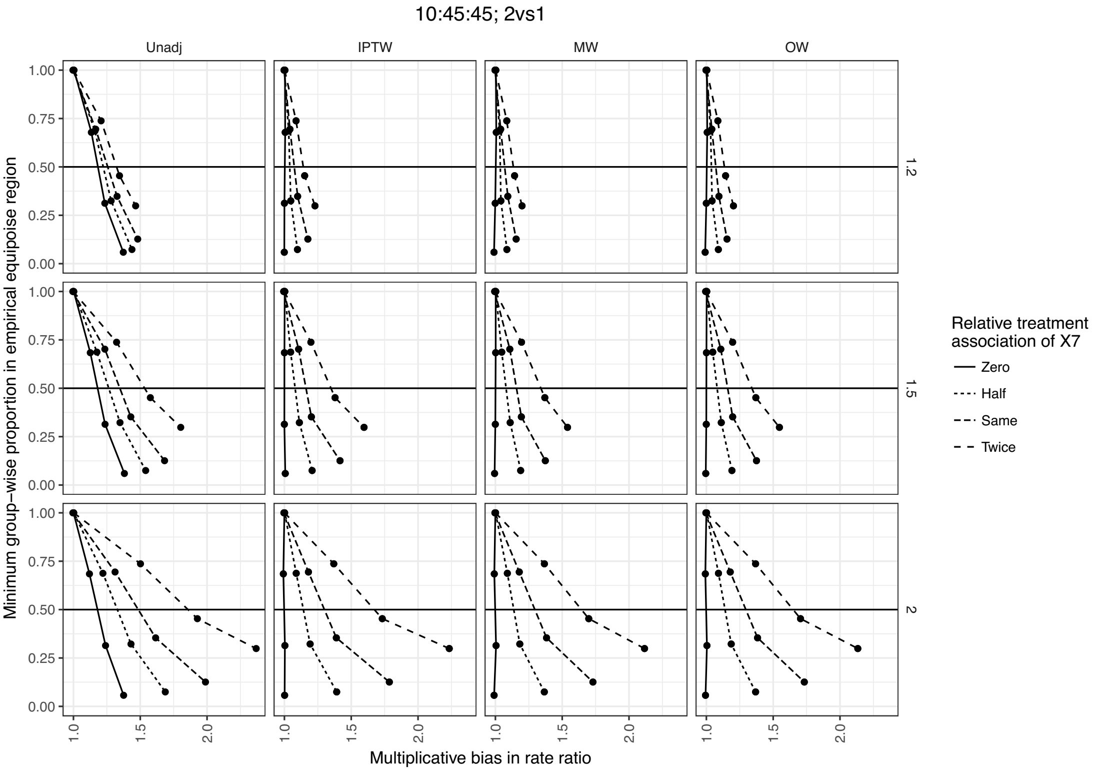


5.2.2 Unequal group sizes 10:45:45



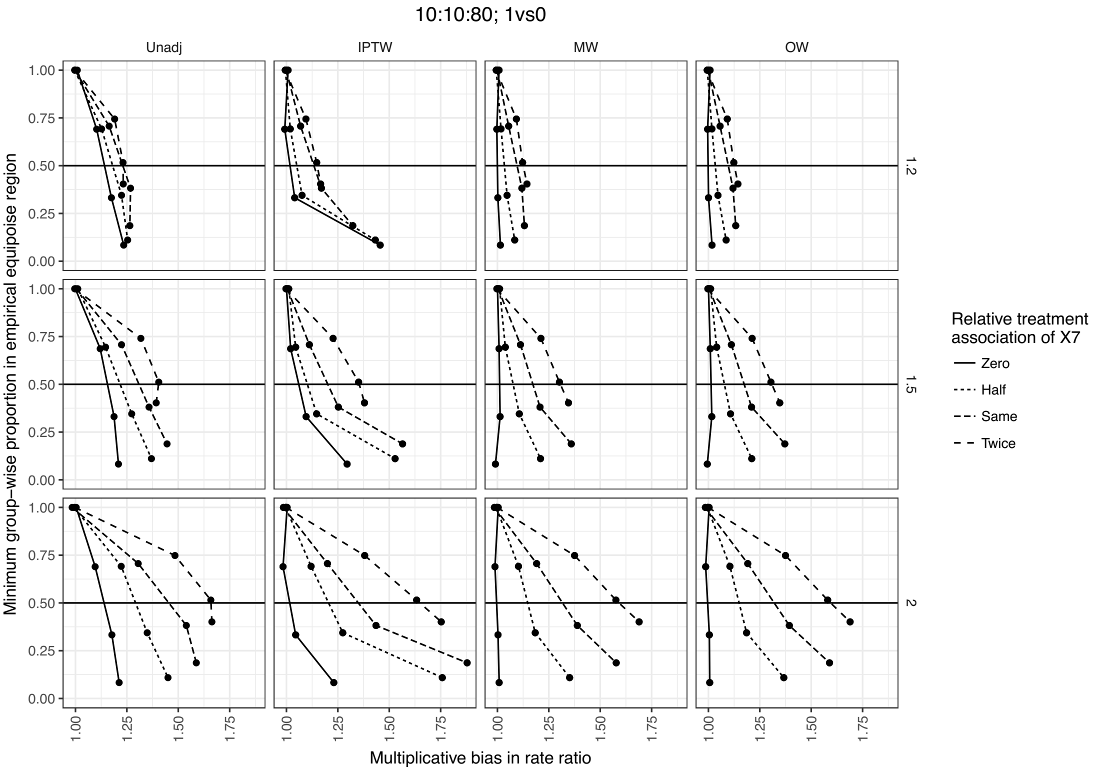


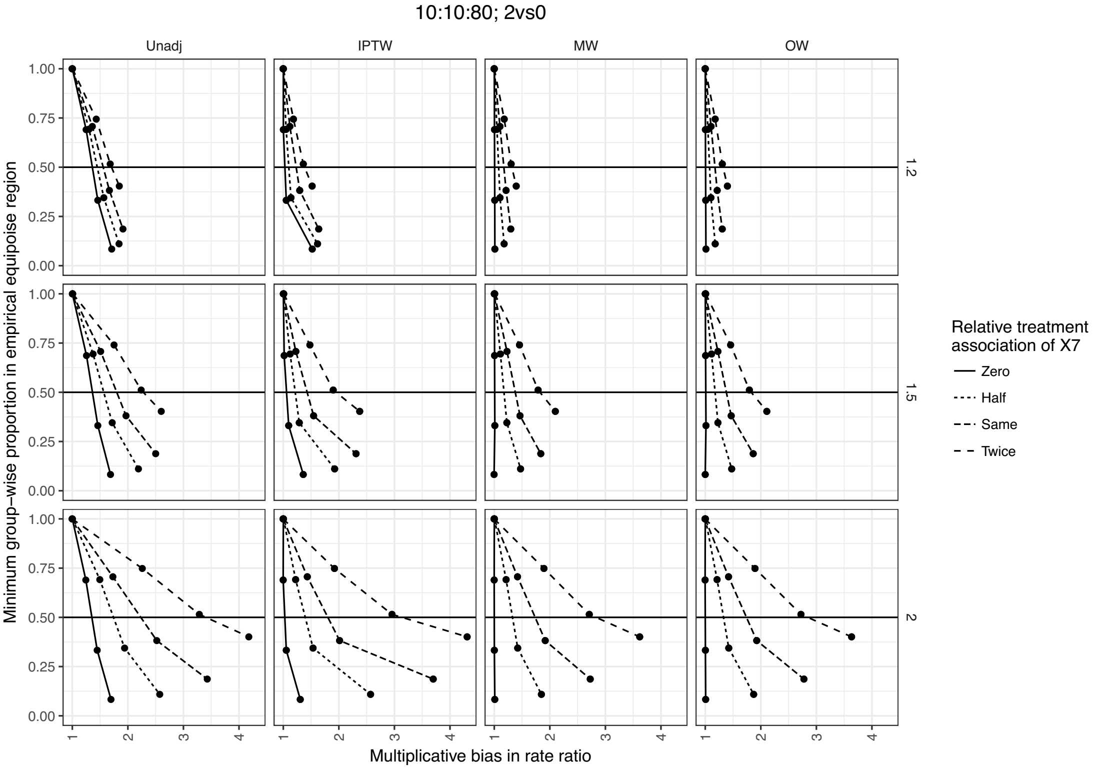
10:45:45; 2vs0

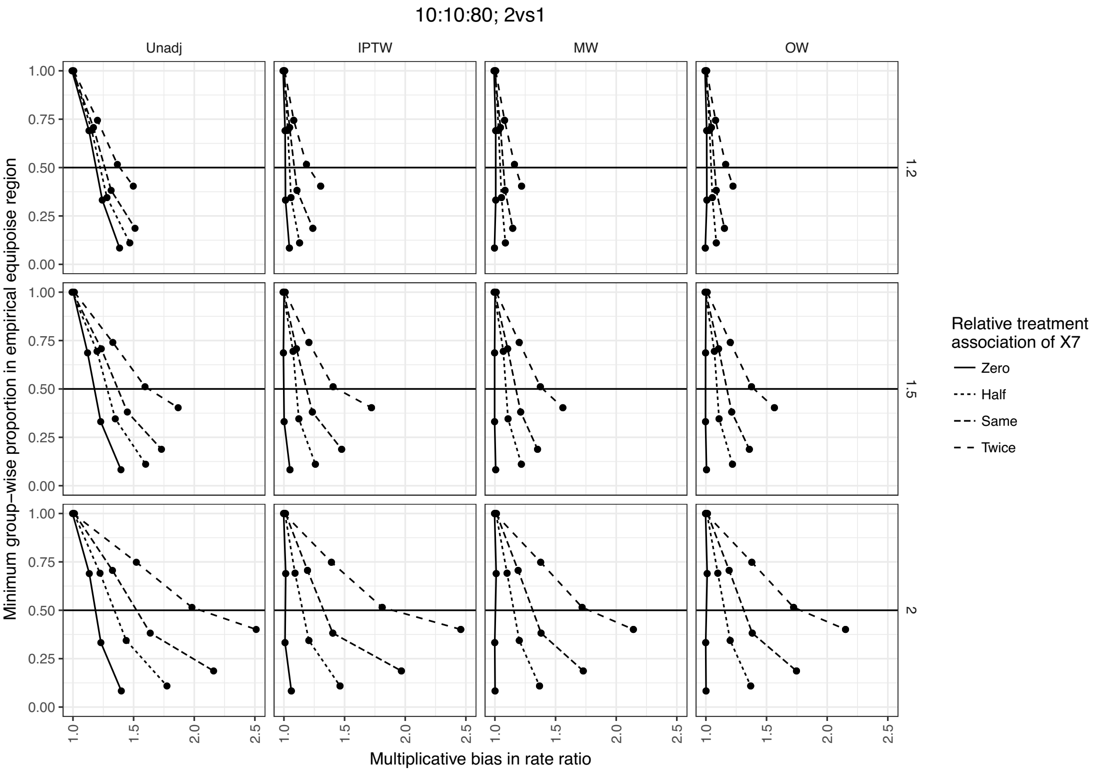


10:45:45; 2vs1

5.2.3 Unequal group sizes 10:10:80



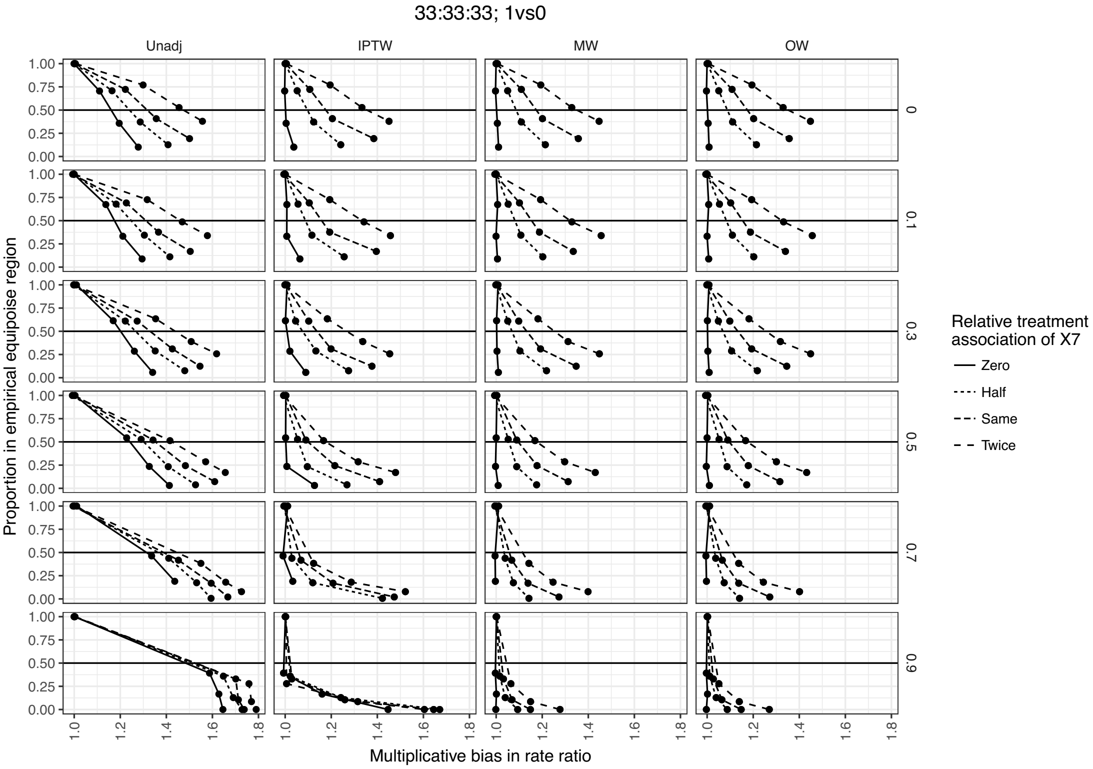


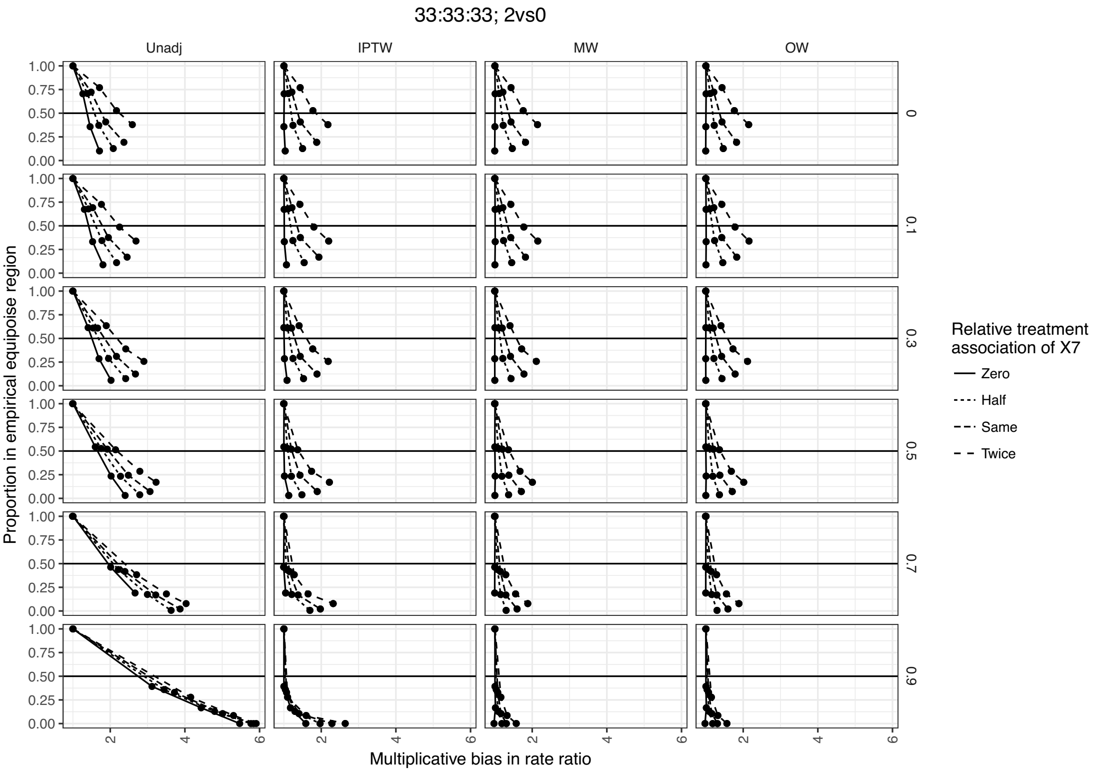


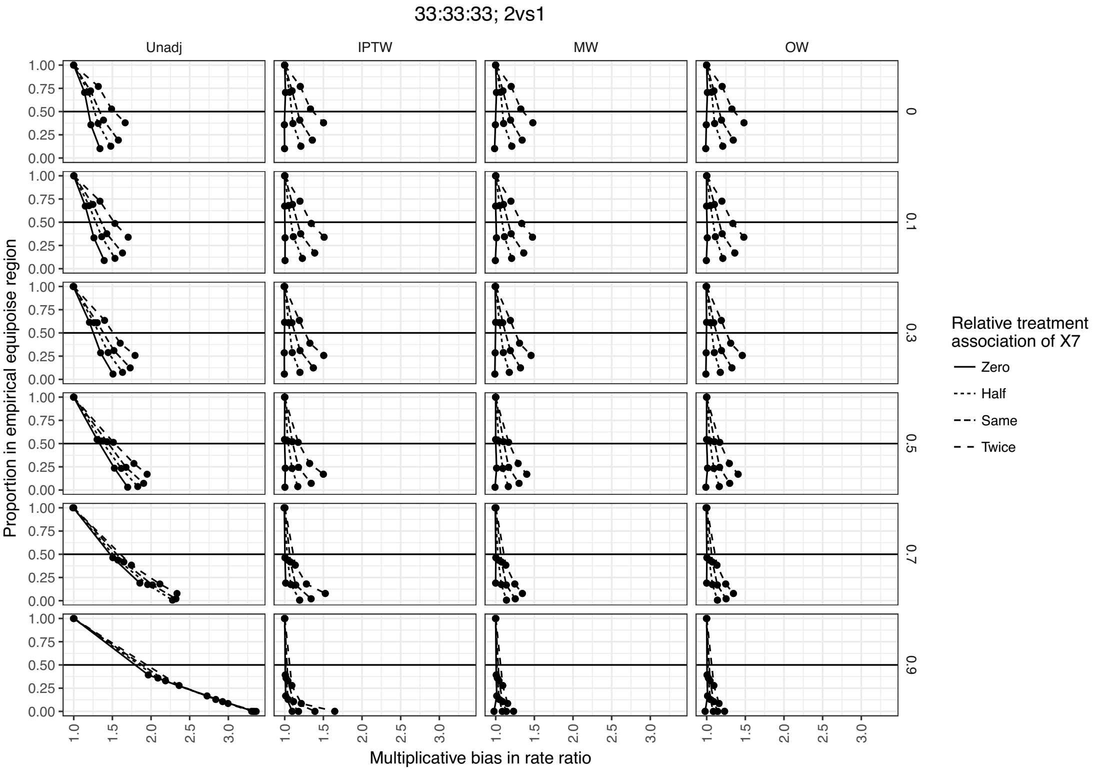
5.3 Different correlation structures

Here the correlation among covariates was varied from 0 to 0.9 (rows of the panels). The RR for the unmeasured variable was kept at 1.5.

