Web Material

Multinomial Extension of Propensity Score Trimming Methods: A Simulation Study

Kazuki Yoshida, Daniel H. Solomon, Sebastien Haneuse, Seoyoung C. Kim, Elisabetta Patorno, Sara K. Tedeschi, Houchen Lyu, Jessica M. Franklin, Til Stürmer, Sonia Hernandez-Diaz, and Robert J. Glynn

Web Appendix 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 [1] 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* [2] 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) [3] 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 [4], 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* [5] 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* [1] 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* [6]. 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

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

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

Also

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

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

$$= \frac{\pi_i p}{(1-\pi_i)(1-p) + \pi_i p}$$

If we rewrite the trimming definition in terms of PS, we obtain the following.

$$\begin{aligned} I_w &= \{i \in I : \pi_i \in [\alpha_w, 1-\alpha_w]\} \\ &= \left\{ i \in I : e_i \in \left[\frac{\frac{\alpha_w}{1-\alpha_w}\frac{p}{1-p}}{1 + \frac{\alpha_w}{1-\alpha_w}\frac{p}{1-p}}, \frac{\frac{1-\alpha_w}{\alpha_w}\frac{p}{1-p}}{1 + \frac{1-\alpha_w}{\alpha_w}\frac{p}{1-p}} \right] \right\} \\ &= \left\{ i \in I : e_i \in \left[\frac{\alpha_w p}{(1-\alpha_w)(1-p) + \alpha_w p}, \frac{(1-\alpha_w)p}{\alpha_w(1-p) + (1-\alpha_w)p} \right] \right\} \end{aligned}$$

Web Appendix 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.

$$
\begin{aligned}
I_{1,s} &= \left\{ i \in I : e_{ji} \geq F^{-1}_{e_{ji}|A_i}(\alpha_{J,s}|j) \ \forall \ j \in \{0,1\} \right\} \\
&= \left\{ \begin{array}{l} i \in I : \\ e_{0i} \geq F^{-1}_{e_{0i}|A_i}(\alpha_{1,s}|0), \\ e_{1i} \geq F^{-1}_{e_{1i}|A_i}(\alpha_{1,s}|1) \end{array} \right\} \\
&\quad \text{Since } e_{0i} = 1 - e_{1i} \\
&\quad e_{0i} \geq 100 \times \alpha_{1,s}\text{-th percentile of } e_{0i} \text{ among } A_i = 0 \\
&\quad \text{and} \\
&\quad e_{1i} \leq 100 \times (1 - \alpha_{1,s})\text{-th percentile of } e_{1i} \text{ among } A_i = 0 \\
&\quad \text{are equivalent conditions (see figures below)} \\
&= \left\{ \begin{array}{l} i \in I : \\ e_{1i} \leq F^{-1}_{e_{1i}|A_i}(1 - \alpha_{1,s}|0), \\ e_{1i} \geq F^{-1}_{e_{1i}|A_i}(\alpha_{1,s}|1) \end{array} \right\} \\
&= \left\{ i \in I : e_{1i} \in \left[F^{-1}_{e_{1i}|A_i}(\alpha_{1,s}|1), F^{-1}_{e_{1i}|A_i}(1 - \alpha_{1,s}|0) \right] \right\} \\
&\quad \text{Note } e_{1i} = e_i \text{ (regular two-group PS).} \\
&\quad \text{For } \alpha_{1,s} = \alpha_s \\
&= \left\{ i \in I : e_i \in \left[F^{-1}_{e_i|A_i}(\alpha_s|1), F^{-1}_{e_i|A_i}(1 - \alpha_s|0) \right] \right\} \\
&= \text{ original two-group definition}
\end{aligned}
$$

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}.

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

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

$$\begin{aligned}
\frac{\pi_{ji}}{\pi_{0i}} &= \frac{e_{ji}}{p_j}\frac{p_0}{e_{0i}} \\
\pi_{ji} &= \pi_{0i}\frac{e_{ji}}{p_j}\frac{p_0}{e_{0i}}
\end{aligned}$$

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

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.

Preference score is recovered as follows.

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

$$
\begin{aligned}
I_{1,w} &= \{i \in I : \pi_{ji} \geq \alpha_{J,w} \ \forall \ j \in \{0,1\}\} \\
&= \{i \in I : \pi_{0i} \geq \alpha_{J,w}, \pi_{1i} \geq \alpha_{J,w}\} \\
&\quad \text{Since } \pi_{0i} = 1 - \pi_{1i} \\
&= \{i \in I : 1 - \pi_{1i} \geq \alpha_{J,w}, \pi_{1i} \geq \alpha_{J,w}\} \\
&= \{i \in I : \pi_{1i} \leq 1 - \alpha_{J,w}, \pi_{1i} \geq \alpha_{J,w}\} \\
&= \{i \in I : \alpha_{J,w} \leq \pi_{1i} \leq 1 - \alpha_{J,w}\} \\
&= \{i \in I : \pi_{1i} \in [\alpha_{1,w}, 1 - \alpha_{1,w}]\} \\
&\quad \text{Note } \pi_{1i} = \pi_i \text{ (two-group preference score).} \\
&\quad \text{For } \alpha_{1,w} = \alpha_w \\
&= \{i \in I : \pi_i \in [\alpha_w, 1 - \alpha_w]\} \\
&= \text{ original two-group definition}
\end{aligned}
$$