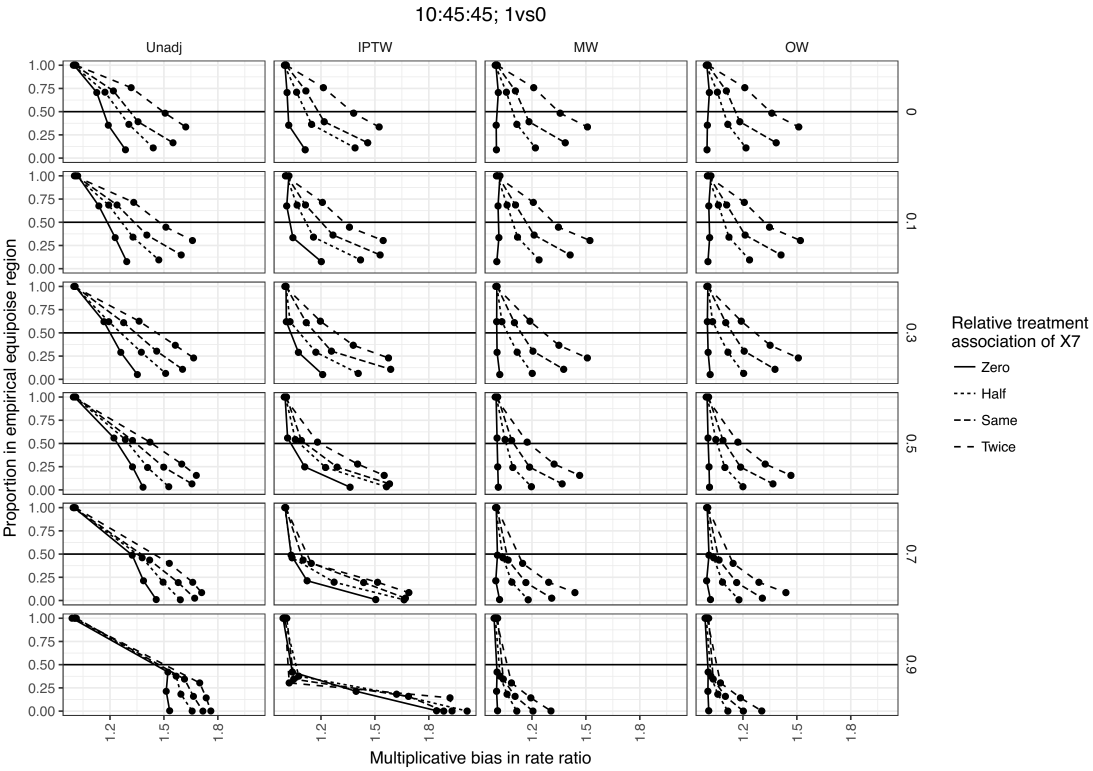
5.3.1 Equal group sizes 33:33:33

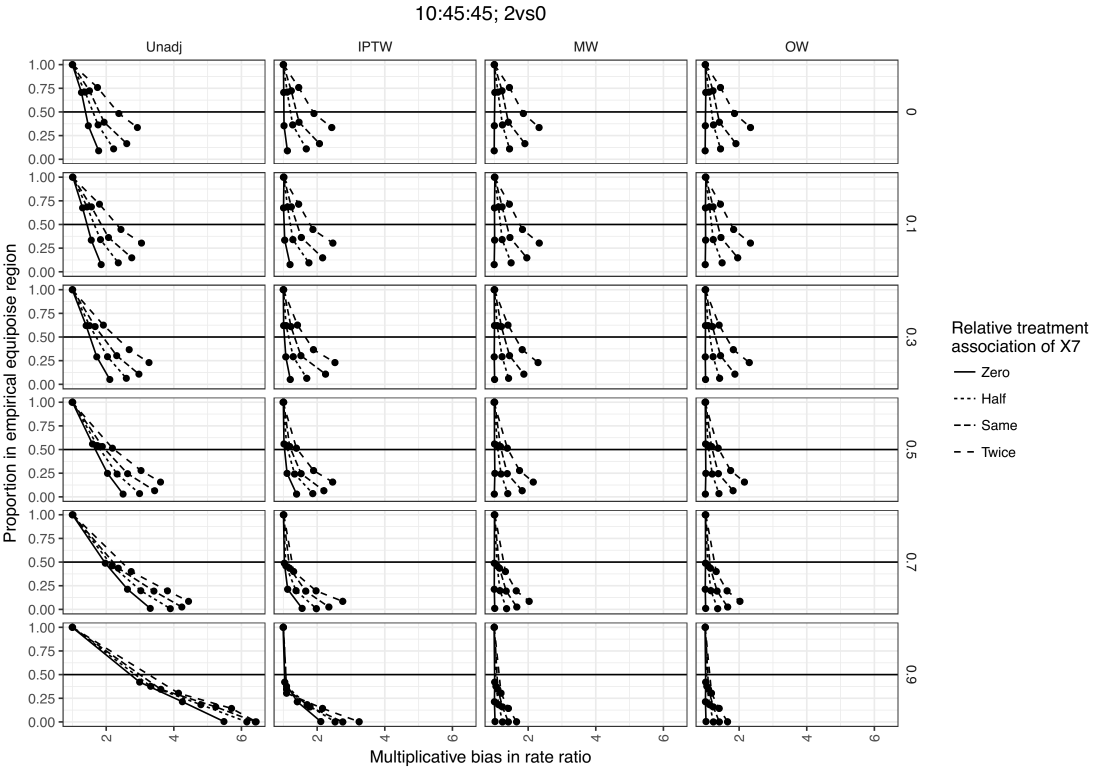


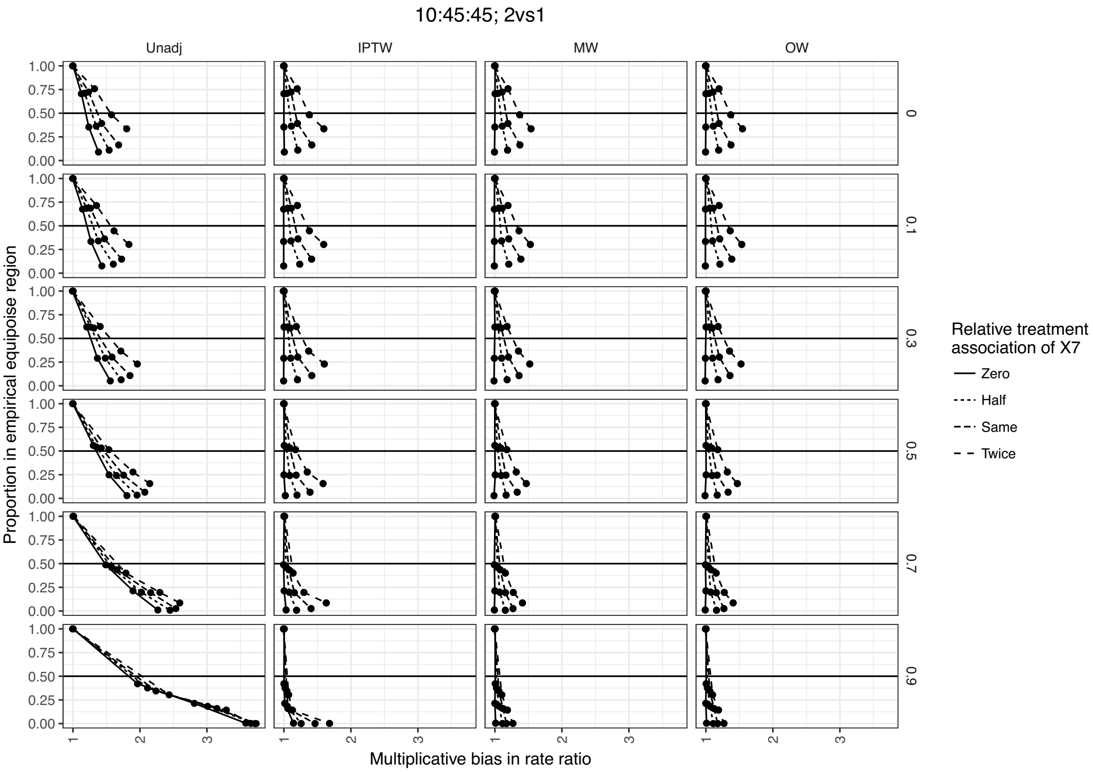

33:33:33; 2vs0
Unadj
IPTW
MW
OW
0
0.1
0.3
0.5
0.7
0.9
Proportion in empirical equipoise region
1.00
0.75
0.50
0.25
0.00
Multiplicative bias in rate ratio
2
4
6
Relative treatment association of X7
Zero
Half
Same
Twice


33:33:33; 2vs1
Unadj
IPTW
MW
OW
0
0.1
0.3
0.5
0.7
0.9
Proportion in empirical equipoise region
1.00
0.75
0.50
0.25
0.00
Multiplicative bias in rate ratio
1.0
1.5
2.0
2.5
3.0
Relative treatment association of X7
Zero
Half
Same
Twice

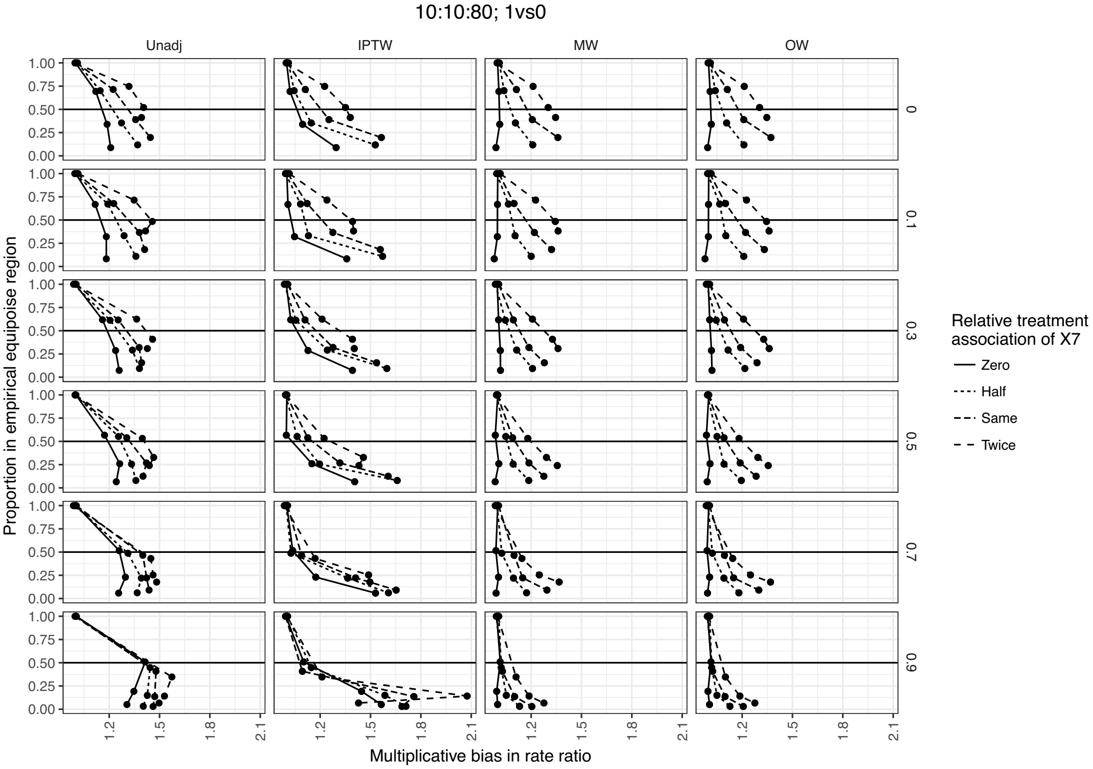
5.3.2 Unequal group sizes 10:45:45

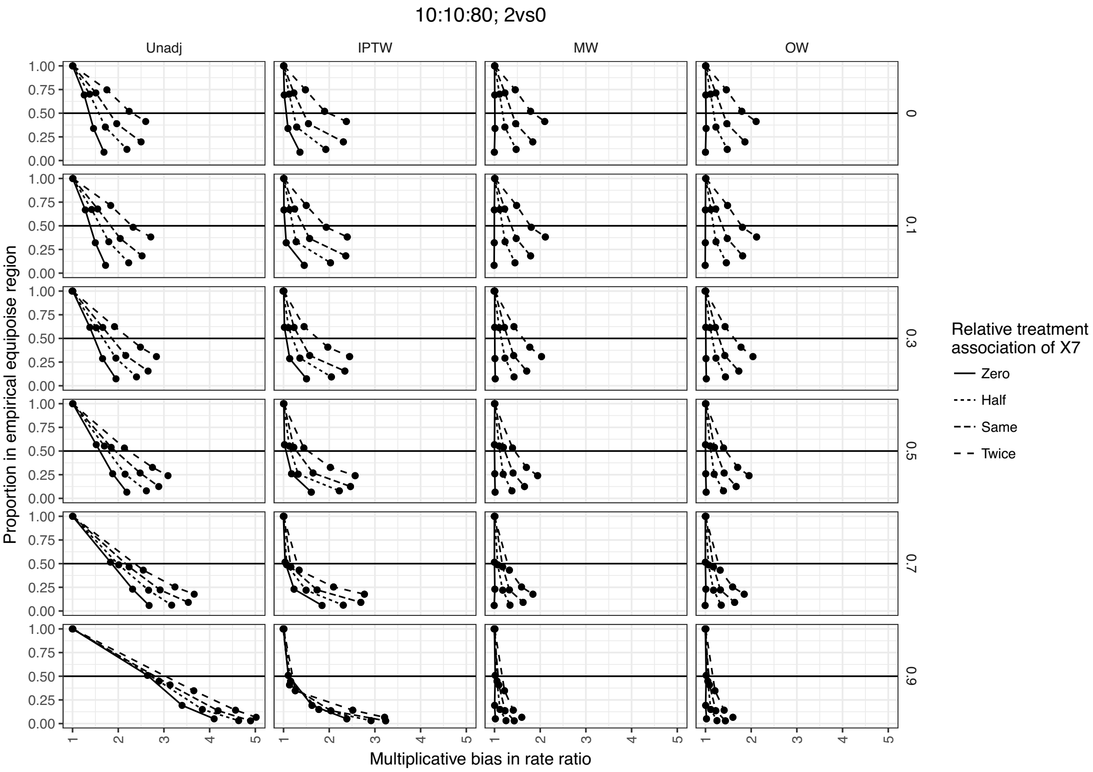


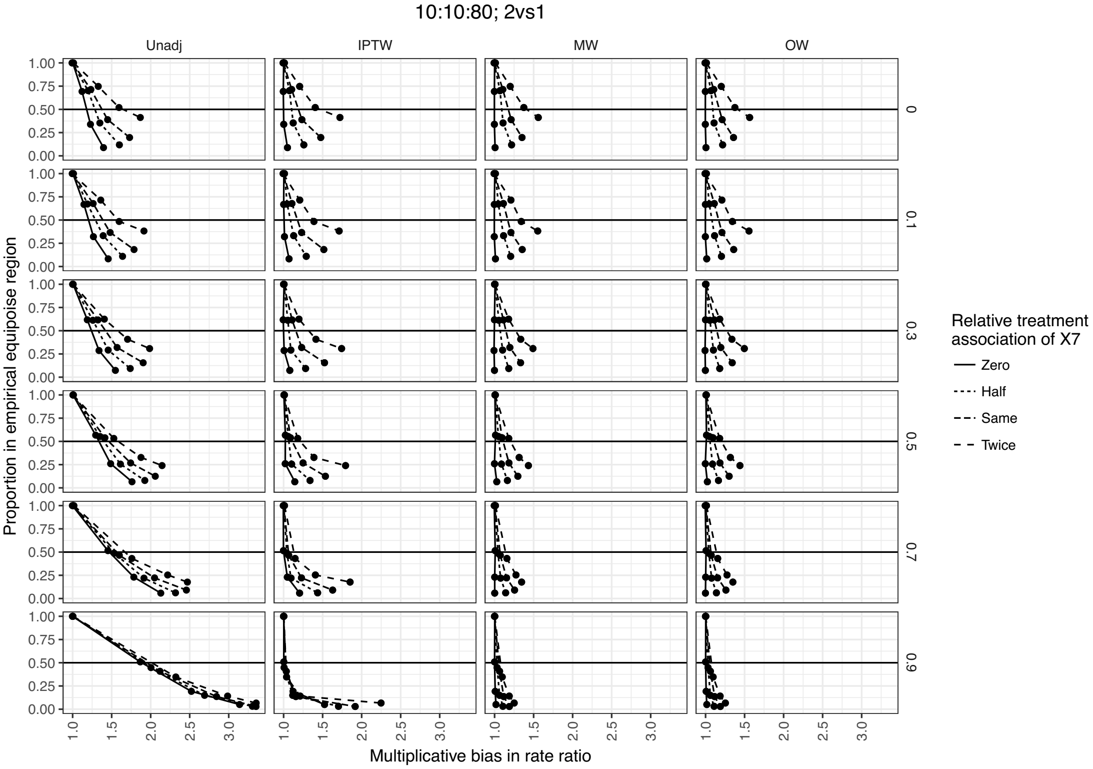




5.3.3 Unequal group sizes 10:10:80







6 Bibliography

[Greenland et al., 1999] Greenland, S., Robins, J. M., and Pearl, J. (1999). Confounding and Collapsibility in Causal Inference. *Statistical Science*, 14(1):29–46.

[Hamilton, 2017] Hamilton, N. (2017). Ggtern: An Extension to 'ggplot2', for the Creation of Ternary Diagrams.

[Hosmer et al., 2013] Hosmer, D. W., Lemeshow, S., and Sturdivant, R. X. (2013). *Applied Logistic Regression*. Wiley series in probability and statistics. Wiley, Hoboken, New Jersey, third edition edition.

[Kang et al., 2018] Kang, E. H., Jin, Y., Brill, G., Lewey, J., Patorno, E., Desai, R. J., and Kim, S. C. (2018). Comparative Cardiovascular Risk of Abatacept and Tumor Necrosis Factor Inhibitors in Patients With Rheumatoid Arthritis With and Without Diabetes Mellitus: A Multidatabase Cohort Study. *J Am Heart Assoc*, 7(3).

[Kim et al., 2017] Kim, S. C., Solomon, D. H., Rogers, J. R., Gale, S., Klearman, M., Sarsour, K., and Schneeweiss, S. (2017). Cardiovascular Safety of Tocilizumab Versus Tumor Necrosis Factor Inhibitors in Patients With Rheumatoid Arthritis: A Multi-Database Cohort Study. *Arthritis & Rheumatology (Hoboken, N.J.)*, 69(6):1154–1164.

[Li and Li, 2018] Li, F. and Li, F. (2018). Propensity Score Weighting for Causal Inference with Multi-valued Treatments. *arXiv:1808.05339 [stat]*.

[Li et al., 2016] Li, F., Morgan, K. L., and Zaslavsky, A. M. (2016). Balancing Covariates via Propensity Score Weighting. *Journal of the American Statistical Association*, 0(0):1–11.

[Li et al., 2018] Li, F., Thomas, L. E., and Li, F. (2018). Addressing Extreme Propensity Scores via the Overlap Weights. *Am. J. Epidemiol.*

[Li and Greene, 2013] Li, L. and Greene, T. (2013). A weighting analogue to pair matching in propensity score analysis. *Int J Biostat*, 9(2):215–234.

[Prentice and Pyke, 1979] Prentice, R. L. and Pyke, R. (1979). Logistic Disease Incidence Models and Case-Control Studies. *Biometrika*, 66(3):403–411.

[Robins et al., 2000] Robins, J. M., Hernán, M. A., and Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11(5):550–560.

[Scott and Wild, 1986] Scott, A. J. and Wild, C. J. (1986). Fitting Logistic Models Under Case-Control or Choice Based Sampling. *J Royal Stat Soc*, 48(2):170–182.

[Singh et al., 2016] Singh, J. A., Saag, K. G., Bridges, S. L., Akl, E. A., Bannuru, R. R., Sullivan, M. C., Vaysbrot, E., McNaughton, C., Osani, M., Shmerling, R. H., Curtis, J. R., Furst, D. E., Parks, D., Kavanaugh, A., O'Dell, J., King, C., Leong, A., Matteson, E. L., Schousboe, J. T., Drevlow, B., Ginsberg, S., Grober, J., St Clair, E. W., Tindall, E., Miller, A. S., McAlindon, T., and American College of Rheumatology (2016). 2015 American College of Rheumatology Guideline for the Treatment of Rheumatoid Arthritis. *Arthritis Care Res (Hoboken)*, 68(1):1–25.

[Smolen et al., 2017] Smolen, J. S., Landewé, R., Bijlsma, J., Burmester, G., Chatzidionysiou, K., Dougados, M., Nam, J., Ramiro, S., Voshaar, M., van Vollenhoven, R., Aletaha, D., Aringer, M., Boers, M., Buckley, C. D., Buttgereit, F., Bykerk, V., Cardiel, M., Combe, B., Cutolo, M., van Eijk-Hustings, Y., Emery, P., Finckh, A., Gabay, C., Gomez-Reino, J., Gossec, L., Gottenberg, J.-E., Hazes, J. M. W., Huizinga, T., Jani, M., Karateev, D., Kouloumas, M., Kvien, T., Li, Z., Mariette, X., McInnes, I., Mysler, E., Nash, P., Pavelka, K., Poór, G., Richez, C., van Riel, P., Rubbert-Roth, A., Saag, K., da Silva, J., Stamm, T., Takeuchi, T., Westhovens, R., de Wit, M., and van der Heijde, D. (2017). EULAR recommendations for the management of rheumatoid arthritis with synthetic and biological disease-modifying antirheumatic drugs: 2016 update. *Ann. Rheum. Dis.*, 76(6):960–977.

[Solomon et al., 2010] Solomon, D. H., Rassen, J. A., Glynn, R. J., Lee, J., Levin, R., and Schneeweiss, S. (2010). The comparative safety of analgesics in older adults with arthritis. *Arch. Intern. Med.*, 170(22):1968–1976.

[Walker et al., 2013] Walker, A. M., Patrick, A. R., Lauer, M. S., Hornbrook, M. C., Marin, M. G., Platt, R., Roger, V. L., Stang, P., and Schneeweiss, S. (2013). A tool for assessing the feasibility of comparative effectiveness research. *Comp Eff Res*, page 11.

[Yoshida et al., 2017] Yoshida, K., Hernandez-Diaz, S., Solomon, D. H., Jackson, J. W., Gagne, J. J., Glynn, R. J., and Franklin, J. M. (2017). Matching Weights to Simultaneously Compare Three Treatment Groups: Comparison to Three-way Matching. *Epidemiology*, 28(3):387–395.

Chapter 4: Multinomial Extension of Propensity Score Trimming Methods: A Simulation Study

AUTHORS: Kazuki Yoshida(1,2,3), Daniel H. Solomon(3,4), Sebastien Haneuse(2), Seoyoung C. Kim(3,4), Elisabetta Patorno(4), Sara K. Tedeschi(3), Houchen Lyu(3), Jessica M. Franklin(4), Til Stürmer(5), Sonia Hernandez-Diaz(1), Robert J. Glynn(2, 4)

AFFILIATIONS

1. Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, United States.
2. Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, United States.
3. Division of Rheumatology, Immunology and Allergy, Brigham and Women's Hospital, Boston, Massachusetts, United States.
4. Division of Pharmacoepidemiology and Pharmacoeconomics, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts, United States.
5. Department of Epidemiology, Gillings School of Global Public Health, University of North Carolina at Chapel Hill, North Carolina, United States.

DISCLOSURES:
KY received financial support for his doctoral study from Harvard T.H. Chan School of Public Health (partially supported by training grants from Pfizer, Takeda, and Bayer) and Honjo International Scholarship Foundation.
RJG received research support in the form of grants to his institution for clinical trial design, monitoring, and analysis from AstraZeneca, Kowa, Novartis, and Pfizer.
DHS receives salary support from institutional research grants from Amgen, Pfizer, Genentech, Bristol-Myers Squibb, and Corrona. He also receives royalties from UpToDate, and serves in unpaid roles in studies funded by Pfizer.
SCK received research grants to the Brigham and Women's Hospital from Pfizer, Roche/Genentech and Bristol-Myers Squibb for unrelated studies.
EP receives support from investigator-initiated grants to the Brigham and Women's Hospital from GSK and Boehringer Ingelheim, not related to the topic of the submitted work.

FUNDING INFORMATION:
KY received financial support for his doctoral study from Harvard T.H. Chan School of Public Health (partially supported by training grants from Takeda, Pfizer, and Bayer) and Honjo International Scholarship Foundation.
DHS receives salary support from NIH-K24AR055989 and NIH-P30AR072577.
EP was supported by a career development grant K08AG055670 from the National Institute on Aging.

ABSTRACT

Crump *et al*. (*Biometrika* 2009;96:187), Stürmer *et al*. (*Am J Epidemiol* 2010;172:843), and Walker *et al*. (*Comp Eff Res* 2013;3:11) proposed propensity score (PS) trimming methods as measures to improve efficiency (Crump) or reduce confounding (Stürmer and Walker). We generalized the trimming definitions by considering multinomial PSs, one for each treatment option and proved these proposed definitions reduce to the original binary definitions when we only have two treatment groups. We then examined the performance of the proposed multinomial trimming methods in the three treatment-group setting in which subjects with extreme PSs more likely had *unmeasured* confounders. Inverse probability of treatment weights (IPTW), matching weights (MW), and overlap weights (OW) were used to control for *measured* confounders. All three methods reduced bias regardless of the weighting methods in most scenarios. Multinomial Stürmer and Walker trimming were more successful in bias reduction when the three treatment groups had very different sizes (10:10:80). Variance reduction, seen in all methods with IPTW but not with MW or OW, was more successful with multinomial Crump and Stürmer trimming. In conclusion, our proposed definitions of multinomial PS trimming methods were successful within our simulation settings. Further validation in both empirical and simulated data are warranted.

INTRODUCTION

Epidemiologists utilize propensity score (PS) methods(1–3) to evaluate the comparability of subjects in alternative exposure groups and to aid in control of imbalances between groups. Several authors(4–6) suggested trimming the tails of the PS distribution. Crump *et al.*(4) suggested trimming to improve imprecision of inverse probability of treatment weight (IPTW)(7) estimator. Stürmer *et al.*(5) developed their trimming method to reduce bias by unmeasured confounders. Walker *et al.*(6) proposed a covariate overlap assessment tool that also serves as a trimming tool. They all focused on two-group comparisons.

Many diseases now have three or more treatment options, from which patients and physicians have to choose. Conducting head-to-head clinical trials is the ideal way to establish equivalence or differences of efficacy and safety. However, it is not generally feasible to compare more than two medications in head-to-head trials. As such, observational comparative effectiveness/safety research (CER) studies are increasingly used for comparing multiple treatment choices.

Multiple-group CER, conducted among three or more active treatment agents, seeks to answer the question: “Given a population of patients requiring treatment and without contraindications to any of several approved options, *which* treatment is most appropriate among a range of available options?” Although the active comparator design(8) is a useful design to improve covariate balance, the presence of unmeasured confounders remains a concern. As reasoned by the authors above(5,6), PS trimming has the potential to mitigate the bias by unmeasured confounders by focusing on a subset of subjects with better treatment equipoise. However, PS trimming strategies, as well as their performance, are not well established in the context of multiple-group CER. In this paper, we propose general strategies for PS trimming for

CER involving three or more treatment groups, illustrate their characteristics in empirical data examples, and evaluate how they perform in simulated scenarios with three treatment groups.

METHODS

Existing PS trimming methods in the two-group setting

To our knowledge, there are at least three PS trimming strategies often considered in epidemiological studies involving PS methods(4–6) (**Figure 4-1**, **Table 4-1**, and **eAppendix** Section 1). Let I be the set of indices {1, …, n} indexing individuals in the entire study sample of sample size n. Let $A_i \in \{0,1\}$ be the binary treatment indicator for individual i and $e_i = \mathrm{P}[A_i = 1 \mid \mathbf{X}_i]$ be the PS for this individual given the covariate vector status $\mathbf{X}_i$. Crump's trimming method is defined as follows(4).

$$I_c = \{i \in I: e_i \in [\alpha_c, 1 - \alpha_c]\}$$

Crump *et al.* proved that the estimated treatment effect based on IPTW has the optimal precision for a specific choice of α_c. In practice, they suggested using $\alpha_c = 0.1$ as a rule-of-thumb threshold that worked in a wide range of PS distributions in achieving near-optimal precision. At this threshold, the trimming method dictates that everyone who receives an IPTW of greater than 10 or less than 10/9 be removed.

Using the inverse of the cumulative distribution function of PS conditional on the treatment group $F_{ei|Ai}$, Stürmer's asymmetric trimming method can be written as follows(5).

$$I_s = \{i \in I: e_i \in \left[F_{e_i|A_i}^{-1}(\alpha_s|1), F_{e_i|A_i}^{-1}(1-\alpha_s|0)\ \right]\}$$

Rather than defining symmetric retention region around 0.5 as in Crump, this definition is based on the distribution of the PS in two treatment groups. The lower bound is defined by the 100×α_sth percentile of PS in the *treated*, and the upper bound is defined by the 100×(1-α_s)th percentile of PS in the *untreated*. Importantly, once this retention region [100×α_sth percentile in the treated, 100×(1-α_s)th percentile in the untreated] is constructed, every individual, *both* treated and untreated, outside this region is removed from the analysis dataset. This is necessary to avoid artificially introducing PS non-overlap. They examined 0.01, 0.025, and 0.05 for α_s. The rationale for this trimming strategy is to remove those who received treatment choice that is on the contrary to the prediction: low PS treated individuals and high PS untreated individuals. They argued that these individuals were more likely to have strong unmeasured risk factors influencing the observed treatment choice.

Another trimming strategy proposed by Walker *et al.*(6) is defined on the scale of the *preference score*, which is a monotone transformation of the PS, adjusting for treatment prevalence *p* and denoted as π_i here.

$$I_w = \{i \in I : \pi_i \in [\alpha_w, 1 - \alpha_w]\}$$

They used $\alpha_w = 0.3$ although it was not validated. The rationale for this trimming strategy is to keep patients with PS close to the mean PS in the trimmed cohort. The mean PS in the population equals the treatment prevalence (**eAppendix** 1.3). Therefore, one can argue that those individuals with $e_i = p$ are the *average* patients most representative of the population of interest. The preference score transformation re-centers the distribution around such average patients. As a result, the trimming thresholds on the *preference score* scale are symmetric around 0.5.

Extension to the multinomial setting

In the two group setting of treated vs untreated, we only need to consider one scalar PS for the probability of being *treated,* $P[A_i = 1 \mid X_i]$. However, in the multinomial setting with *J*+1 treatment groups, it helps to consider a PS vector $\boldsymbol{e_i} = (e_{0i}, e_{1i}, \dots , e_{Ji})^T$ having one probability of assignment for each one of the *J*+1 treatment groups(9) where $e_{ji} = P[A_i = j \mid X_i]$ for $j \in \{0,1, \dots, J\}$. The sum of the *J*+1 elements is constrained to one. We introduced corresponding generalization of the preference score transformation using the group prevalence p_j (**eAppendix** 2.3).

We can extend the definition of trimming using these generalized definitions of scores. The proposed definitions for the setting with $J + 1$ treatment groups are given in **Table 4-1**. Multinomial Crump trimming retains subjects who have *all* PSs above the threshold $\alpha_{J,c}$. Multinomial Stürmer trimming is asymmetric in that the lower threshold for each PS is different unlike multinomial Crump trimming. The lower threshold is the $100\alpha_{J,c}$ percentile of each PS in the corresponding treatment group. Multinomial Walker trimming is similar to multinomial Crump trimming except the use of a preference score in place of PS. We only define the lower threshold. Trimming the upper tail is implicit because individuals who have a very *high* PS for one treatment have very *low* PSs for the other treatments. These definitions reduce to the original definitions when there are only two groups (**eAppendix** Section 2). These lower thresholds are indexed with J to indicate the need to adjust for the number of groups $J + 1$. This adjustment is required because the threshold values used in the two-group setting can become too strict as the number of treatment groups increases (**eAppendix** 2.4). We used tentative values for our three-group empirical illustration (**Table 4-2**).

Empirical data illustration in the three-group setting

We illustrated how the trimming methods worked in the three-group setting using observational datasets(10,11) and visualization with ternary plots(12). A ternary plot is a triangle-shaped two-dimensional representation of three-dimensional data that sum to a constant (**eAppendix** Section 3). A point distant from a corner, for example, far from the top corner labeled 0, represents an individual with a low probability of being in group 0. The mid-point represents an individual with equal probabilities for all three groups. An interactive web application that emulates a PS distribution is available at https://kaz-yos.shinyapps.io/shiny_trim_ternary/ (**eAppendix** 3.4).

Figure 4-2 shows the results of the three trimming methods on three different observational datasets. All proposed multinomial trimming methods resulted in triangular retention regions. Crump trimming resulted in fixed trimming bounds regardless of the PS distribution. The other two methods were adaptive to the observed PS distribution. In the example of three COX2 selective inhibitors(10), all three groups were of similar sizes (32,684 celecoxib users, 24,124 rofecoxib users, and 26,582 valdecoxib users) and had comparable distributions of patient characteristics, resulting in a concentrated cluster of all three groups on top of each other. Crump trimming retained all subjects. The other two methods retained most subjects.

We found 23,532 naproxen users, 21,880 ibuprofen users, and 5,261 diclofenac users in the non-selective NSAID example(10). Users were still similar across treatment groups as illustrated by their clustering, but the small size of diclofenac group resulted in the off-centered location of the observations and off-centered bounds for Stürmer and Walker trimming. All three methods trimmed similar proportions in this specific instance.

When the indications were different, as expected in the diabetes medication example(11), the distribution of PS became more visibly separated (distinct colors), leading small percentages of subjects remaining after trimming. This can reduce efficiency, but more importantly, it may be necessary to narrow cohort eligibility criteria to provide more comparable groups. We had a disproportionately large sulfonylurea group (n = 113,429), followed by the insulin (n = 18,294), and the GLP-1 agonist (n = 14,278) groups. This imbalance again resulted in off-centered bounds for Stürmer and Walker trimming.

Simulation setup

We conducted a simulation study to examine the influence of the proposed multinomial PS trimming methods in combination with different PS confounding adjustment methods on bias and efficiency in the setting of three treatment group CER. The reporting follows Morris *et al.*'s recommendation(13). The simulation suite written in R is available at https://github.com/kaz-yos/multinomial-ps-trimming.

Data generating mechanism. We detailed the formulation of the data generating models as well as their parameters in the **eAppendix** Section 4. Briefly, to introduce unmeasured confounders in the tails of the PS distribution, we extended the data generating mechanism developed by Stürmer in the two-group setting to the three-group setting(5). Covariates X_1 through X_6 were considered the base variables that were measured, whereas covariates X_7 through X_9 were considered the rare confounders that remained unmeasured. As in Stürmer *et al.*, we calculated a *tentative* PS based on the measured covariates. The unmeasured binary covariates were then generated based on the tentative PS such that X_7=1 was more prevalent in those who had a high tentative propensity for group 0; X_8=1 was more prevalent in those who had a high tentative propensity for group 1; and X_9=1 was more prevalent in those who had a high tentative propensity for group 2. After constructing the full set of covariates both measured and unmeasured, the *true* PS was assigned based on coefficients given to all the covariates. The unmeasured covariates had strong "contraindication effects." For example, when X_7 was present in an individual with a high tentative propensity of receiving treatment 0, this treatment assignment became much less likely (X_7 serving as a strong contraindication to an otherwise preferred treatment). Treatment A_i was then generated as a three-group multinomial random variable taking on one of {0, 1, 2}. The outcome Y_i was a Poisson count random variable based on a linear predictor dependent on all the covariates and treatment. A log-link model was chosen to eliminate the problem of non-collapsibility(14), which complicates the calculation of true effects (**eAppendix** 4.4).

Methods to be evaluated. We compared the three types of multinomial PS trimming methods defined above in combination with different confounding adjustment methods. Each trimming method was examined at several trimming thresholds to compare alternative cutoffs (**eAppendix** 4.3). We used the three-group IPTW(7), matching weights (MW)(15,16), and overlap weights (OW) (17–19) as confounding adjustment methods. Consideration of these three weighting schemes permitted evaluation of the sensitivity of any observed benefit of trimming to this choice.

Estimand of interest. We estimated the alternatively weighted log rate ratios for group 1 vs. group 0, group 2 vs. group 0, and group 2 vs. group 1 contrasts in the overall study population as well as PS trimmed cohort (**eAppendix** 4.4).

Performance measures. The trimmed sample size, bias, standard error (SE), and mean squared errors (MSE) were examined (**eAppendix** 4.5).

RESULTS

We examined nine scenarios of varying data configurations, each run 500 times. **Figure 4-3** shows the sample size decrease after trimming (methods as the columns of panels) at different thresholds (X-axis). The strength of unmeasured confounding did not affect the proportion of trimmed observations because this strength of unmeasured confounding was manipulated by changing the coefficients for the outcome-generating model, but not the treatment-generating model. The size of trimmed cohorts after trimming differed by the treatment prevalence in the Crump trimming because these trimming thresholds did not adapt to the skewed distribution of PS as seen in the empirical examples (**Figure 4-2**). In the 10:10:80 setting, in particular, the center of the PS distribution was close to group 2 (right lower corner in the ternary plot), resulting in a larger proportion of the cohort trimmed by this method. Walker trimming provided most similar numbers of patients remaining in the cohort regardless of the treatment prevalence. This is because the Walker trimming region is around the average PS, *i.e.*, a region where the treatment prevalence coincides with the full-sample prevalence.

Figure 4-4 illustrates the bias in the setting of moderate unmeasured confounding with different treatment prevalence, various trimming methods, and trimming thresholds. The bias in the unadjusted analysis at trimming threshold zero (no trimming) shows the direction and magnitude of the total confounding including both measured and unmeasured confounding. As expected from the principle of restriction as a measure to control confounding (if variables do not vary in the analysis cohort, they cannot confound), trimming reduced the bias in unadjusted analyses until the threshold where the cohort became too small for outcome analyses. Use of MW and OW resulted in a reduction of bias even without trimming. However, small bias persisted in the other direction except for the 1 vs. 0 contrast. Bias of similar magnitude appeared in the other direction with IPTW. Reduction in residual confounding was seen for all weighting methods. The only exception was that in the 10:10:80 treatment prevalence scenario, the bias increased for 2 vs. 0 and 2 vs. 1 contrasts with Crump trimming beyond the 1/60 threshold. The reason for exacerbated bias seems to be the very skewed PS distribution. The average PS vector corresponded to the marginal prevalence, *i.e.*, $(0.1, 0.1, 0.8)^T$. Therefore, group 2 would distribute closer to the left lower corner in the ternary plot, preferentially trimmed by Crump trimming. Estimation was less reliable for contrasts involving group 2 as a result. Stürmer and Walker trimming performed similarly regardless of the treatment prevalence. Overcorrection occurred with PS trimming in contrast 1 vs 0, in which MW and OW did not have residual confounding. Further trimming resulted in a return to less biased estimates.

Figure 4-5 illustrates the corresponding simulation standard error (SE) of estimates. IPTW SE took a convex shape, initially benefiting from trimming, but eventually increasing due to the small sample sizes after trimming. This IPTW SE reduction appeared in all three trimming methods although only Crump trimming was proposed for improved precision. Among the thresholds examined in the simulation, the smallest SE was attained at around 0.07 for Crump, 0.03 for Stürmer, and 0.1 for Walker trimming, suggesting the rule-of-thumb threshold of 0.2 for Walker trimming may be too strict and increases SE. Neither MW nor OW SE clearly benefited from trimming. Stürmer trimming, in particular, resulted in a quick increase in SE with MW and OW. Compared to other methods, Crump trimming seemed to offer the minimum IPTW SE in the absence of unmeasured confounding (**eAppendix** 5.2.1).

Figure 4-6 illustrates the MSE of the estimators calculated as variance + bias^2, which represents the variability around the true value of the parameter. The variance term dominated the bias term with moderate unmeasured confounding. For IPTW, the minimum MSE was achieved at around 0.07 with Crump trimming, 0.017 for Stürmer trimming, and 0.05-0.10 for Walker trimming. The results for MW and OW were similar although the initial decrease in MSE was seen only in some settings (2 vs. 0 and 2 vs. 1 contrasts, particularly with 10:45:45 treatment prevalence). For the 1 vs. 0 contrast, no apparent benefit was observed with any of the trimming methods or thresholds.

DISCUSSION

Several PS trimming methods have been proposed to improve the validity and efficiency of two-group observational studies requiring PS-based confounding control(4–6). We extended these trimming methods to the multinomial treatment setting and conducted a simulation study in the three-group setting. We specifically examined the interplay of bias introduced by confounders present in the tails of PS distribution and the variance of estimators with increasing trimming. All methods reduced bias in IPTW, MW, and OW estimators in most scenarios. However, multinomial Stürmer and Walker trimming were more successful in bias reduction when three treatment groups had very different sizes (10:10:80) skewing the PS distribution. Trimming a small fraction of observations in all three methods decreased variance for IPTW, not for MW or OW. At the proposed rule-of-thumb thresholds, multinomial Crump and Stürmer trimming achieved variance reduction better.

For the specific purpose of reducing bias by unmeasured confounders in the tails of multinomial PS distributions, Stürmer and Walker trimming may be better suited when the prevalence of treatment groups is quite different. Stürmer *et al*. suggested that this type of unmeasured confounding bias might be a reason for apparent "treatment effect heterogeneity" (truly a bias) seen in the tails of binary PS in the two-group observational study setting.(5,20) This bias can also happen in the multinomial setting in the presence of a strong indication for one of the drugs or a strong contraindication against one of the drugs that are unmeasured. Diabetes medications provide an illustrative example (**Figure 4-2**). Those who have severe diabetes and observable clinical indications for insulin may be found in one of the oral medication groups. Such patients are more likely to have unobserved contraindications for insulin such as frailty, which could strongly influence many outcomes. We simulated this type of setting and demonstrated that trimming reduced the bias by strong unmeasured contraindications.

Progressively stricter trimming reduced bias, but this was at the cost of efficiency once the trimmed sample size became too small. In the simulation scenarios that we examined, we found that relatively limited PS trimming gave the best balance of bias and variance as assessed by MSE (**Figure 4-7**). The rule-of-thumb threshold for Walker trimming may be overly strict.

Another critical trade-off is the changing estimand when treatment effect heterogeneity exists. The target of inference, the population of individuals for whom we estimate the treatment effects, changes with trimming. Although PS trimming, a form of restriction, is expected to improve the validity of inference as long as all groups are trimmed in the same manner (21), the generalizability may be compromised. However, the type of patients retained after trimming can be argued as patients with reasonable chances of being assigned to any of the treatment groups, *i.e.*, individuals for whom CER is most relevant(6). In practice, one should vary the trimming threshold to examine the sensitivity of the results related to progressively stricter trimming thresholds(22).

Our focus was bias by unmeasured risk factors that were more prevalent in the tails of PS distribution. This focus can be considered a multinomial equivalent of what Stürmer *et al.* examined(5). Importantly, the original intentions of the other Crump *et al.*'s(4) and Walker *et al.'s*(6) methods were somewhat different from Stürmer *et al.*'s(5). Crump *et al.*'s paper(4) emphasized the efficiency argument given that the PS model was correct and unmeasured confounding was absent. Their method's strength is the proven minimum variance with IPTW under some constraints although multinomial Crump trimming also reduced residual bias in most settings in our simulation. Interestingly, multinomial Stürmer and Walker trimming also reduced the variance of the IPTW estimator albeit to a lesser extent. MW(15,16) and OW(17–19) were more efficient than IPTW, thus, no trimming methods examined improved the efficiency of MW or OW estimators. One might argue that PS trimming is of little benefit for MW and OW. However, small bias reduction did occur even for MW and OW. Walker *et al.*'s paper(6) focused primarily on identifying CER settings where unmeasured confounding might be less of a concern. The tool's role as a trimming tool was secondary. In our simulation study focusing on reducing unmeasured confounding bias in a given dataset, we found that smaller thresholds ([0.05, 1.0] to [0.10, 1.0]) were sufficient to reduce confounding.

Another potential approach to unmeasured confounding worth mentioning is PS calibration(23,24). The important difference here is the requirement for an additional external validation dataset which contains variables that are unmeasured confounders in the main dataset. Our use of PS trimming to control for unmeasured confounding instead relies on the assumption that the tails of the PS contain individuals with unmeasured factors.

Although our definitions of multinomial PS trimming are natural extensions of the original binary PS trimming, they are not the only extensions. For example, PS trimming can be extended by considering all possible pairwise PS rather than the single multinomial PS. However, the complexity of implementation increases more rapidly for the pairwise definition than for the multinomial definition. Importantly, all pairwise PS must be defined for all patients. The pairwise PS for the A vs B contrast is estimated on groups A and B. However, we must assign this pairwise PS for the A vs B contrast even for those who are in group C. This counterintuitive approach is necessary to define the same retention region for all treatment groups and to capture those who are in equipoise for all treatment options. Otherwise, the principle of PS methods, assuring similar distribution of covariates in all treatment groups, is violated. The multinomial approach considers all treatment groups simultaneously, thus, it is not unnatural to assign all J+1 probabilities of treatment assignment for each individual. It also has the advantage of having only one PS model rather than all possible pairwise PS models, which need to be fit separately on relevant pairwise subsets of the entire dataset.

Our study assumed that the relevant *a priori* clinical question was the comparison of treatment among subjects who had some chance of receiving any one of the multiple treatments. This assumption was an important rationale for modeling all groups in one multinomial PS model. On the other hand, we could construct pairwise PS and a pairwise PS trimmed cohort for each one of the pairwise contrast. The potential problem here is that each pairwise comparison may have a different target population. Having different target populations could cause non-transitive results, for example, A is better than B; B is better than C; but A is worse than C (25). The pairwise approach is more acceptable when we have one group that is the reference group or the drug-of-interest group. In this case, only the pairwise contrasts involving this one group are relevant, making non-transitivity less of a concern. These two approaches may result in similar and transitive effect estimates if those who are in pairwise equipoise are also in equipoise among all groups. If this does not hold, the multinomial trimming likely results in a small trimmed cohort as the separation between groups in the PS space may be greater. Ideally, investigators should assess the appropriateness of a multi-group CER question *a priori*. When multinomial PS trimming results in a much smaller cohort than the original, one may need to reconsider whether the data and eligibility criteria give sufficient overlap among groups to justify multi-group CER (6,26).

The implication of a simulation study should be considered within the limitations of the data generation process. We introduced unmeasured confounding in the tails of PS distributions similarly to Stürmer *et al.*(5). Further studies are required for other specific data generation mechanisms. The use of a count outcome in our simulation was for simplicity and consistency with a previous study(5) and the trimming methods are not limited to this outcome type.

However, how these multinomial trimming methods perform in practice with other outcome types needs a future examination.

In conclusion, we proposed a multinomial extension of the existing two-group PS trimming methods and examined their performance with three treatment groups. The extensions of Stürmer and Walker's PS trimming methods reduced bias in 3-group exposure settings even with highly imbalanced treatment frequencies. In practice, examining how effect estimates vary at various trimming thresholds can be a useful sensitivity analysis to assess potential unmeasured confounding in the tails of a multinomial PS.

References

1. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika*. 1983;70(1):41–55.

2. Rosenbaum PR, Rubin DB. Reducing Bias in Observational Studies Using Subclassification on the Propensity Score. *J Am Stat Assoc*. 1984;79(387):516.

3. Rosenbaum PR, Rubin DB. Constructing a Control Group Using Multivariate Matched Sampling Methods That Incorporate the Propensity Score. *The American Statistician*. 1985;39(1):33–38.

4. Crump RK, Hotz VJ, Imbens GW, et al. Dealing with limited overlap in estimation of average treatment effects. *Biometrika*. 2009;96(1):187–199.

5. Stürmer T, Rothman KJ, Avorn J, et al. Treatment effects in the presence of unmeasured confounding: dealing with observations in the tails of the propensity score distribution--a simulation study. *Am. J. Epidemiol.* 2010;172(7):843–854.

6. Walker AM, Patrick AR, Lauer MS, et al. A tool for assessing the feasibility of comparative effectiveness research. *Comp Eff Res*. 2013;11.

7. Robins JM, Hernán MA, Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology*. 2000;11(5):550–560.

8. Yoshida K, Solomon DH, Kim SC. Active-comparator design and new-user design in observational studies. *Nat Rev Rheumatol*. 2015;11(7):437–441.

9. Imbens GW. The role of the propensity score in estimating dose-response functions. *Biometrika*. 2000;87(3):706–710.

10. Solomon DH, Rassen JA, Glynn RJ, et al. The comparative safety of analgesics in older adults with arthritis. *Arch. Intern. Med.* 2010;170(22):1968–1976.

11. Patorno E, Everett BM, Goldfine AB, et al. Comparative cardiovascular safety of glucagon-like peptide-1 receptor agonists versus other antidiabetic drugs in routine care: a cohort study. *Diabetes Obes Metab*. 2016;18(8):755–765.

12. Hamilton N. ggtern: An Extension to "ggplot2", for the Creation of Ternary Diagrams. 2017 (Accessed January 25, 2018).(https://cran.r-project.org/web/packages/ggtern/index.html). (Accessed January 25, 2018)

13. Morris TP, White IR, Crowther MJ. Using simulation studies to evaluate statistical methods. *arXiv:1712.03198 [stat]* [electronic article]. 2017;(http://arxiv.org/abs/1712.03198). (Accessed January 9, 2018)

14. Greenland S, Robins JM, Pearl J. Confounding and Collapsibility in Causal Inference. *Statistical Science*. 1999;14(1):29–46.

15. Li L, Greene T. A weighting analogue to pair matching in propensity score analysis. *Int J Biostat*. 2013;9(2):215–234.

16. Yoshida K, Hernandez-Diaz S, Solomon DH, et al. Matching Weights to Simultaneously Compare Three Treatment Groups: Comparison to Three-way Matching. *Epidemiology*. 2017;28(3):387–395.

17. Li F, Morgan KL, Zaslavsky AM. Balancing Covariates via Propensity Score Weighting. *Journal of the American Statistical Association*. 2016;0(0):1–11.

18. Li F, Thomas LE, Li F. Addressing Extreme Propensity Scores via the Overlap Weights. *Am. J. Epidemiol.* 2018;

19. Li F, Li F. Propensity Score Weighting for Causal Inference with Multi-valued Treatments. *arXiv:1808.05339 [stat]* [electronic article]. 2018;(http://arxiv.org/abs/1808.05339). (Accessed August 23, 2018)

20. Kurth T, Walker AM, Glynn RJ, et al. Results of multivariable logistic regression, propensity matching, propensity adjustment, and propensity-based weighting under conditions of nonuniform effect. *Am. J. Epidemiol.* 2006;163(3):262–270.

21. Schneeweiss S, Patrick AR, Stürmer T, et al. Increasing levels of restriction in pharmacoepidemiologic database studies of elderly and comparison with randomized trial results. *Med Care*. 2007;45(10 Supl 2):S131-142.

22. Wyss R, Stürmer T, Joshua GJ, et al. Propensity Score Trimming to Enhance Validity in Comparative E ectiveness Research [Abstract]. *Pharmacoepidemiol Drug Saf.* 2017;26:92.

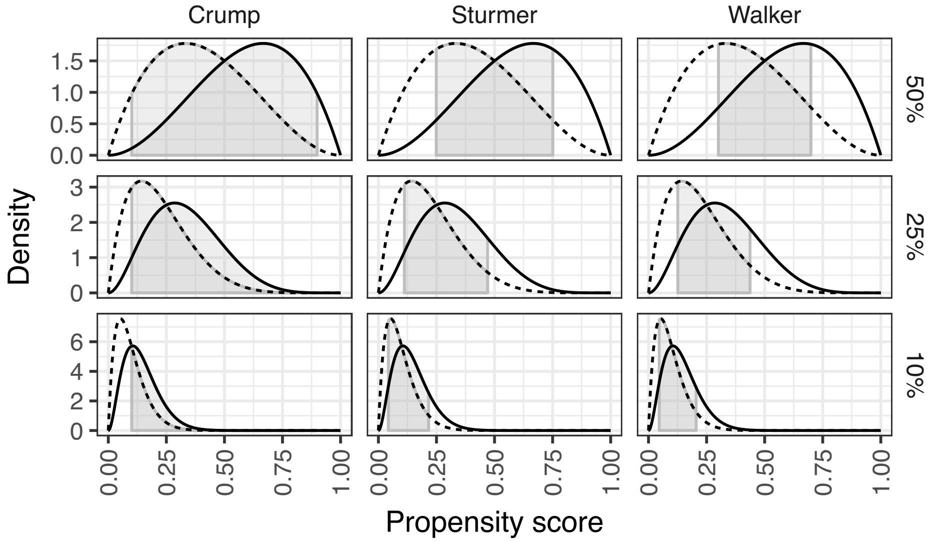
23. Stürmer T, Schneeweiss S, Avorn J, et al. Adjusting effect estimates for unmeasured confounding with validation data using propensity score calibration. *Am. J. Epidemiol.* 2005;162(3):279–289.

24. Stürmer T, Schneeweiss S, Rothman KJ, et al. Performance of propensity score calibration--a simulation study. *Am. J. Epidemiol.* 2007;165(10):1110–1118.

25. Rassen JA, Shelat AA, Franklin JM, et al. Matching by propensity score in cohort studies with three treatment groups. *Epidemiology*. 2013;24(3):401–409.

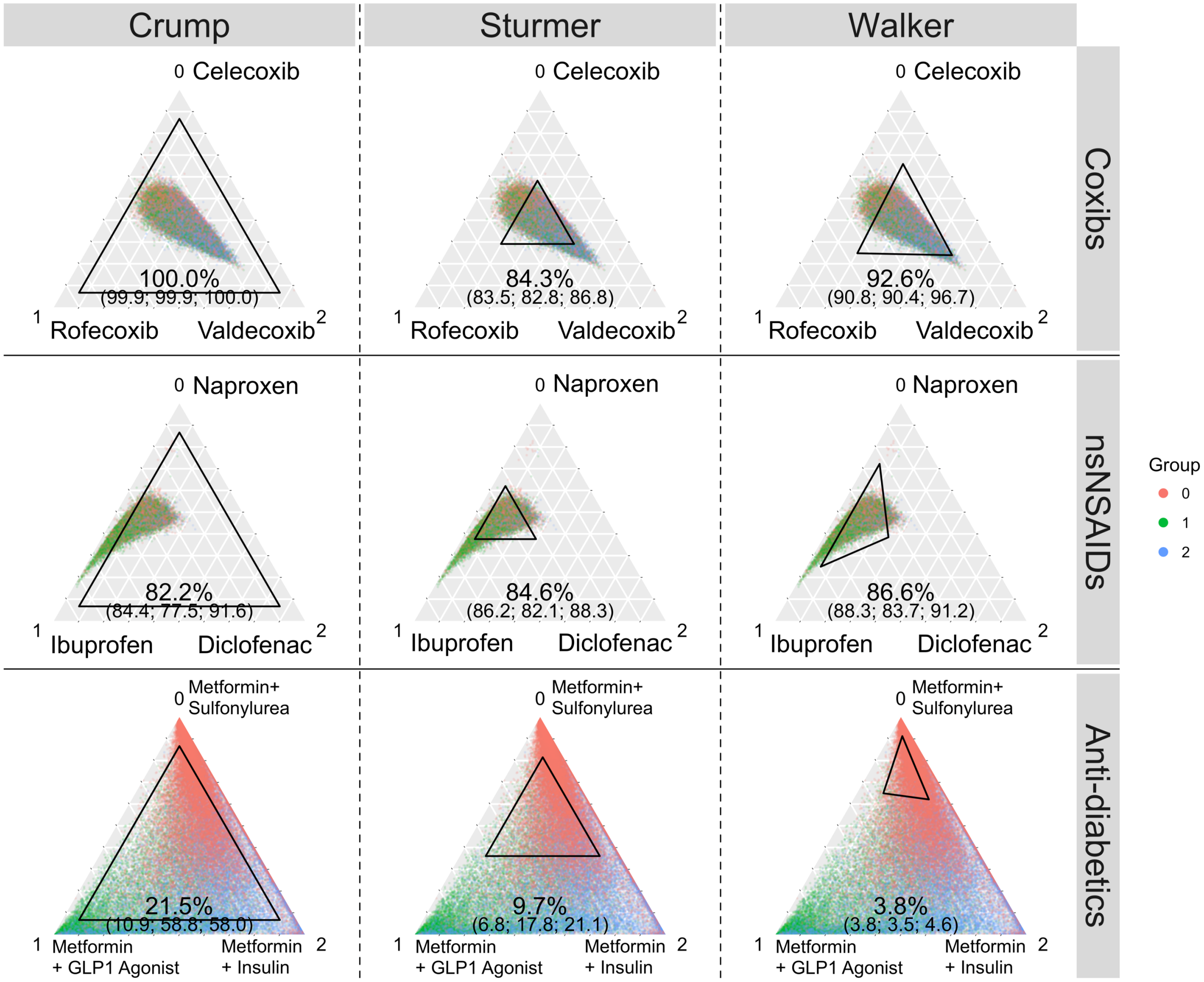
26. Girman CJ, Faries D, Ryan P, et al. Pre-study feasibility and identifying sensitivity analyses for protocol pre-specification in comparative effectiveness research. *J Comp Eff Res*. 2014;3(3):259–270.

Figure 4-1. Visual explanation of three existing two-group trimming methods.



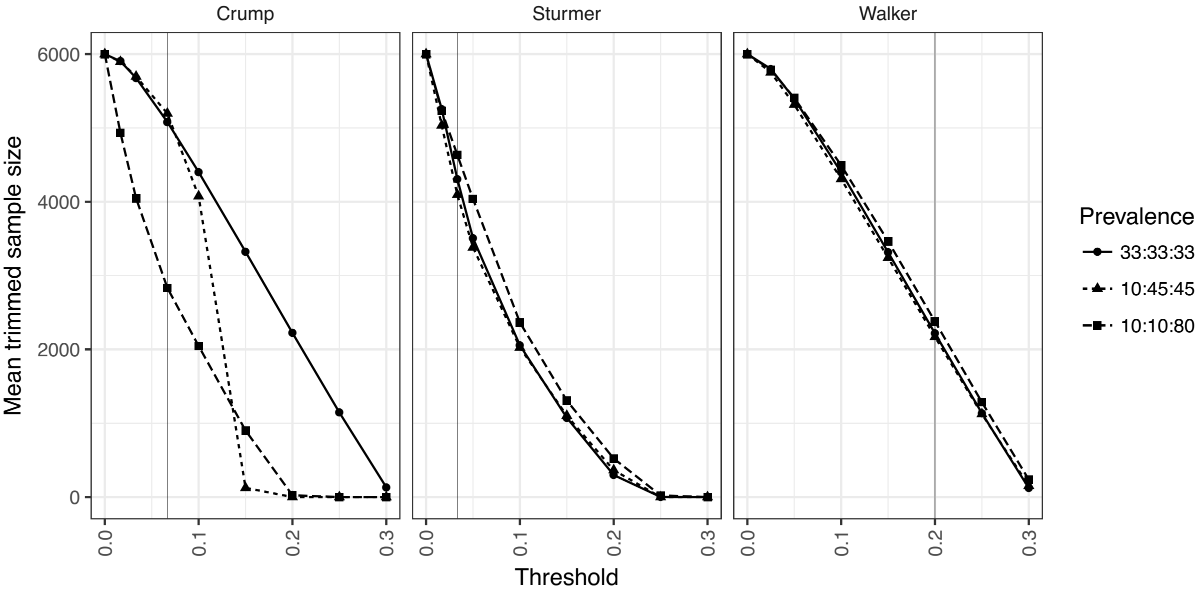
The hypothetical PS distribution densities were generated from beta distributions. The dotted line represents the propensity score (PS) density in the untreated, whereas the solid line represents the PS density in the treated. The columns of panels represent three trimming methods. The rows of panels represent treatment prevalence (50%, 25%, and 10%). In each panel, the gray region represents the retention region that applies to *both* treated and untreated groups. Individuals outside the retention region are removed *regardless of their treatment status*. Crump trimming is the same regardless of the prevalence, whereas the other two methods adapt to skewed PS distributions due to less frequent treatment. See **eAppendix** 1.4 for further examples.

Figure 4-2. Ternary plots of trimming results in empirical datasets.



The rows represent datasets: coxibs, non-selective non-steroidal anti-inflammatory drugs (nsNSAIDs), and anti-diabetics. The columns represent the trimming methods: Crump, Stürmer, and Walker. The inner black triangles are the trimming thresholds. The numbers in the triangles indicate the proportion (%) of the original cohort that remained after trimming as well as group-wise proportions. The groups are: (0) celecoxib, (1) rofecoxib, and (2) valdecoxib for coxibs; (0) naproxen, (1) ibuprofen, and (2) diclofenac for nsNSAIDs; and (0) sulfonylurea + metformin, (1) glucagon-like peptide-1 receptor agonist + metformin, and (2) insulin + metformin for anti-diabetics.

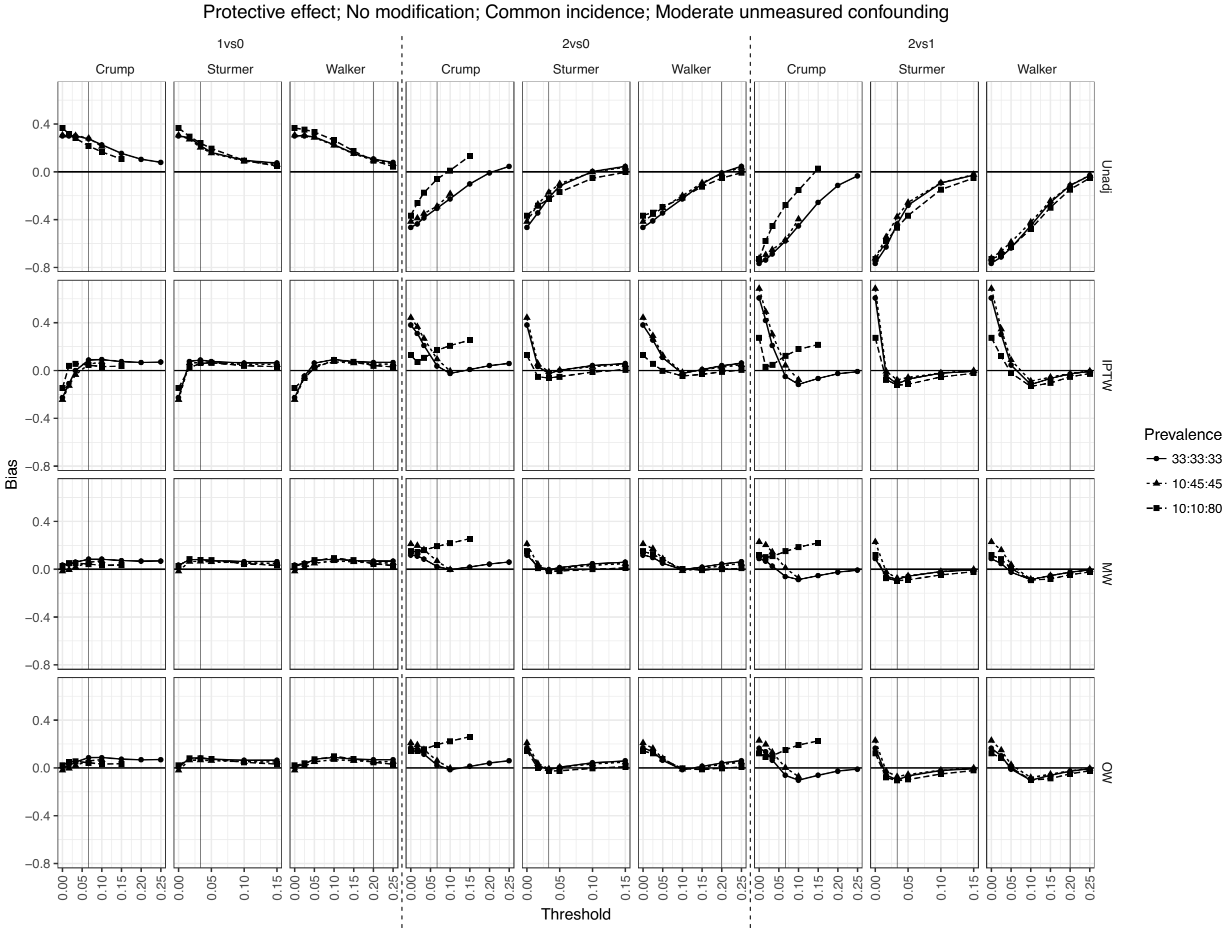
Figure 4-3. Simulated samples size after trimming at different thresholds.



The scales for the thresholds were the propensity score (PS) scale for Crump, quantiles of PS for Stürmer, and the preference score scale for Walker. The vertical hairlines are at the tentative thresholds used for the empirical data illustration (**Figure 4-1**).

The original sample size was n = 6,000 in all prevalence scenarios. Both Stürmer and Walker methods trimmed similarly regardless of treatment prevalence as they accommodated skewed PS distribution**s**. Crump trimming, on the other hand, trimmed differently at the same trimming threshold across treatment prevalence scenarios.

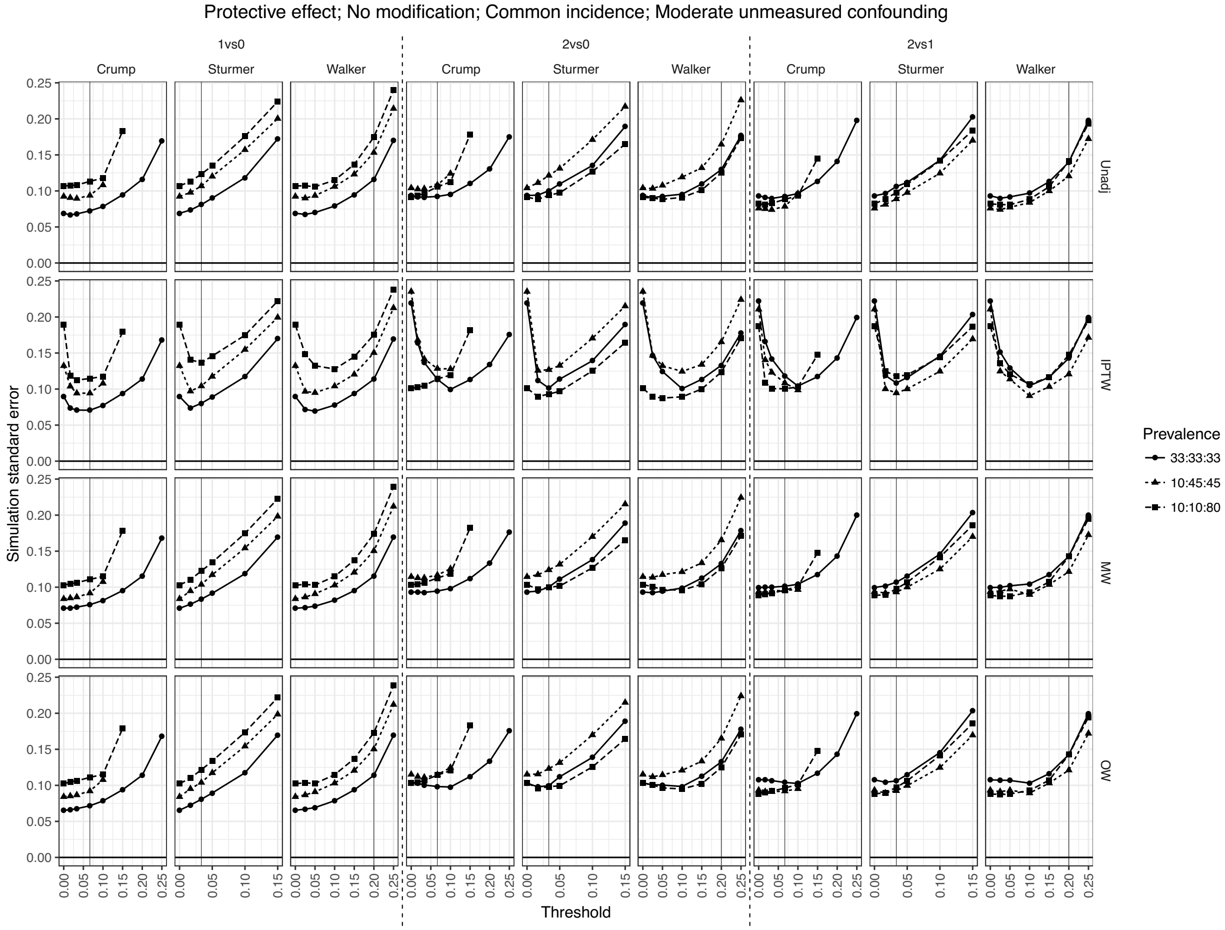
Figure 4-4. Bias results from the moderate unmeasured confounding setting.



Abbreviations: 1vs0: group 1 vs group 0 treatment contrast; 2vs0: group 2 vs group 0 treatment contrast; 2vs1: group 2 vs group 1 treatment contrast; IPTW: inverse probability of treatment weights; MW: matching weights; OW: overlap weights.

The rows of panels represent confounding adjustment methods: unadjusted, IPTW, MW, and OW. The columns of panels represent the group contrast and then trimming methods. Within each panel, the X-axis represents progressive increase in trimming threshold (more observations are trimmed off). The vertical hairlines are at the tentative thresholds used for the empirical data illustration (**Figure 4-1**).

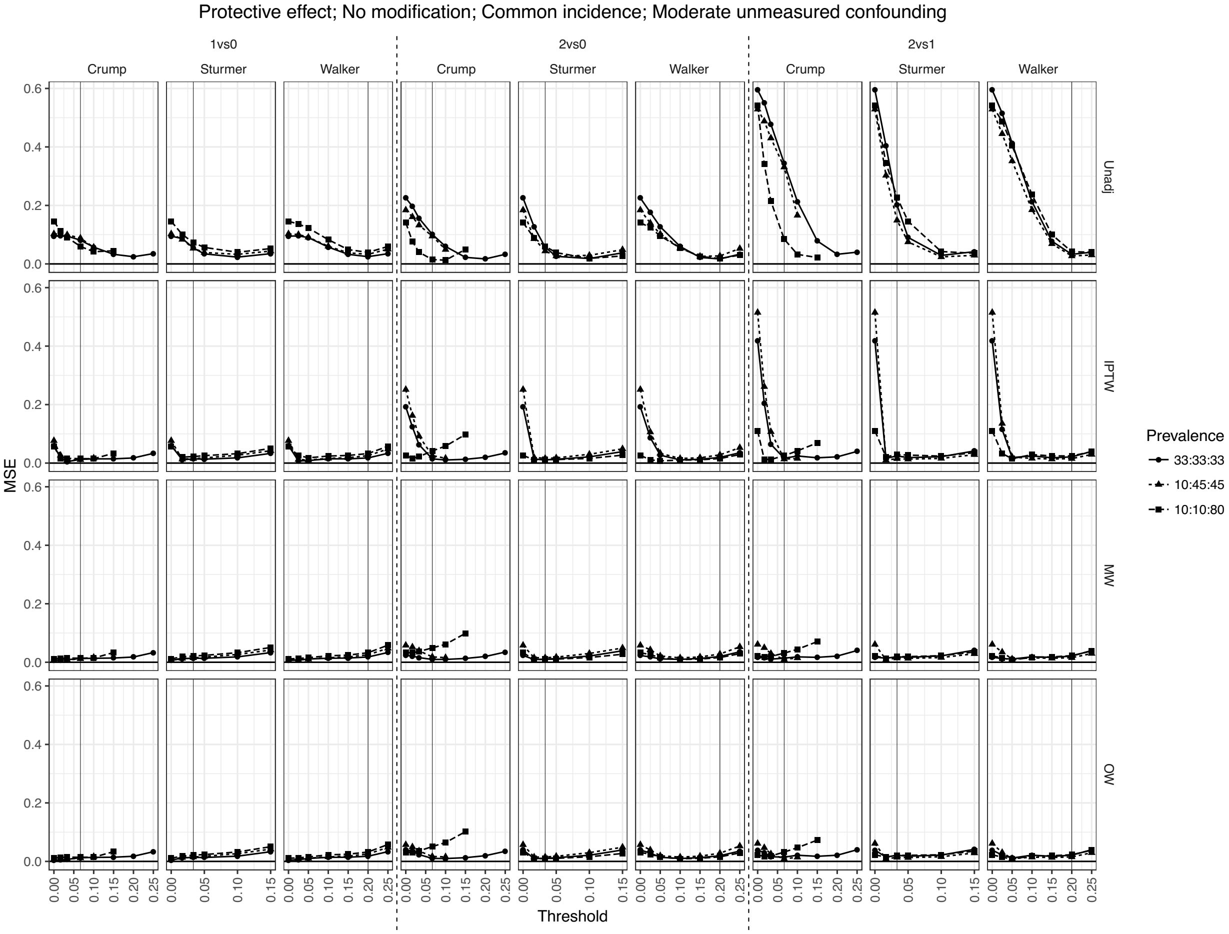
Figure 4-5. Standard error results from the moderate unmeasured confounding setting.



Abbreviations: 1vs0: group 1 vs group 0 treatment contrast; 2vs0: group 2 vs group 0 treatment contrast; 2vs1: group 2 vs group 1 treatment contrast; IPTW: inverse probability of treatment weights; MW: matching weights; OW: overlap weights.

The rows of panels represent confounding adjustment methods: unadjusted, IPTW, MW, and OW. The columns of panels represent the group contrast and then trimming methods. Within each panel, the X-axis represents progressive increase in trimming threshold (more observations are trimmed off). The vertical hairlines are at the tentative thresholds used for the empirical data illustration (**Figure 4-1**).

Figure 4-6. Mean squared error (MSE) results from the moderate unmeasured confounding setting.



Abbreviations: 1vs0: group 1 vs group 0 treatment contrast; 2vs0: group 2 vs group 0 treatment contrast; 2vs1: group 2 vs group 1 treatment contrast; IPTW: inverse probability of treatment weights; MW: matching weights; OW: overlap weights.

The rows of panels represent confounding adjustment methods: unadjusted, IPTW, MW, and OW. The columns of panels represent the group contrast and then trimming methods. Within each panel, the X-axis represents progressive increase in trimming threshold (more observations are trimmed off). The vertical hairlines are at the tentative thresholds used for the empirical data illustration (**Figure 4-1**).

Table 4-1. Existing propensity score trimming method definitions for a binary treatment and proposed propensity score trimming method definitions for a multinomial treatment

	Original binary definition	**Proposed multinomial definition**
Crump *et al.*(4)	$I_c = \{i \in I: e_i \in [\alpha_c, 1-\ \alpha_c]\}$ Their rule-of-thumb threshold for α_c was 0.1.	$I_{J,c} = \{i \in I_J: e_{ji} \geq \alpha_{J,c}\ \forall\ j \in \{0,1,\dots,J\}\}$
Stürmer *et al.*(5)	$I_s = \{i \in I: e_i \in [F^{-1}_{e_i\|A_i}(\alpha_s\|1), F^{-1}_{e_i\|A_i}(1-\alpha_s\|0)\ ]\}$ α_s were 0.01, 0.025, and 0.05 in their simulation.	$I_{J,s} = \{i \in I_J: e_{ji} \geq F^{-1}_{e_{ji}\|A_i}(\alpha_{J,s}\|j)\ \forall\ j \in \{0,1,\dots,J\}\}$
Walker *et al.*(6)	$I_w = \{i \in I: \pi_i \in [\alpha_w, 1-\ \alpha_w]\}$ Their rule-of-thumb threshold for α_w was 0.3.	$I_{J,w} = \{i \in I_J: \pi_{ji} \geq \alpha_{J,w}\ \forall\ j \in \{0,1,\dots,J\}\}$
Notations	$I = \{1, \dots, n\}$ Set of individual indices I_x Subset of individual indices retained by method x $A_i \in \{0,1\}$ Binary treatment indicator for individual i e_i Propensity score for individual i $F^{-1}_{ei\|Ai}$ Inverse cumulative distribution function of e_i conditional on A_i π_i Preference score for individual i (See text) α_x Trimming threshold by method x	$I_J = \{1, \dots, n\}$ Set of individual indices with $J + 1$ groups $I_{J,x}$ Subset of individual indices retained by method x $A_i \in \{0,1, \dots, J\}$ Multinomial treatment indicator for individual i e_{ji} Propensity score for individual i for treatment j $F^{-1}_{eji\|Ai}$ Inverse cumulative distribution function of e_{ji} conditional on A_i π_{ji} Preference score for individual i for treatment j $\alpha_{J,x}$ Trimming threshold by method x with $J + 1$ groups

See text for interpretation and eAppendix Section 1 for rationale for each method. In all original and proposed methods, the same retention region is applied to every treatment group. See **eAppendix** Section 2 for equivalence of the proposed methods to the original binary methods and proposed tentative thresholds in the multinomial setting

Table 4-2. Tentative threshold values for each method and number of groups.

Number of groups	Crump	Stürmer	Walker
2 (original)	0.10	0.050	0.30
3 (our study)	0.07	0.033	0.20
4 (not examined)	0.05	0.025	0.15
Scale	PS	Group-specific PS quantile	Preference score

Abbreviation: PS Propensity score.
See **eAppendix** 2.4 for details.

Contents

1 Base trimming methods for the two-group setting

Here we consider a two group setting where the treatment is defined as $A_i \in \{0, 1\}$ and the propensity score (PS) as a function of the covariate vector $\mathbf{X}_i$ is defined as $e_i = P[A_i = 1|\mathbf{X}_i] \in (0, 1)$. Let $I = \{1, ..., n\}$ be the set of indices indexing all the individuals in the study cohort.

Let $F_{e_i}(\cdot)$ be the cumulative distribution function (CDF) of the PS. A preference score [Walker et al., 2013] π_i is a one-to-one transformation of PS such that $\text{logit}(\pi_i) = \text{logit}(e_i) - \text{logit}(p)$ where $p = P[A_i = 1] = E[e_i]$ is the prevalence of treatment, which equals the mean PS.

Using these notations, the subset of indices retained after each trimming method can be written as follows.

Method	Definition
Crump	$I_c = \{i \in I : e_i \in [\alpha_c, 1 - \alpha_c]\}$
Stürmer	$I_s = \left\{i \in I : e_i \in \left[F^{-1}_{e_i\|A_i}(\alpha_s\|1), F^{-1}_{e_i\|A_i}(1 - \alpha_s\|0)\right]\right\}$
Walker	$I_w = \{i \in I : \pi_i \in [\alpha_w, 1 - \alpha_w]\}$

The rationale and detailed definition for each method is given in the following.

1.1 Crump trimming

1.1.1 Rationale

[Crump et al., 2009] used trimming for precision. They specifically utilized trimming to deal with the limited overlap of the PS distributions between the treated and the untreated patients. The inverse probability of treatment weight (IPTW) [Robins et al., 2000] can result in an imprecise estimate of the average treatment effect (ATE) due to this lack of overlap. They developed their trimming method to select the optimal subset of subjects for whom the ATE can be estimated most precisely. They proved that their trimming gives the most precise estimate under the assumptions of no unmeasured confounding, positivity [Petersen et al., 2012], homoscedastic outcome.

1.1.2 Definition

The Crump trimming method is defined with fixed bounds on the PS scale as follows.

$$I_c = \{i \in I : e_i \in [\alpha_c, 1 - \alpha_c]\}$$

Those who have PS outside the retention region $[\alpha_c, 1 - \alpha_c]$ are trimmed. The most precise estimate is obtained at a specific choice of α_c that has to be estimated. In practice, they suggested using $\alpha_c = 0.1$ as a rule-of-thumb threshold that is a good approximation for a wide range of PS distributions. We adopted this threshold.

1.2 Stürmer trimming

1.2.1 Rationale

[Stürmer et al., 2010] used trimming for confounding control. Specifically, they reasoned that those with a treatment choice contrary to the choice predicted by the working PS model might have unmeasured risk factors, such as frailty, that motivated the treatment decision. Treated individuals with very low PSs and untreated individuals with very high PSs raise such concerns. They designed their trimming method such that those with a treatment choice contrary to their PSs are removed.

1.2.2 Definition

Their trimming method is defined as follows using the $100 \times \alpha_s$ th percentile of the PS among the treated patients $F^{-1}_{e_i|A_i}(\alpha_s|1)$ and the $100 \times \alpha_s$ th percentile of the PS among the untreated $F^{-1}_{e_i|A_i}(1 - \alpha_s|0)$.

$$I_s = \left\{i \in I : e_i \in \left[F^{-1}_{e_i|A_i}(\alpha_s|1), F^{-1}_{e_i|A_i}(1 - \alpha_s|0)\right]\right\}$$

Note that the retention region $[L, U]$ where $L = F^{-1}_{e_i|A_i}(\alpha_s|1)$ and $U = F^{-1}_{e_i|A_i}(1 - \alpha_s|0)$ applies to *both* untreated and treated. That is, the range restriction on PS is the *same* for the untreated and treated groups although this point may be somewhat unclear in the original paper. Their simulation examined $\alpha_s \in \{0.01, 0.025, 0.05\}$. We adopted $\alpha_s = 0.05$.

1.3 Walker trimming

1.3.1 Rationale

[Walker et al., 2013] proposed a covariate overlap assessment tool based on the PS as a surrogate marker for the potential for unmeasured confounding. They defined the proportion of patients in the medium range of the *preference score* (prevalence-adjusted transformation of PS) as a measure of *empirical equipoise.* Empirical equipoise can be interpreted as the observed surrogate of the underlying level of *clinical equipoise* [Freedman, 1987]. Clinical equipoise is defined as "a state of collective uncertainty among medical providers regarding the best treatment option for a specific patient population."

Walker and colleagues reasoned that similar patients can be assigned to different treatments under this setting, resulting in a reduced concern for confounding by indication. After this initial assessment for the risk of confounding by indication, they recommended using the patients within the medium range of preference score as the analysis cohort. Therefore, this approach also constitutes another PS trimming method.

1.3.2 Definition

Their trimming method is defined on the scale of the preference score π_i.

$$I_w = \{i \in I : \pi_i \in [\alpha_w, 1 - \alpha_w]\}$$

They suggested using $\alpha_w = 0.3$ as rule-of-thumb thresholds although this value has not been systematically validated. The following equation defines the preference score π_i in terms of the PS e_i and treatment prevalence p.

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \log\left(\frac{e_i}{1-e_i}\right) - \log\left(\frac{p}{1-p}\right)$$

Note that the prevalence p is the mean PS.

$$\begin{aligned} p &= P[A_i = 1] \\ &= E\left[A_i\right] \\ &= E\left[E\left[A_i|\mathbf{X}_i\right]\right] \\ &= E\left[P\left[A_i = 1|\mathbf{X}_i\right]\right] \\ &= E\left[e_i\right] \end{aligned}$$

We can solve for π_i and e_i as follows.

$$\begin{aligned} \log\left(\frac{\pi_i}{1-\pi_i}\right) &= \log\left(\frac{e_i}{1-e_i}\right) - \log\left(\frac{p}{1-p}\right) \\ &= \log\left(\frac{e_i}{1-e_i}\Big/\frac{p}{1-p}\right) \end{aligned}$$

As log is increasing, we have

$$\begin{aligned} \frac{\pi_i}{1-\pi_i} &= \frac{e_i}{1-e_i}\Big/\frac{p}{1-p} \\ &= \frac{e_i}{p}\frac{1-p}{1-e_i} \end{aligned}$$

$$
\begin{aligned}
\pi_i &= \frac{\frac{e_i}{p}\frac{1-p}{1-e_i}}{1+\frac{e_i}{p}\frac{1-p}{1-e_i}} \\
&= \frac{\frac{e_i}{p}}{\frac{1-e_i}{1-p}+\frac{e_i}{p}} \\
&\text{Also} \\
\frac{e_i}{1-e_i} &= \frac{\pi_i}{1-\pi_i}\frac{p}{1-p} \\
e_i &= \frac{\frac{\pi_i}{1-\pi_i}\frac{p}{1-p}}{1+\frac{\pi_i}{1-\pi_i}\frac{p}{1-p}} \\
&= \frac{\pi_i p}{(1-\pi_i)(1-p)+\pi_i p}
\end{aligned}
$$

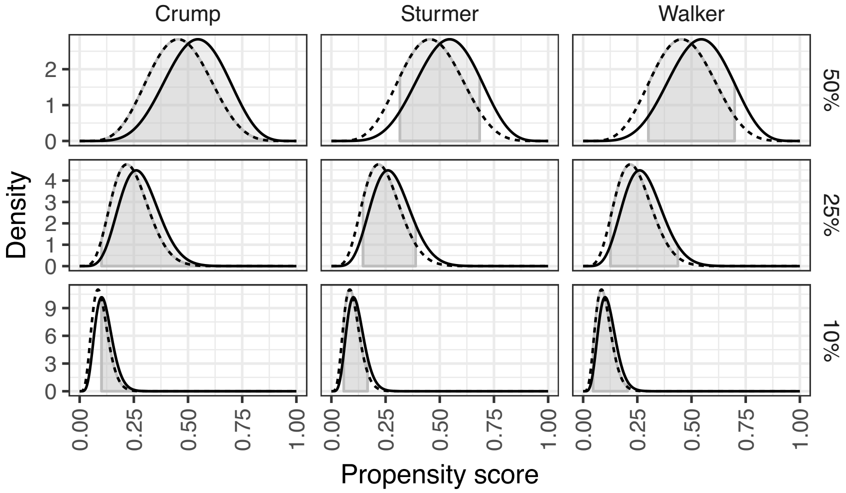
If we rewrite the trimming definition in terms of PS, we obtain the following.

$$
\begin{aligned}
I_w &= \{i \in I : \pi_i \in [\alpha_w, 1-\alpha_w]\} \\
&= \left\{ i \in I : e_i \in \left[\frac{\frac{\alpha_w}{1-\alpha_w}\frac{p}{1-p}}{1+\frac{\alpha_w}{1-\alpha_w}\frac{p}{1-p}}, \frac{\frac{1-\alpha_w}{\alpha_w}\frac{p}{1-p}}{1+\frac{1-\alpha_w}{\alpha_w}\frac{p}{1-p}} \right] \right\} \\
&= \left\{ i \in I : e_i \in \left[\frac{\alpha_w p}{(1-\alpha_w)(1-p)+\alpha_w p}, \frac{(1-\alpha_w)p}{\alpha_w(1-p)+(1-\alpha_w)p} \right] \right\}
\end{aligned}
$$

1.4 Visual comparison of methods

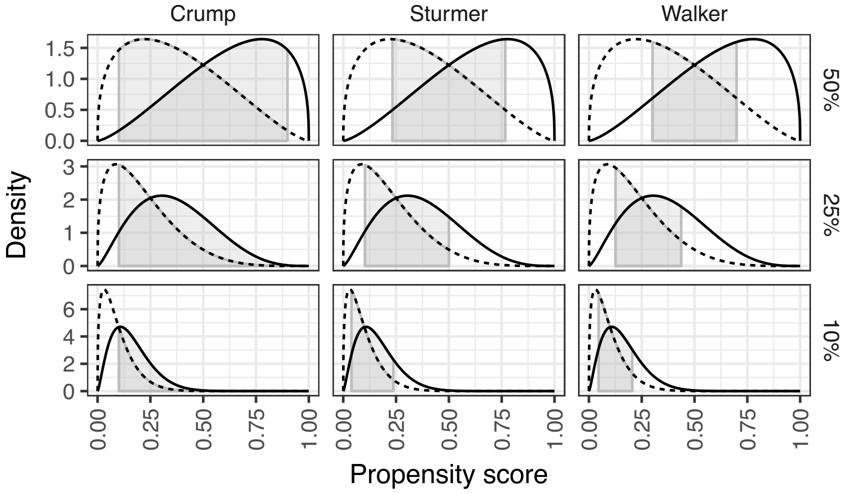
Here we provide a visual comparison of the three methods using hypothetical PS distributions. The PS distributions were generated from different beta distribution to emulate different treatment prevalence as well as covariate balance between the treated and untreated. Note in all methods, the same retention region applies to *both* treated and untreated. This uniform application of the retention region to both groups is important in avoiding artificially creating PS non-overlap regions.

1.4.1 More similar treatment groups



This example emulates a setting where covariates are *more* similar across treatment groups than the example in the main text, that is, the treatment assignment mechanism is closer to random (less confounding). In this type of setting, Walker trimming tends to be less strict (wider retention region) than Stürmer trimming.

1.4.2 Less similar treatment groups



This example emulates a setting where covariates are *less* similar across treatment groups than the example in the main text, that is, covariates affect treatment assignment more strongly (more confounding). In this type of setting, Walker trimming tends to be more strict (narrower retention region) than Stürmer trimming.

2 Extended trimming methods for the multiple-group setting

When we have multiple treatment groups ($J+1$ groups indexed with $j \in \{0, ..., J\}$), it is easier to consider all PSs, that is, all conditional probabilities of treatment assignment given the covariates.

$$\begin{aligned}
&\text{Let}\\
&A_i \in \{0, 1, ..., J\}\\
&e_{ji} = P[A_i = j|\mathbf{X}_i]\\
&\text{where} \sum_{j=0}^{J} e_{ji} = 1
\end{aligned}$$

Each individual has an individual-specific PS vector $\mathbf{e}_i = (e_{0i}, ..., e_{Ji})^T$. Using the group count-specific threshold value $\alpha_{J,c}$, $\alpha_{J,s}$, and $\alpha_{J,w}$, the proposed multinomial definitions can be written as follows.

Method	Definition
Crump	$I_{J,c} = \{i \in I : e_{ji} \geq \alpha_{J,c} \; \forall \; j \in \{0, ..., J\}\}$
Stürmer	$I_{J,s} = \left\{i \in I : e_{ji} \geq F^{-1}_{e_{ji}\|A_i}(\alpha_{J,s}\|j) \; \forall \; j \in \{0, ..., J\}\right\}$
Walker	$I_{J,w} = \{i \in I : \pi_{ji} \geq \alpha_{J,w} \; \forall \; j \in \{0, ..., J\}\}$

Notice only the lower threshold is set for each PS as opposed to the base two-group definitions. However, this is sufficient because we define the constraint for every one of the all $J+1$ PSs. As shown in the following parts, having a lower threshold for each one of the two PSs in the two-group setting is equivalent to having both upper and lower thresholds for one non-redundant PS.

2.1 Crump trimming

$$I_{J,c} = \{i \in I : e_{ji} \geq \alpha_{J,c} \; \forall \; j \in \{0, ..., J\}\}$$

This definition means that we select a subset of subjects for whom all their PSs are greater than or equal to some threshold $\alpha_{J,c}$. We can check this definition reduces to the original definition in the two group setting ($J = 1$) as follows.

$$\begin{aligned}
I_{1,c} &= \{i \in I : e_{ji} \geq \alpha_{J,c} \ \forall \ j \in \{0, 1\}\} \\
&= \{i \in I : e_{0i} \geq \alpha_{1,c}, e_{1i} \geq \alpha_{1,c}\} \\
&\quad \text{Since } e_{0i} = 1 - e_{1i} \\
&= \{i \in I : 1 - e_{1i} \geq \alpha_{1,c}, e_{1i} \geq \alpha_{1,c}\} \\
&= \{i \in I : e_{1i} \leq 1 - \alpha_{1,c}, e_{1i} \geq \alpha_{1,c}\} \\
&= \{i \in I : \alpha_{1,c} \leq e_{1i} \leq 1 - \alpha_{1,c}\} \\
&= \{i \in I : e_{1i} \in [\alpha_{1,c}, 1 - \alpha_{1,c}]\} \\
&\quad \text{Note } e_{1i} = e_i \text{ (regular two-group PS).} \\
&\quad \text{For } \alpha_{1,c} = \alpha_c \\
&= \{i \in I : e_i \in [\alpha_c, 1 - \alpha_c]\} \\
&= \text{ original two-group definition}
\end{aligned}$$

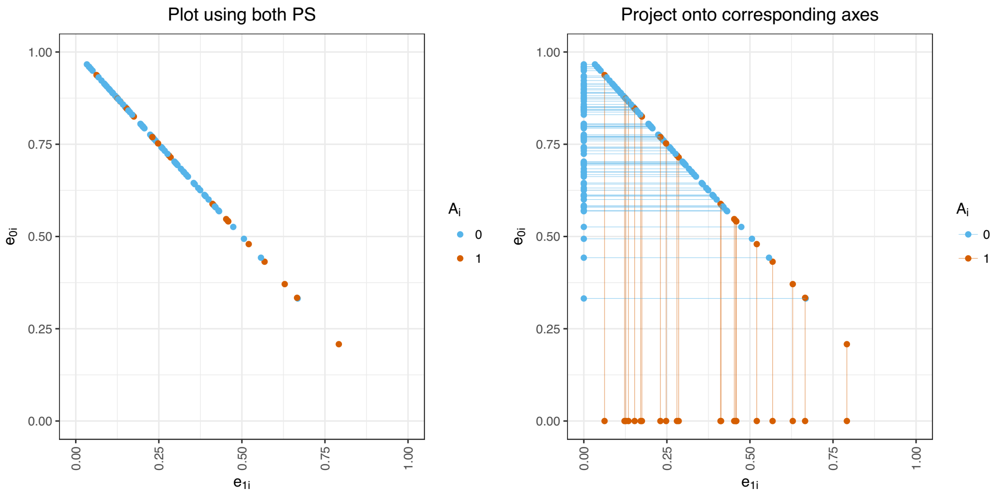
2.2 Stürmer trimming

$$I_{J,s} = \left\{ i \in I : e_{ji} \geq F^{-1}_{e_{ji}|A_i}(\alpha_{J,s}|j) \ \forall \ j \in \{0, ..., J\} \right\}$$

Note the bound is now $F^{-1}_{e_{ji}|A_i}(\alpha_{J,s}|j)$ for the corresponding multinomial PS e_{ji}. That is, for PS for treatment j (e_{ji}), the bound is determined by the lower $\alpha_{J,s}$ quantile of the PS for treatment j in the group actually received treatment j. We can check this definition reduces to the original definition in the two group setting as follows.

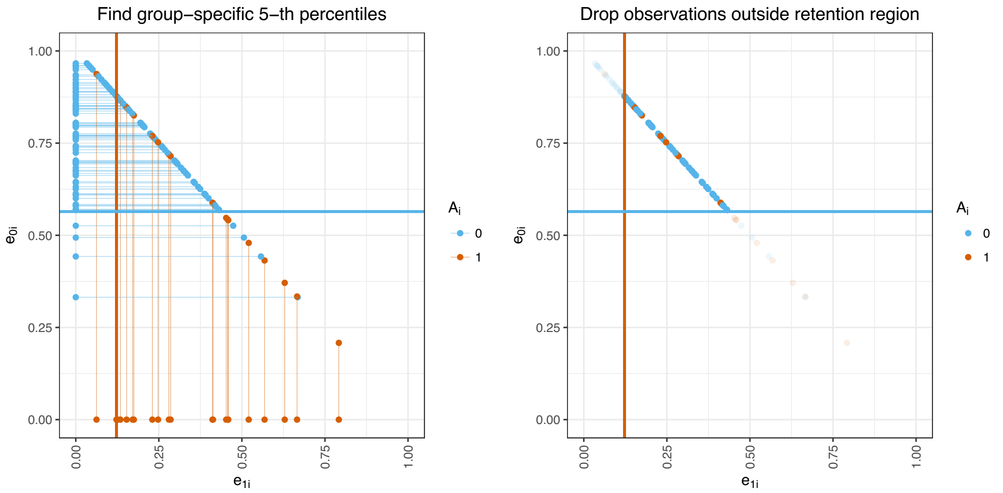
$$\begin{aligned}
I_{1,s} &= \left\{ i \in I : e_{ji} \geq F^{-1}_{e_{ji}|A_i}(\alpha_{J,s}|j) \ \forall \ j \in \{0, 1\} \right\} \\
&= \left\{ \begin{array}{l} \quad i \in I : \\ e_{0i} \geq F^{-1}_{e_{0i}|A_i}(\alpha_{1,s}|0), \\ e_{1i} \geq F^{-1}_{e_{1i}|A_i}(\alpha_{1,s}|1) \end{array} \right\} \\
&\quad \text{Since } e_{0i} = 1 - e_{1i} \\
&\quad e_{0i} \geq 100 \times \alpha_{1,s}\text{-th percentile of } e_{0i} \text{ among } A_i = 0 \\
&\quad \text{and} \\
&\quad e_{1i} \leq 100 \times (1 - \alpha_{1,s})\text{-th percentile of } e_{1i} \text{ among } A_i = 0 \\
&\quad \text{are equivalent conditions (see figures below)} \\
&= \left\{ \begin{array}{l} \quad i \in I : \\ e_{1i} \leq F^{-1}_{e_{1i}|A_i}(1 - \alpha_{1,s}|0), \\ e_{1i} \geq F^{-1}_{e_{1i}|A_i}(\alpha_{1,s}|1) \end{array} \right\} \\
&= \left\{ i \in I : e_{1i} \in \left[F^{-1}_{e_{1i}|A_i}(\alpha_{1,s}|1), F^{-1}_{e_{1i}|A_i}(1 - \alpha_{1,s}|0) \right] \right\} \\
&\quad \text{Note } e_{1i} = e_i \text{ (regular two-group PS).} \\
&\quad \text{For } \alpha_{1,s} = \alpha_s \\
&= \left\{ i \in I : e_i \in \left[F^{-1}_{e_i|A_i}(\alpha_s|1), F^{-1}_{e_i|A_i}(1 - \alpha_s|0) \right] \right\} \\
&= \text{ original two-group definition}
\end{aligned}$$

This reduction in the two-group case can be visually understood as follows. Plot data points in a 2-dimensional coordinate using both of the two PSs (left). By the constraint that $e_{0i} + e_{1i} = 1$, the data points lines up along the diagonal line. Project data points onto either axis depending on the treatment group, *i.e.*, $A_i = 1$ onto the e_{1i} axis and $A_i = 0$ onto the e_{0i} axis (right).


Plot using both PS

Project onto corresponding axes

Determine the fifth percentile in each treatment group (left), that is, the fifth percentile of e_{1i} for $A_i = 1$ and the fifth percentile of e_{0i} for $A_i = 0$. On the original diagonal line, only keep observations inside the two fifth percentile thresholds.


Find group−specific 5−th percentiles

Drop observations outside retention region

If we consider the 1-dimensional representation using e_{1i} only, what we have done is identical to the asymmetric PS trimming, *i.e.*, trim both treated and untreated observations below the fifth percentile of e_{1i} among $A_i = 1$ (treated) and *above* the ninety-fifth percentile of e_{1i} among $A_i = 0$ (untreated). The latter condition is equivalent to dropping both treated and untreated observations *below* the fifth percentile of e_{0i} among $A_i = 0$ (untreated) because of the relationship $e_{1i} = 1 - e_{0i}$.

2.3 Walker trimming

Using the multinomial preference scores, the definition is written as follows.

$$I_{J,w} = \{i \in I : \pi_{ji} \geq \alpha_{J,w} \ \forall \ j \in \{0, ..., J\}\}$$

Each multinomial preference score is defined as follows.

$$\pi_{ji} = \frac{\frac{e_{ji}}{p_j}}{\sum_{k=0}^{J} \frac{e_{ki}}{p_k}}$$

This proposed definition came from the following proposed generalization of the defining equations (J simultaneous equations) using the baseline logit multinomial logistic regression in place of the binary logistic regression in the two-group definition.

$$\begin{gathered}
\text{For } j \in \{1, ..., J\} \\
\log\left(\frac{\pi_{ji}}{\pi_{0i}}\right) = \log\left(\frac{e_{ji}}{e_{0i}}\right) - \log\left(\frac{p_j}{p_0}\right) \\
\text{where} \\
\sum_{k=0}^{J} \pi_{ki} = 1
\end{gathered}$$

The sum constraint is necessary to maintain the interpretation as the prevalence-adjusted PS. For each $j \in \{1, ..., J\}$, we have the following.

$$\begin{aligned}
\log\left(\frac{\pi_{ji}}{\pi_{0i}}\right) &= \log\left(\frac{e_{ji}}{e_{0i}}\right) - \log\left(\frac{p_j}{p_0}\right) \\
&= \log\left(\frac{e_{ji}}{e_{0i}} \Big/ \frac{p_j}{p_0}\right) \\
\frac{\pi_{ji}}{\pi_{0i}} &= \frac{e_{ji}}{e_{0i}} \Big/ \frac{p_j}{p_0} \\
&= \frac{e_{ji}}{p_j} \frac{p_0}{e_{0i}}
\end{aligned}$$

First solve for π_{0i}.

$$\begin{gathered}
\text{Sum } J \text{ equations} \\
\sum_{j=1}^{J} \frac{\pi_{ji}}{\pi_{0i}} = \sum_{j=1}^{J} \frac{e_{ji}}{p_j} \frac{p_0}{e_{0i}} \\
\frac{\sum_{j=1}^{J} \pi_{ji}}{\pi_{0i}} = \sum_{j=1}^{J} \frac{e_{ji}}{p_j} \frac{p_0}{e_{0i}} \\
\text{By } \sum_{j=0}^{J} \pi_{ji} = 1 \\
\frac{1 - \pi_{0i}}{\pi_{0i}} = \sum_{j=1}^{J} \frac{e_{ji}}{p_j} \frac{p_0}{e_{0i}} \\
\frac{\pi_{0i}}{1 - \pi_{0i}} = \frac{1}{\sum_{j=1}^{J} \frac{e_{ji}}{p_j} \frac{p_0}{e_{0i}}}
\end{gathered}$$

$$\begin{aligned}
\pi_{0i} &= \frac{\frac{1}{\sum_{j=1}^{J} \frac{e_{ji}}{p_j} \frac{p_0}{e_{0i}}}}{1 + \frac{1}{\sum_{j=1}^{J} \frac{e_{ji}}{p_j} \frac{p_0}{e_{0i}}}} \\
&= \frac{1}{1 + \sum_{j=1}^{J} \frac{e_{ji}}{p_j} \frac{p_0}{e_{0i}}} \\
&= \frac{\frac{e_{0i}}{p_0}}{\frac{e_{0i}}{p_0} + \sum_{j=1}^{J} \frac{e_{ji}}{p_j}} \\
&= \frac{\frac{e_{0i}}{p_0}}{\sum_{j=0}^{J} \frac{e_{ji}}{p_j}}
\end{aligned}$$

Now solve for an arbitrary $j \in \{1, ..., J\}$.

$$\begin{aligned}
\frac{\pi_{ji}}{\pi_{0i}} &= \frac{e_{ji}}{p_j} \frac{p_0}{e_{0i}} \\
\pi_{ji} &= \pi_{0i} \frac{e_{ji}}{p_j} \frac{p_0}{e_{0i}} \\
&= \pi_{0i} \frac{e_{ji}}{p_j} \frac{p_0}{e_{0i}} \\
&\text{Substitute } \pi_{0i} \\
&= \frac{\frac{e_{0i}}{p_0}}{\sum_{k=0}^{J} \frac{e_{ki}}{p_k}} \frac{e_{ji}}{p_j} \frac{p_0}{e_{0i}} \\
&= \frac{1}{\sum_{k=0}^{J} \frac{e_{ki}}{p_k}} \frac{e_{ji}}{p_j} \\
&= \frac{\frac{e_{ji}}{p_j}}{\sum_{k=0}^{J} \frac{e_{ki}}{p_k}}
\end{aligned}$$

Taken together, for $j \in \{0, 1, ..., J\}$,

$$\pi_{ji} = \frac{\frac{e_{ji}}{p_j}}{\sum_{k=0}^{J} \frac{e_{ki}}{p_k}}$$

We can check this definition reduces to the original definition in the two group setting as follows.

$$\text{Preference score is recovered as follows.}$$
$$\log\left(\frac{\pi_{1i}}{\pi_{0i}}\right) = \log\left(\frac{e_{1i}}{e_{0i}}\right) - \log\left(\frac{p_1}{p_0}\right)$$

$$\begin{aligned}\log\left(\frac{\pi_{1i}}{1-\pi_{1i}}\right) &= \log\left(\frac{e_{1i}}{1-e_{1i}}\right) - \log\left(\frac{p_1}{1-p_1}\right)\\ \log\left(\frac{\pi_{i}}{1-\pi_{i}}\right) &= \log\left(\frac{e_{i}}{1-e_{i}}\right) - \log\left(\frac{p}{1-p}\right)\end{aligned}$$

$$\begin{aligned}I_{1,w} &= \{i \in I : \pi_{ji} \geq \alpha_{J,w}\ \forall\ j \in \{0,1\}\}\\ &= \{i \in I : \pi_{0i} \geq \alpha_{J,w}, \pi_{1i} \geq \alpha_{J,w}\}\\ &\quad \text{Since } \pi_{0i} = 1 - \pi_{1i}\\ &= \{i \in I : 1 - \pi_{1i} \geq \alpha_{J,w}, \pi_{1i} \geq \alpha_{J,w}\}\\ &= \{i \in I : \pi_{1i} \leq 1 - \alpha_{J,w}, \pi_{1i} \geq \alpha_{J,w}\}\\ &= \{i \in I : \alpha_{J,w} \leq \pi_{1i} \leq 1 - \alpha_{J,w}\}\\ &= \{i \in I : \pi_{1i} \in [\alpha_{1,w}, 1 - \alpha_{1,w}]\}\\ &\quad \text{Note } \pi_{1i} = \pi_i \text{ (two-group preference score).}\\ &\quad \text{For } \alpha_{1,w} = \alpha_w\\ &= \{i \in I : \pi_i \in [\alpha_w, 1 - \alpha_w]\}\\ &= \text{ original two-group definition}\end{aligned}$$

2.4 Tentative threshold values

In the two group setting, the rule-of-thumb thresholds are $[0.1, 0.9]$ for Crump trimming [Crump et al., 2009], 5-th and 95-th percentiles for the Stürmer trimming [Stürmer et al., 2010], and $[0.3, 0.7]$ on the preference score scale for the Walker trimming [Walker et al., 2013]. However, using the same lower threshold value causes the multinomial trimming methods to become progressively stricter as the number of groups increases. This problem is most easily understood with Crump trimming rule. Once there are 11 groups, it is not possible to have $e_{ji} \geq 0.1$ for all PSs ($j \in \{0, ..., 10\}$) because of the constraint $\sum_{j=0}^{11} e_{ji} = 1$. Therefore, we considered the following scaling of the threshold values using the number of groups $J + 1$ for the graphical demonstration in the empirical data illustration.

Groups	J	Crump ($\alpha_{J,c}$)	Stürmer ($\alpha_{\mathrm{J,s}}$)	Walker ($\alpha_{J,w}$)
2	1	0.10	0.050	0.30
3	2	0.07	0.033	0.20
4	3	0.05	0.025	0.15
5	4	0.04	0.020	0.12
6	5	0.03	0.017	0.10
		$\vdots$		
$J + 1$	J	$\frac{1}{J+1}\frac{1}{5}$	$\frac{1}{J+1}\frac{1}{10}$	$\frac{1}{J+1}\frac{3}{5}$

Crump lower bounds are on the multinomial PS, Stürmer lower bounds are on multinomial PS quantile, and Walker lower bounds are on the multinomial preference score.

3 Empirical data illustration

3.1 Datasets

We used three characteristics datasets, each consisting of three treatment groups, to provide an intuitive understanding of the trimming methods and to illustrate how the three trimming methods differ depending on the distribution of PS among three treatment groups.

- The first example was the Medicaid non-steroidal anti-inflammatory drugs (NSAIDs) dataset [Solomon et al., 2010], the users of the three types of COX2 selective inhibitors (celecoxib, rofecoxib, and valdecoxib). The dataset was restricted to the calendar period when all of them were available (1/1/2002 - 9/30/2004).

- The second example was non-selective NSAIDs dataset derived from the same Medicaid data, and included naproxen, ibuprofen, and diclofenac as three treatment groups.

- The third dataset consisted of diabetes patients who were started on either one of sulfonylurea, glucagon-like peptide receptor agonist (GLP1-RA), or insulin in addition to metformin [Patorno et al., 2016].

3.2 Propensity score calculation and trimming

We estimated the generalized PS in each example using the baseline logit multinomial logistic regression using VGAM R package [Yee, 2010]. Three predicted probabilities were estimated for each individual. The generalized preference score was then obtained by the following equation using the generalized PS and the respective prevalence of each treatment.

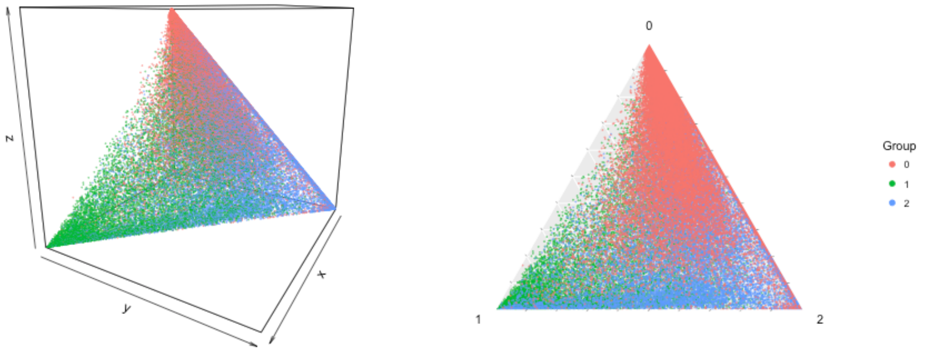
$$\widehat{\pi}_{ji} = \frac{\frac{\widehat{e}_{ji}}{\widehat{p}_j}}{\sum\limits_{k=0}^{2} \frac{\widehat{e}_{ki}}{\widehat{p}_k}} \text{ for } j \in \{0, 1, 2\}$$

Trimming was then performed at the proposed thresholds of $\alpha_{J,c} = 1/15$, $\alpha_{J,s} = 1/30$, and $\alpha_{J,w} = 1/5$. The proportion of subjects remained after trimming was recorded for the entire cohort as well as each treatment group.

3.3 Visualization with a ternary plot

The generalized PSs in the three-group setting is a vector of three elements $(e_{0i}, e_{1i}, e_{2i})^T$. As three dimensional data, individual subjects can be plotted in a three-dimensional cube $[0, 1]^3$ (left). The Z-axis represents e_{0i}, X-axis represents e_{1i}, and Y-axis represents e_{2i}. As seen in the three-dimensional plot (left), the points only occupy the diagonal triangular plane. This is because of the constraint $e_{0i} + e_{1i} + e_{2i} = 1$ for all i. In this case, we know what e_{2i} is as soon as we know e_{0i} and e_{1i}. That is, although the data are three-dimensional, the information carried is only two dimensional.

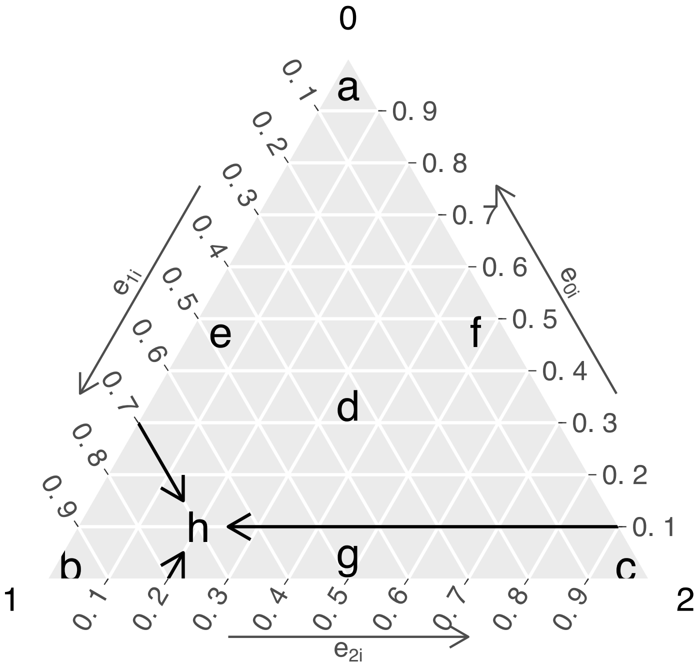
Therefore, we can take out this triangular plane in the left plot and represent as a two-dimensional plot (right). This two-dimensional representation is called a *ternary plot*. We used the ggtern R package for ternary plots [Hamilton, 2017].



The coordinate systems is explained here. The top corner of the triangle (a) is $\mathbf{e}_i = (1, 0, 0)$, *i.e.*, 100% probability of being in Group 0. The left lower corner (b) is $\mathbf{e}_i = (0, 1, 0)$ and the right lower corner (c) is $\mathbf{e}_i = (0, 0, 1)$. The mid-point in the triangle (d) is $\mathbf{e}_i = (1/3, 1/3, 1/3)$. That is, equal probability of being in any of the three groups. The mid points on the edges are: (e) $\mathbf{e}_i = (1/2, 1/2, 0)$, (f) $\mathbf{e}_i = (1/2, 0, 1/2)$, and (g) $\mathbf{e}_i = (0, 1/2, 1/2)$.

To look up point (h), all three axes have to be looked up. The e_{0i} axis is on the right edge. Use the horizontal guide lines because the labels (0.1, etc) are horizontal. Point (h) is at $e_{0i} = 0.1$. The e_{1i} axis is on the left edge. Use the guide lines going into the lower right direction as the labels indicate. Point (h) is at $e_{1i} = 0.7$. The e_{2i} axis is on the bottom edge. Use the guide lines going into the upper right direction as the labels indicate. Point (h) is at $e_{2i} = 0.2$. As a result, Point (h) is at $\mathbf{e}_i = (0.1, 0.7, 0.2)$.

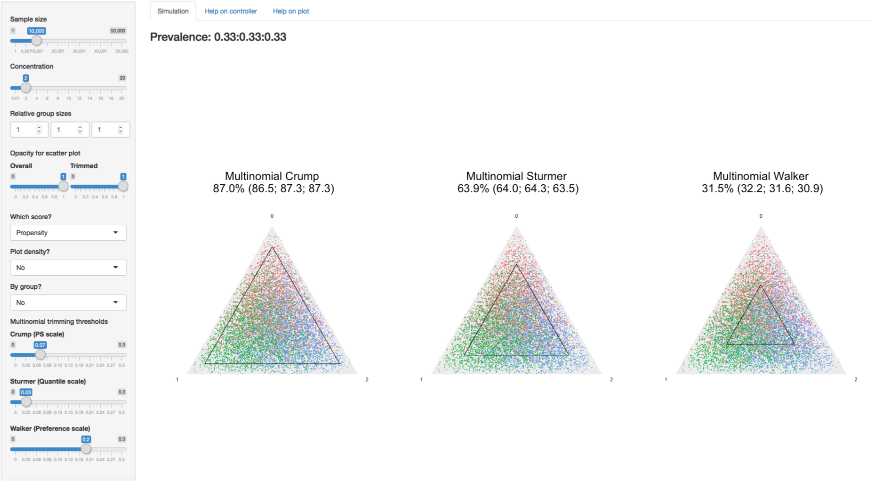
We omitted the axis labels in the empirical examples since we did not need precise value lookup. The general intuition is that being far from a given corner, for example, the top corner labeled 0, means having a low probability of being in that group.



	e_{0i}	e_{1i}	e_{2i}
(a)	1	0	0
(b)	0	1	0
(c)	0	0	1
(d)	1/3	1/3	1/3
(e)	1/2	1/2	0
(f)	1/2	0	1/2
(g)	0	1/2	1/2
(h)	0.1	0.7	0.2

3.4 Web application

The empirical datasets [Solomon et al., 2010, Patorno et al., 2016] cannot be disclosed due to data use agreement. Instead, we made an interactive web application available with a simplistic PS distribution simulation. The web application is online at `https://kaz-yos.shinyapps.io/shiny_trim_ternary/` and the source code is available at [The page will be online at publication].


Sample size: 10,000

Concentration: 2

Relative group sizes: 1 1 1

Opacity for scatter plot

Overall: 1

Trimmed: 1

Which score?: Propensity

Plot density?: No

By group?: No

Multinomial trimming thresholds

Crump (PS scale): 0.07

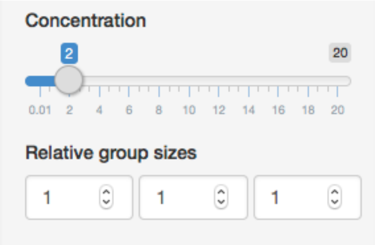
Sturmer (Quantile scale): 0.03

Walker (Preference scale): 0.2

Prevalence: 0.33:0.33:0.33

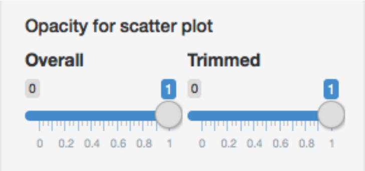
The distribution of three-group PS $\boldsymbol{e}_i = (e_{0i}, e_{1i}, e_{2i})^T$ where $e_{0i} + e_{1i} + e_{2i} = 1$ is simulated from a Dirichlet distribution Dirichlet$(c\alpha_0, c\alpha_1, c\alpha_2)$. The Dirichlet distribution is a multivariate generalization of the beta distribution [Gelman et al., 2013]. Treatment A_i is then chosen as 0, 1, or 2 based on $\boldsymbol{e}_i$, which is the treatment assignment probability.

The marginal mean of $\boldsymbol{e}_i$ generated this way is $\left(\frac{\alpha_0}{\sum_{j=0}^{2} \alpha_j}, \frac{\alpha_1}{\sum_{j=0}^{2} \alpha_j}, \frac{\alpha_2}{\sum_{j=0}^{2} \alpha_j}\right)^T$. This corresponds to the marginal prevalence of three treatment groups as already explained. In the web application, **Concentration** controls the multiplication factor c, whereas **Relative group sizes** decide α_0, α_1, and α_2. The default value for all α's is 1. The default value for c is 2. A smaller c value gives a more separated distribution of the PS and treatment groups (poor PS overlap). A larger c value gives a more concentrated distribution of the PS and treatment groups (good PS overlap). The observed prevalence of three treatment groups is shown above the plots.


Concentration: 2

Relative group sizes: 1 1 1

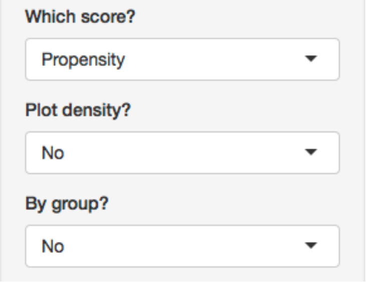
Opacity Overall controls the fading of the points in each scatter plot. **Opacity Trimmed** can be used to further fade points outside the trimming region.


Opacity for scatter plot

Overall: 1

Trimmed: 1

Choosing **Preference** in **Which score?** changes the scale to the preference score. Choosing Yes in **Plot density** results in a contour plot instead of a scatter plot. Selecting Yes in **By group** separates groups into three panels.



Which score?

Propensity

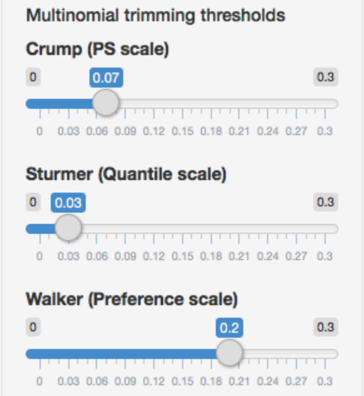
Plot density?

No

By group?

No

Multinomial trimming thresholds controls the trimming threshold for each method. The default values are the tentative values for the three-group setting stated above. The trimming boundaries are shown visually for each method. The proportion of data points retained (overall and by group) are displayed above each plot.



Multinomial trimming thresholds

Crump (PS scale)

0 0.07 0.3

Sturmer (Quantile scale)

0 0.03 0.3

Walker (Preference scale)

0 0.2 0.3

The following settings can give approximations of the empirical PS distributions by the Dirichlet distribution. They were derived from the MLE assuming a Dirichlet distribution and manual adjustment for visual similarity.

Data	c	α_0	α_1	α_2
Coxibs	5	4	3	3
nsNSAIDs	5	4.5	4.5	1
Anti-diabetics	2	0.8	0.1	0.1

4 Simulation design

The description follows the reporting recommendation in [Morris et al., 2017].

4.1 Aim

The aim of this simulation study was to assess whether the extended definitions of the PS trimming methods reduce bias due to unmeasured confounders.

4.2 Data generating mechanisms

We extended the data generating mechanism in [Stürmer et al., 2010], which they used to induce unmeasured confounders in the tails of distribution, considering three treatment groups. In the two-group setting, their data generation mechanism produces data like the following. An unmeasured binary confounder X_7 is present in the lower tail, particularly those who were actually treated. The other unmeasured binary confounder X_8 is present in the upper tail, particularly those who were left untreated.

4.2.1 Outline

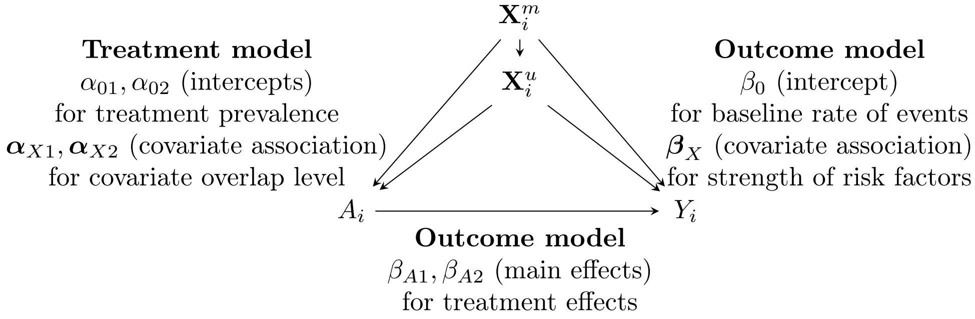
Let $i \in 1, ..., n$ index individuals.

Measured covariates

$$\mathbf{X}_i^m = \begin{bmatrix} X_{1i} & X_{2i} & X_{3i} & X_{4i} & X_{5i} & X_{6i} \end{bmatrix}^T$$

Unmeasured covariates

$$\mathbf{X}_i^u = \begin{bmatrix} X_{7i} & X_{8i} & X_{9i} \end{bmatrix}^T$$

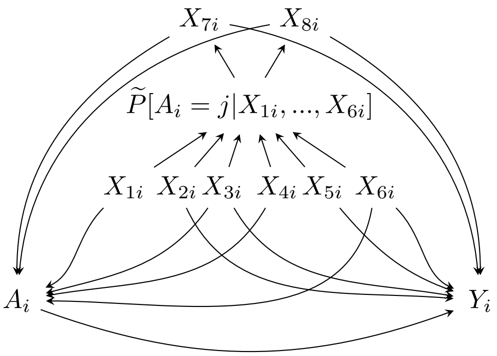

$\mathbf{X}_i^m$
Treatment model
α_{01}, α_{02} (intercepts)
for treatment prevalence
$\boldsymbol{\alpha}_{X1}, \boldsymbol{\alpha}_{X2}$ (covariate association)
for covariate overlap level
$\mathbf{X}_i^u$
Outcome model
β_0 (intercept)
for baseline rate of events
$\boldsymbol{\beta}_X$ (covariate association)
for strength of risk factors
A_i
Y_i
Outcome model
β_{A1}, β_{A2} (main effects)
for treatment effects

The following elements were varied, resulting in 3 × 3 = 9 simulation scenarios.

- **Exposure distribution**: {(33:33:33), (10:45:45), (10:10:80)}
- **Unmeasured confounding**: {none, moderate, strong}

4.2.2 Covariate generation

[Stürmer et al., 2010] used the following structure to calculate the tentative PS $\widetilde{P}[A_i = j|X_1, ..., X_6]$ based only on the base covariates $X_{1i}, ..., X_{6i}$. The tentative PS was then used to determine the probabilities of the unmeasured binary covariates X_{7i} and X_{8i}.



The base covariates $X_{1i}, ..., X_{6i}$ were generated independently using the same mechanism as [Stürmer et al., 2010].

$$X_{1i} \sim \text{Bernoulli}(0.1)$$
$$X_{2i} \sim \text{Bernoulli}(0.1)$$
$$X_{3i} \sim \text{Bernoulli}(0.1)$$
$$X_{4i} \sim \text{Normal}(0, 1)$$

$$X_{5i} \sim \text{Normal}(0,1)$$
$$X_{6i} \sim \text{Normal}(0,1)$$

Based on these measured base variables $\mathbf{X}_i^m$, the tentative PS vector $\tilde{\mathbf{e}}_i$ was calculated in a multinomial logistic regression model as follows.

$$\begin{cases} \widetilde{\eta}_{A1i} = \log\left(\dfrac{\widetilde{P}[A_i = 1|\mathbf{X}_i^m]}{\widetilde{P}[A_i = 0|\mathbf{X}_i^m]}\right) = \alpha_{01} + (\mathbf{X}_i^m)^T \boldsymbol{\alpha}_{X^m 1} \\ \widetilde{\eta}_{A2i} = \log\left(\dfrac{\widetilde{P}[A_i = 2|\mathbf{X}_i^m]}{\widetilde{P}[A_i = 0|\mathbf{X}_i^m]}\right) = \alpha_{02} + (\mathbf{X}_i^m)^T \boldsymbol{\alpha}_{X^m 2} \end{cases}$$

$$\begin{cases} \widetilde{e}_{0i} = \widetilde{P}[A_i = 0|\mathbf{X}_i^m] = \dfrac{1}{1 + \exp(\widetilde{\eta}_{A1i}) + \exp(\widetilde{\eta}_{A2i})} \\ \widetilde{e}_{1i} = \widetilde{P}[A_i = 1|\mathbf{X}_i^m] = \dfrac{\exp(\widetilde{\eta}_{A1i})}{1 + \exp(\widetilde{\eta}_{A1i}) + \exp(\widetilde{\eta}_{A2i})} \\ \widetilde{e}_{2i} = \widetilde{P}[A_i = 2|\mathbf{X}_i^m] = \dfrac{\exp(\widetilde{\eta}_{A2i})}{1 + \exp(\widetilde{\eta}_{A1i}) + \exp(\widetilde{\eta}_{A2i})} \end{cases}$$

$$\tilde{\mathbf{e}}_i = \begin{bmatrix} \widetilde{e}_{0i} & \widetilde{e}_{1i} & \widetilde{e}_{2i} \end{bmatrix}^T$$

The parameter values used in this part were the following.

$$\boldsymbol{\alpha}_{X^m 1} = (\log(2.0), \log(1.0), \log(0.2), \log(1.5), \log(1.0), \log(0.5))^T$$
$$\boldsymbol{\alpha}_{X^m 2} = (-\log(2.0), -\log(1.0), -\log(0.2), -\log(1.5), -\log(1.0), -\log(0.5))^T$$

$$(\alpha_{01}, \alpha_{02}) = \begin{cases} (-0.2, -0.5) & \text{for prevalence 33:33:33} \\ (+1.25, +0.95) & \text{for prevalence 10:45:45} \\ (-0.7, +2.1) & \text{for prevalence 10:10:80} \end{cases}$$

These tentative PSs were then used as follows to define the additional binary covariates X_{7i} through X_{9i}, which were designed as rare unmeasured conditions.

$$X_{7i} := I(U_{0i} \leq [\widetilde{e}_{0i} - \delta_0])$$
$$X_{8i} := I(U_{1i} \leq [\widetilde{e}_{1i} - \delta_1])$$
$$X_{9i} := I(U_{2i} \leq [\widetilde{e}_{2i} - \delta_2])$$

U_{ji}'s were independent $U(0,1)$ variables to introduce randomness and δ_j's were manipulated to achieve the desired marginal prevalence of 1% for each unmeasured covariate. The actual chosen values are shown below.

$$(\delta_0, \delta_1, \delta_2) = \begin{cases} (0.37, 0.63, 0.70) & \text{for prevalence 33:33:33} \\ (0.11, 0.80, 0.85) & \text{for prevalence 10:45:45} \\ (0.13, 0.35, 0.92) & \text{for prevalence 10:10:80} \end{cases}$$

4.2.3 Treatment generation

Treatment A_i was assigned based on all covariates $\mathbf{X}_i$ including both measured $\mathbf{X}_i^m$ and unmeasured $\mathbf{X}_i^u$.

$$\begin{cases} \eta_{A1i} = \log\left(\dfrac{P[A_i = 1|\mathbf{X}_i]}{P[A_i = 0|\mathbf{X}_i]}\right) = \alpha_{01} + \mathbf{X}_i^T \boldsymbol{\alpha}_{X1} \\ \eta_{A2i} = \log\left(\dfrac{P[A_i = 2|\mathbf{X}_i]}{P[A_i = 0|\mathbf{X}_i]}\right) = \alpha_{02} + \mathbf{X}_i^T \boldsymbol{\alpha}_{X2} \end{cases}$$

$$\begin{cases} e_{0i} = P(A_i = 0|\mathbf{X}_i) = \dfrac{1}{1 + \exp(\eta_{A1i}) + \exp(\eta_{A2i})} \\ e_{1i} = P(A_i = 1|\mathbf{X}_i) = \dfrac{\exp(\eta_{A1i})}{1 + \exp(\eta_{A1i}) + \exp(\eta_{A2i})} \\ e_{2i} = P(A_i = 2|\mathbf{X}_i) = \dfrac{\exp(\eta_{A2i})}{1 + \exp(\eta_{A1i}) + \exp(\eta_{A2i})} \end{cases}$$

$$A_i \in \{0, 1, 2\} \sim \text{Multinomial}\left((e_{0i}, e_{1i}, e_{2i})^T, 1\right)$$

The intercept and measured covariate coefficients were the same as before. The coefficients for the additional unmeasured covariates were the following.

For prevalence 33:33:33

$$\begin{cases} \boldsymbol{\alpha}_{X^u1} = (+10, -10, +3)^T \\ \boldsymbol{\alpha}_{X^u2} = (+10, +2, -10)^T \end{cases}$$

For prevalence 10:45:45

$$\begin{cases} \boldsymbol{\alpha}_{X^u1} = (+10, -10, +2)^T \\ \boldsymbol{\alpha}_{X^u2} = (+10, +2, -10)^T \end{cases}$$

For prevalence 10:10:80

$$\begin{cases} \boldsymbol{\alpha}_{X^u1} = (+10, -10, +2)^T \\ \boldsymbol{\alpha}_{X^u2} = (+10, +2, -10)^T \end{cases}$$

- X_{7i}, which was more common with a high $\widetilde{e}_{0i}$, had positive coefficients for both linear predictors, meaning treatment assignment was strongly driven away from group 0 when $X_{7i} = 1$.
- X_{8i}, which was more common with a high $\widetilde{e}_{1i}$, had a negative coefficient for the first linear predictor, but positive for the second, meaning treatment assignment was manipulated such that group 0 was strongly preferred over 1 and group 2 was preferred over 0 in effect driving assignment away from group 1 when $X_{8i} = 1$.
- X_{9i}, which was more common with a high $\widetilde{e}_{2i}$, had a positive coefficient for the first linear predictor, but negative for the second, meaning treatment assignment was manipulated such that group 1 was preferred over 0 and group 0 was strongly preferred over 2 in effect driving assignment away from group 2 when $X_{9i} = 1$.

In more clinical term, $X_{7i} = 1$ was a contraindication for treatment 0, $X_{8i} = 1$ was a contraindication for treatment 1, and $X_{9i} = 1$ was a contraindication for treatment 2.

4.2.4 Outcome generation

The linear predictor (log rate) for the Poisson count outcome was assigned based on all covariates and treatment. The log link was used to avoid the issue of non-collapsibility of the logit link [Greenland et al., 1999].

$$\begin{aligned}\eta_{Yi} &= \beta_0 + \beta_{A1} I(A_i = 1) + \beta_{A2} I(A_i = 2)\\ &\quad + \mathbf{X}_i^T \boldsymbol{\beta}_X + I(A_i = 1)\mathbf{X}_i^T \boldsymbol{\beta}_{XA1} + I(A_i = 2)\mathbf{X}_i^T \boldsymbol{\beta}_{XA2}\\ \\ Y_i &\sim \text{Poisson}\left(\exp(\eta_{Yi})\right)\end{aligned}$$

Additionally, the following counterfactual log rates were kept for use in calculating the marginal causal effects.

$$\begin{aligned}\eta_{Y_i^0} &= \beta_0 + \mathbf{X}_i^T \boldsymbol{\beta}_X\\ \eta_{Y_i^1} &= \beta_0 + \beta_{A1} + \mathbf{X}_i^T \boldsymbol{\beta}_X + \mathbf{X}_i^T \boldsymbol{\beta}_{XA1}\\ \eta_{Y_i^2} &= \beta_0 + \beta_{A2} + \mathbf{X}_i^T \boldsymbol{\beta}_X + \mathbf{X}_i^T \boldsymbol{\beta}_{XA2}\end{aligned}$$

The outcome model parameter values were the following.

$$\beta_0 = \log(0.20) \quad \text{Baseline rate}$$

$$(\beta_{A1}, \beta_{A2}) = (\log(0.9), \log(0.6)) \quad \text{Protective main effects}$$

$$\boldsymbol{\beta}_{X^m} = (\log(1.0), \log(2.0), \log(0.2), \log(1.0), \log(1.5), \log(0.5))^T$$

$$\boldsymbol{\beta}_{X^u}^T = \begin{cases} (0,0,0) & \text{No unmeasured confounding}\\ (\log(2), \log(2), \log(2)) & \text{Moderate unmeasured confounding}\\ (\log(10), \log(10), \log(10)) & \text{Strong unmeasured confounding}\end{cases}$$

$$\begin{bmatrix}\boldsymbol{\beta}_{XA1}^T\\ \boldsymbol{\beta}_{XA2}^T\end{bmatrix} = \begin{bmatrix}0&0&0&0&0&0&0&0&0\\0&0&0&0&0&0&0&0&0\end{bmatrix} \quad \text{No effect modification}$$

4.3 Methods to evaluate

4.3.1 Trimming thresholds

The following thresholds were used for each three-group trimming methods to examine the influence of progressively stricter trimming.

Trimming Method	Scale	Thresholds
Crump	Propensity score	{0, 1/60, 1/30, 1/15, 0.10, 0.15, 0.20, 0.30}
Stürmer	Quantile	{0, 1/60, 1/30, 0.05, 0.10, 0.15, 0.20, 0.30}
Walker	Preference score	{0, 1/40, 0.05, 0.10, 0.15, 0.20, 0.30}

4.3.2 Confounding adjustment methods

We used three PS weighting methods as confounding adjustment methods: inverse probability of treatment weights (IPTW) [Robins et al., 2000], matching weights (MW) [Li and Greene, 2013, Yoshida et al., 2017], and overlap weights (OW) [Li et al., 2016, Li et al., 2018, Li and Li, 2018]. The definitions were as follows.

$$
\begin{aligned}
IPTW_i &= \frac{1}{\sum_{j=0}^{2} I(A_i = j) e_{ji}} \\
MW_i &= \frac{\min(e_{0i}, e_{1i}, e_{2i})}{\sum_{j=0}^{2} I(A_i = j) e_{ji}} \\
OW_i &= \frac{\frac{1}{\frac{1}{e_{0i}} + \frac{1}{e_{1i}} + \frac{1}{e_{2i}}}}{\sum_{j=0}^{2} I(A_i = j) e_{ji}}
\end{aligned}
$$

4.4 Estimand

The following outcome model was fit using the glm function with the poisson family and the trimmed and weighted data. The variance estimate was obtained using the sandwich function in the sandwich package. The third contrast (group 2 vs group 1) was calculated as $\widehat{\theta}_{A2} - \widehat{\theta}_{A1}$ and its variance estimate was calculated from the variance covariance matrix accordingly, taking into consideration the covariance.

$$\log(E[Y_i|A_i]) = \theta_0 + \theta_{A1} I(A_i = 1) + \theta_{A2} I(A_i = 2)$$

The estimands (true θ's) were the marginal causal log rate ratio in the respective trimmed and weighted cohorts. These true effects can be calculated from the true coefficients (conditional effects) in the data generation mechanism in the settings without treatment effect modification by other covariates by the virtue of collapsible log link [Greenland et al., 1999]. That is, $\theta_{A1} = \beta_{A1}$ and $\theta_{A2} = \beta_{A2}$.

The simulation framework was designed to be more general as follows. In settings with treatment effect modification, the true effects depended on the covariate distribution in the trimmed and weighted cohort. We utilized the saved counterfactual log rates for each individual (below) in calculating the causal effects.

$$
\begin{aligned}
\eta_{Y_i^0} &= \beta_0 + \mathbf{X}_i^T \boldsymbol{\beta}_X \\
\eta_{Y_i^1} &= \beta_0 + \beta_{A1} + \mathbf{X}_i^T \boldsymbol{\beta}_X + \mathbf{X}_i^T \boldsymbol{\beta}_{XA1} \\
\eta_{Y_i^2} &= \beta_0 + \beta_{A2} + \mathbf{X}_i^T \boldsymbol{\beta}_X + \mathbf{X}_i^T \boldsymbol{\beta}_{XA2}
\end{aligned}
$$

Each remaining individual in the trimmed cohort was cloned three times to represent counterfactuals under three treatments. The treatment variable A_i was forced to be 0, 1, and 2 for the three clones. The outcome variable Y_i was set to be the corresponding counterfactual mean count. For example, $\exp(\eta_{Y_i^0})$ for the clone with $A_i = 0$. The same model fitting procedure was conducted using this augmented dataset containing three counterfactual clones for each original individual to calculate the true effect in the dataset. The calculated log rate ratios were average over simulation iterations.

We focused on the marginal estimands rather than conditional estimands that condition PSs because the latter require explicit modeling of the PS-outcome functional form and PS-treatment interactions. Both of these can become complicated with $J + 1$ PSs, of which J linearly independent PSs must be incorporated.

4.5 Performance measures

The trimmed sample size, bias, simulation standard error (SE), and mean squared errors (MSE) were examined. The bias, SE, and MSE were defined as follows for a true log rate ratio θ and the corresponding estimate $\widehat{\theta}_r$ (r indexing a simulation iteration $1, ..., R$).

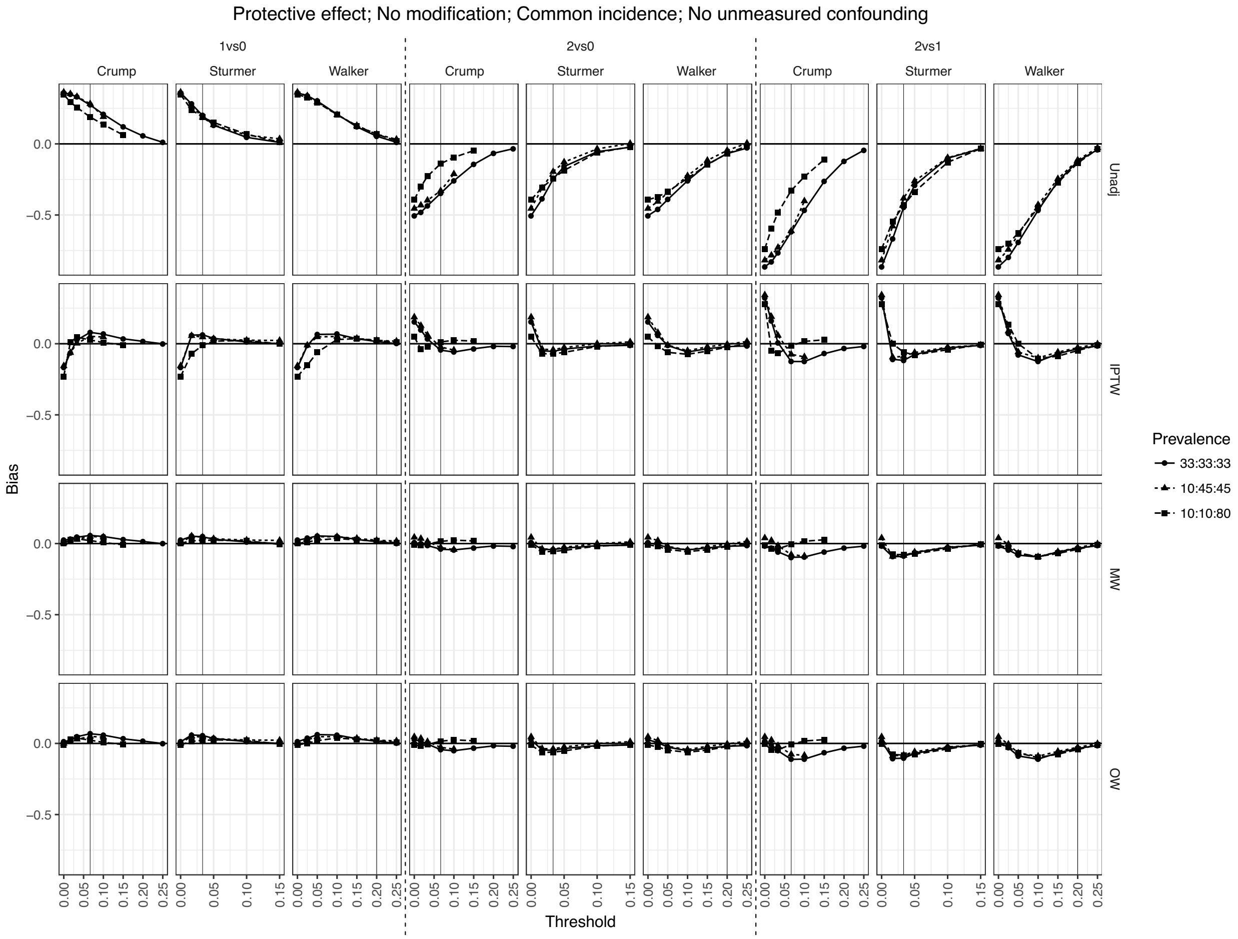
$$
\begin{aligned}
\text{Bias} &= \left(\frac{1}{R}\sum_{r=1}^{R}\widehat{\theta}_r\right) - \theta \\
\text{SE} &= \sqrt{\frac{1}{R-1}\sum_{r=1}^{R}\left(\widehat{\theta}_r - \left(\frac{1}{R}\sum_{r=1}^{R}\widehat{\theta}_r\right)\right)^2} \\
\text{MSE} &= \text{SE}^2 + \text{Bias}^2
\end{aligned}
$$

Bias of the estimators with respect to increasing trimming thresholds was the metric of most interest. Bias was calculated as the the average deviation of the estimate from the truth on the log rate ratio scale. The simulation SE was the variability (standard deviation) of estimates around their mean, whereas the MSE was the variability around the truth. MSE was used to examine the bias-variance trade off of increasing levels of trimming.

5 Additional simulation results

5.1 Bias in log rate ratio estimates

5.1.1 No unmeasured confounding

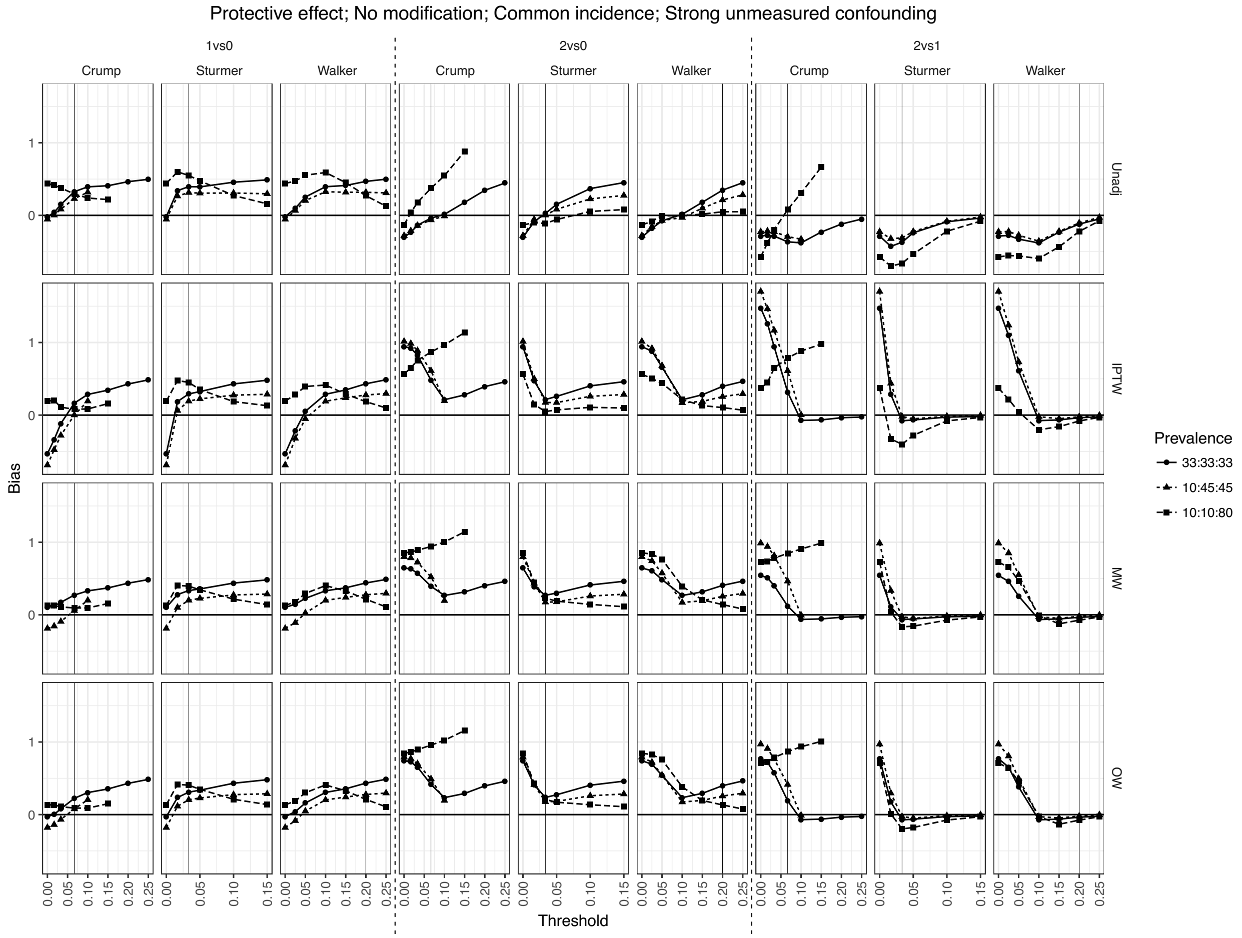


Panel layout: The rows of panels represent confounding adjustment methods: unadjusted, IPTW, MW, and OW. The columns of panels represent the group contrast and then trimming methods. Within each panel, the X axis represents progressive increase in trimming threshold (more observations are trimmed off). The vertical hairlines are at the tentative thresholds used for the empirical data illustration (Figure 1).

Abbreviations: 1vs0: group 1 vs group 0 treatment contrast; 2vs0: group 2 vs group 0 treatment contrast; 2vs1: group 2 vs group 1 treatment contrast; IPTW: inverse probability of treatment weights; MW: matching weights; OW: overlap weights.

In this case without unmeasured confounding by $X_7,\ldots,X_9$, there was a minor increase in bias with trimming after initial decrease although it decreased again with further trimming.

5.1.2 Strong unmeasured confounding



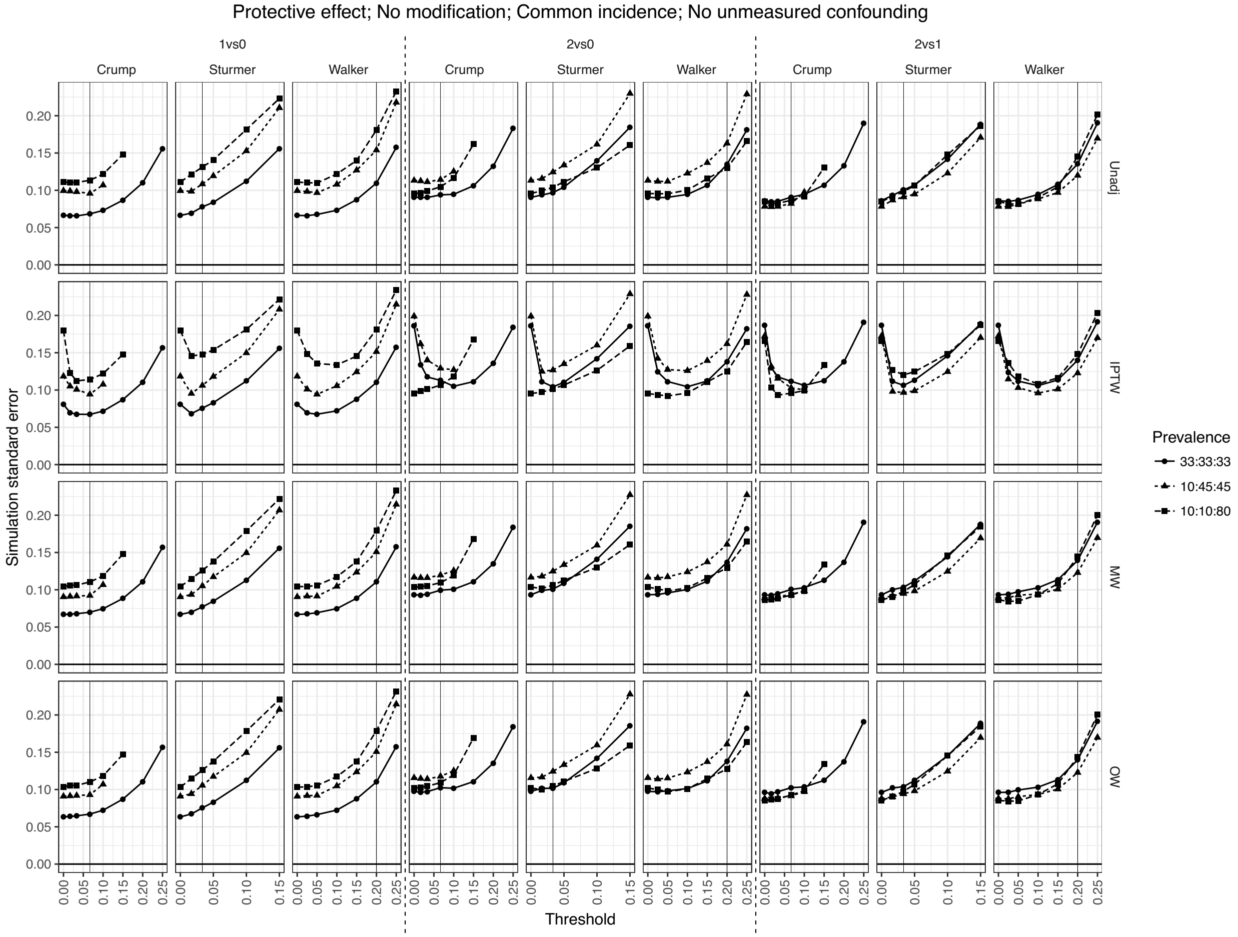
Panel layout: The rows of panels represent confounding adjustment methods: unadjusted, IPTW, MW, and OW. The columns of panels represent the group contrast and then trimming methods. Within each panel, the X axis represents progressive increase in trimming threshold (more observations are trimmed off). The vertical hairlines are at the tentative thresholds used for the empirical data illustration (Figure 1).

Abbreviations: 1vs0: group 1 vs group 0 treatment contrast; 2vs0: group 2 vs group 0 treatment contrast; 2vs1: group 2 vs group 1 treatment contrast; IPTW: inverse probability of treatment weights; MW: matching weights; OW: overlap weights.

In this case with strong unmeasured confounding by $X_7,\ldots,X_9$, the bias reduction with trimming was more apparent with contrasts 2vs0 and 2vs1, which were more biased to begin with. As observed in the moderate unmeasured confounding case, Crump trimming increased bias in the 10:10:80 treatment prevalence.

5.2 Variance of log rate ratio estimates

5.2.1 No unmeasured confounding

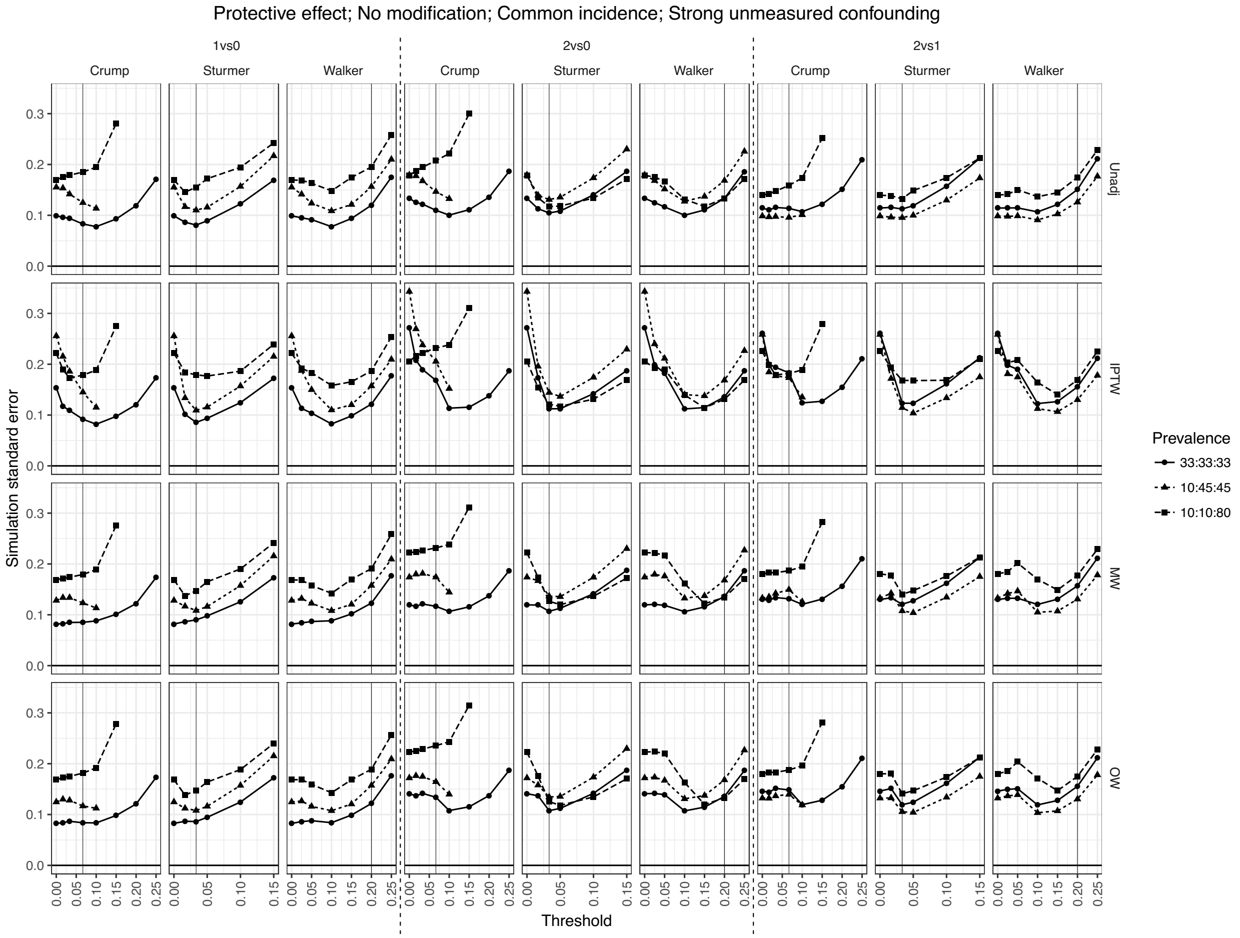


Panel layout: The rows of panels represent confounding adjustment methods: unadjusted, IPTW, MW, and OW. The columns of panels represent the group contrast and then trimming methods. Within each panel, the X axis represents progressive increase in trimming threshold (more observations are trimmed off). The vertical hairlines are at the tentative thresholds used for the empirical data illustration (Figure 1).

Abbreviations: 1vs0: group 1 vs group 0 treatment contrast; 2vs0: group 2 vs group 0 treatment contrast; 2vs1: group 2 vs group 1 treatment contrast; IPTW: inverse probability of treatment weights; MW: matching weights; OW: overlap weights.

Prominent convex patterns were seen in IPTW estimators, indicating that efficiency gain in IPTW was present even in the absence of unmeasured confounding. Much smaller initial decreases in SEs were seen in unadjusted estimators with Crump and Walker trimming. The unadjusted estimators were unweighted, thus, they did not suffer the variance inflation by huge weights in the tails of PSs. Therefore, the very minor initial reductions in unadjusted estimator SEs may be due to the bias reduction property of trimming (see the strong unmeasured confounding case).

5.2.2 Strong unmeasured confounding



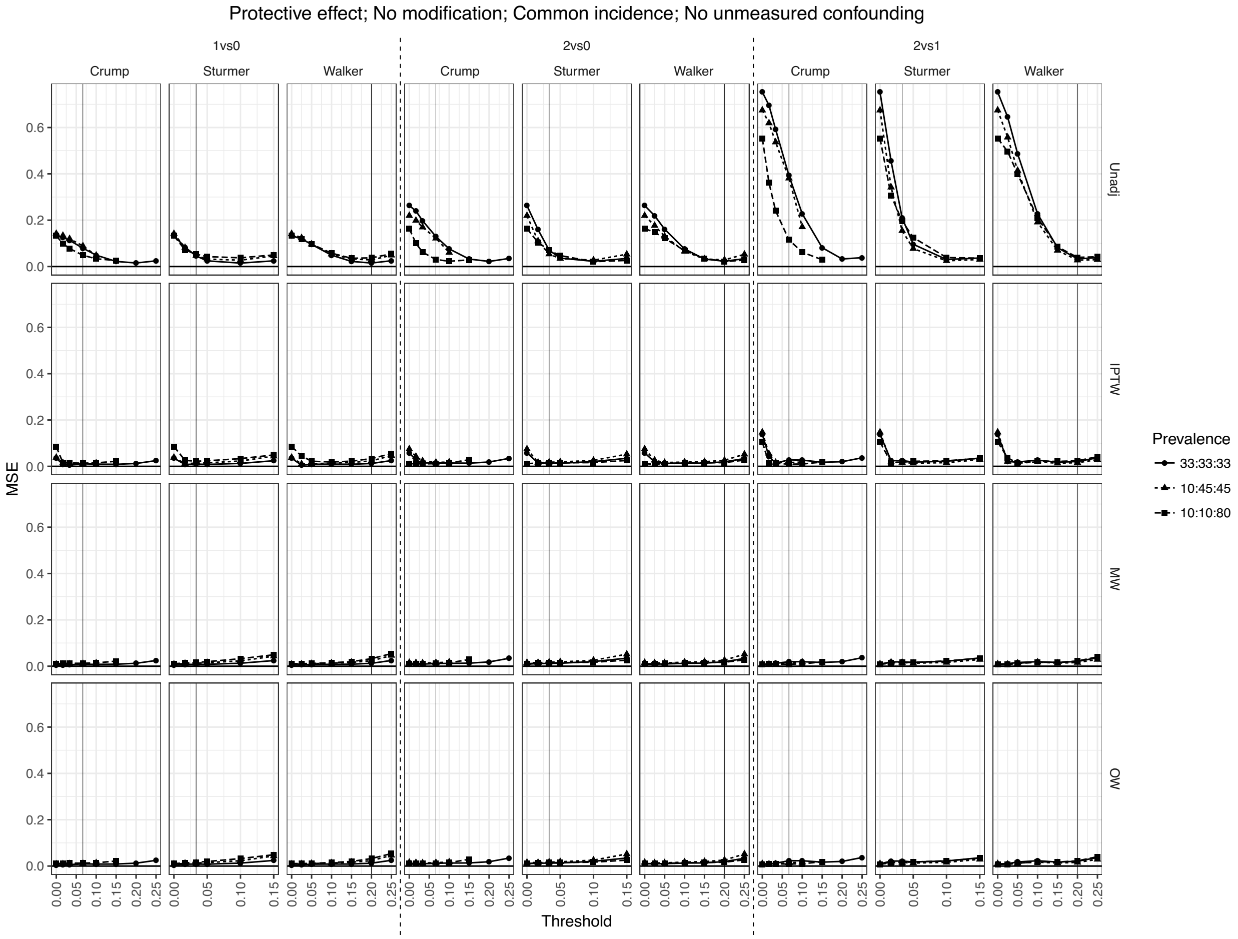
Panel layout: The rows of panels represent confounding adjustment methods: unadjusted, IPTW, MW, and OW. The columns of panels represent the group contrast and then trimming methods. Within each panel, the X axis represents progressive increase in trimming threshold (more observations are trimmed off). The vertical hairlines are at the tentative thresholds used for the empirical data illustration (Figure 1).

Abbreviations: 1vs0: group 1 vs group 0 treatment contrast; 2vs0: group 2 vs group 0 treatment contrast; 2vs1: group 2 vs group 1 treatment contrast; IPTW: inverse probability of treatment weights; MW: matching weights; OW: overlap weights.

When unmeasured confounding was strong, noticeable initial decreases in the SEs were also observed for unadjusted, MW, and OW estimators. The clearest demonstration is in the 2 vs 0 contrast with Stürmer trimming. As none of these three estimators suffer from huge weights, these findings may be explained by bias reduction. That is, when the estimates decreased in magnitude with reduced bias by the virtue of trimming, SEs also shrank (typically, small effects tend to be associated with smaller SEs).

5.3 MSE of log rate ratio estimates

5.3.1 No unmeasured confounding

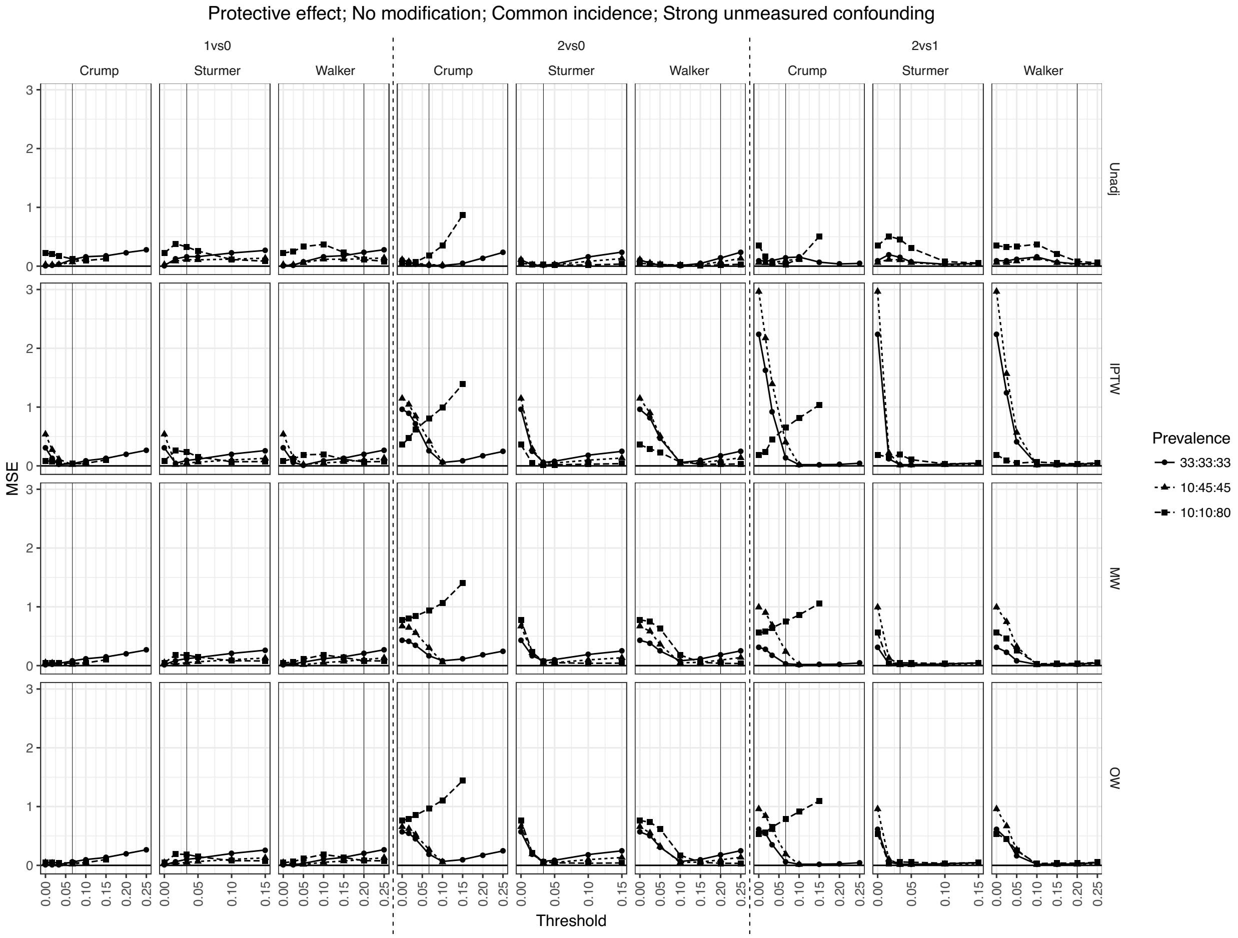


Panel layout: The rows of panels represent confounding adjustment methods: unadjusted, IPTW, MW, and OW. The columns of panels represent the group contrast and then trimming methods. Within each panel, the X axis represents progressive increase in trimming threshold (more observations are trimmed off). The vertical hairlines are at the tentative thresholds used for the empirical data illustration (Figure 1).

Abbreviations: 1vs0: group 1 vs group 0 treatment contrast; 2vs0: group 2 vs group 0 treatment contrast; 2vs1: group 2 vs group 1 treatment contrast; IPTW: inverse probability of treatment weights; MW: matching weights; OW: overlap weights.

The MSE reduction was observed in IPTW, but was not apparent in MW and OW in the setting without unmeasured confounding.

5.3.2 Strong unmeasured confounding



Panel layout: The rows of panels represent confounding adjustment methods: unadjusted, IPTW, MW, and OW. The columns of panels represent the group contrast and then trimming methods. Within each panel, the X axis represents progressive increase in trimming threshold (more observations are trimmed off). The vertical hairlines are at the tentative thresholds used for the empirical data illustration (Figure 1).

Abbreviations: 1vs0: group 1 vs group 0 treatment contrast; 2vs0: group 2 vs group 0 treatment contrast; 2vs1: group 2 vs group 1 treatment contrast; IPTW: inverse probability of treatment weights; MW: matching weights; OW: overlap weights.

When the unmeasured confounding was strong, the MSE was more heavily influenced by bias than variance. As a result, all of IPTW, MW, and OW demonstrated decrease in the MSE for the more biased treatment contrasts (2vs0 and 2vs1). Crump trimming increased the MSE in the 10:10:80 treatment prevalence due to increase in bias.

6 Bibliography

[Crump et al., 2009] Crump, R. K., Hotz, V. J., Imbens, G. W., and Mitnik, O. A. (2009). Dealing with limited overlap in estimation of average treatment effects. *Biometrika*, 96(1):187–199.

[Freedman, 1987] Freedman, B. (1987). Equipoise and the ethics of clinical research. *N. Engl. J. Med.*, 317(3):141–145.

[Gelman et al., 2013] Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian Data Analysis, Third Edition.* CRC Press.

[Greenland et al., 1999] Greenland, S., Robins, J. M., and Pearl, J. (1999). Confounding and Collapsibility in Causal Inference. *Statistical Science*, 14(1):29–46.

[Hamilton, 2017] Hamilton, N. (2017). Ggtern: An Extension to 'ggplot2', for the Creation of Ternary Diagrams.

[Li and Li, 2018] Li, F. and Li, F. (2018). Propensity Score Weighting for Causal Inference with Multi-valued Treatments. *arXiv:1808.05339 [stat]*.

[Li et al., 2016] Li, F., Morgan, K. L., and Zaslavsky, A. M. (2016). Balancing Covariates via Propensity Score Weighting. *Journal of the American Statistical Association*, 0(0):1–11.

[Li et al., 2018] Li, F., Thomas, L. E., and Li, F. (2018). Addressing Extreme Propensity Scores via the Overlap Weights. *Am. J. Epidemiol.*

[Li and Greene, 2013] Li, L. and Greene, T. (2013). A weighting analogue to pair matching in propensity score analysis. *Int J Biostat*, 9(2):215–234.

[Morris et al., 2017] Morris, T. P., White, I. R., and Crowther, M. J. (2017). Using simulation studies to evaluate statistical methods. *arXiv:1712.03198 [stat]*.

[Patorno et al., 2016] Patorno, E., Everett, B. M., Goldfine, A. B., Glynn, R. J., Liu, J., Gopalakrishnan, C., and Kim, S. C. (2016). Comparative cardiovascular safety of glucagon-like peptide-1 receptor agonists versus other antidiabetic drugs in routine care: A cohort study. *Diabetes Obes Metab*, 18(8):755–765.

[Petersen et al., 2012] Petersen, M. L., Porter, K. E., Gruber, S., Wang, Y., and van der Laan, M. J. (2012). Diagnosing and responding to violations in the positivity assumption. *Stat Methods Med Res*, 21(1):31–54.

[Robins et al., 2000] Robins, J. M., Hernán, M. A., and Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11(5):550–560.

[Solomon et al., 2010] Solomon, D. H., Rassen, J. A., Glynn, R. J., Lee, J., Levin, R., and Schneeweiss, S. (2010). The comparative safety of analgesics in older adults with arthritis. *Arch. Intern. Med.*, 170(22):1968–1976.

[Stürmer et al., 2010] Stürmer, T., Rothman, K. J., Avorn, J., and Glynn, R. J. (2010). Treatment effects in the presence of unmeasured confounding: Dealing with observations in the tails of the propensity score distribution–a simulation study. *Am. J. Epidemiol.*, 172(7):843–854.

[Walker et al., 2013] Walker, A. M., Patrick, A. R., Lauer, M. S., Hornbrook, M. C., Marin, M. G., Platt, R., Roger, V. L., Stang, P., and Schneeweiss, S. (2013). A tool for assessing the feasibility of comparative effectiveness research. *Comp Eff Res*, page 11.

[Yee, 2010] Yee, T. W. (2010). The VGAM Package for Categorical Data Analysis. *J Stat Softw*, 29(6):1427–1445.

[Yoshida et al., 2017] Yoshida, K., Hernandez-Diaz, S., Solomon, D. H., Jackson, J. W., Gagne, J. J., Glynn, R. J., and Franklin, J. M. (2017). Matching Weights to Simultaneously Compare Three Treatment Groups: Comparison to Three-way Matching. *Epidemiology*, 28(3):387–395.