2.4 Tentative threshold values

In the two group setting, the rule-of-thumb thresholds are $[0.1, 0.9]$ for Crump trimming [2], 5-th and 95-th percentiles for the Stürmer trimming [5], and $[0.3, 0.7]$ on the preference score scale for the Walker trimming [1]. 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 $\sum_{j=0}^{10} e_{ji} = 1$. 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	0.100	0.050	0.300
3	2	0.067	0.033	0.200
4	3	0.050	0.025	0.150
5	4	0.040	0.020	0.120
6	5	0.033	0.017	0.100
		$\vdots$		
$J+1$	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.

Web Appendix 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 [7], 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 [8].

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 [9]. 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.

$$\widehat{\pi}_{ji} = \frac{\frac{\widehat{e}_{ji}}{\widehat{p}_j}}{\sum_{k=0}^{2} \frac{\widehat{e}_{ki}}{\widehat{p}_k}} \text{ for } j \in \{0, 1, 2\}$$

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.

Web Appendix 4. Simulation design

The description follows the reporting recommendation in [10].

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 [5], 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

The following elements were varied, resulting in $3 \times 3 = 9$ simulation scenarios.

- **Exposure distribution**: {(33:33:33), (10:45:45), (10:10:80)}
- **Unmeasured confounding**: {none, moderate, strong}

4.2.2 Covariate generation

The base covariates $X_{1i}, ..., X_{6i}$ were generated independently using the same mechanism as [5].

$$
\begin{aligned}
X_{1i} &\sim \text{Bernoulli}(0.1) \\
X_{2i} &\sim \text{Bernoulli}(0.1) \\
X_{3i} &\sim \text{Bernoulli}(0.1) \\
X_{4i} &\sim \text{Normal}(0,1) \\
X_{5i} &\sim \text{Normal}(0,1) \\
X_{6i} &\sim \text{Normal}(0,1)
\end{aligned}
$$

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.

$$
\begin{aligned}
\boldsymbol{\alpha}_{X^m 1} &= (\log(2.0), \log(1.0), \log(0.2), \log(1.5), \log(1.0), \log(0.5))^T \\
\boldsymbol{\alpha}_{X^m 2} &= (-\log(2.0), -\log(1.0), -\log(0.2), -\log(1.5), -\log(1.0), -\log(0.5))^T
\end{aligned}
$$

$$
(\alpha_{01}, \alpha_{02}) = \begin{cases}
(-0.2, -0.5) & \text{for prevalence 33:33:33} \\
(+1.25, +0.95) & \text{for prevalence 10:45:45} \\
(-0.7, +2.1) & \text{for prevalence 10:10:80}
\end{cases}
$$

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.

$$
\begin{aligned}
X_{7i} &:= I(U_{0i} \leq [\widetilde{e}_{0i} - \delta_0]) \\
X_{8i} &:= I(U_{1i} \leq [\widetilde{e}_{1i} - \delta_1]) \\
X_{9i} &:= I(U_{2i} \leq [\widetilde{e}_{2i} - \delta_2])
\end{aligned}
$$

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

$$\begin{cases} \boldsymbol{\alpha}_{X^u 1} = (+10, -10, +3)^T \\ \boldsymbol{\alpha}_{X^u 2} = (+10, +2, -10)^T \end{cases}$$

For prevalence 10:45:45

$$\begin{cases} \boldsymbol{\alpha}_{X^u 1} = (+10, -10, +2)^T \\ \boldsymbol{\alpha}_{X^u 2} = (+10, +2, -10)^T \end{cases}$$

For prevalence 10:10:80

$$\begin{cases} \boldsymbol{\alpha}_{X^u 1} = (+10, -10, +2)^T \\ \boldsymbol{\alpha}_{X^u 2} = (+10, +2, -10)^T \end{cases}$$

- 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 [11].

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

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

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.

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

$$(\beta_{A1}, \beta_{A2}) = (\log(0.9), \log(0.6)) \quad \text{Protective main effects}$$

$$\boldsymbol{\beta}_{X^m} = (\log(1.0), \log(2.0), \log(0.2), \log(1.0), \log(1.5), \log(0.5))^T$$

$$\boldsymbol{\beta}_{X^u}^T = \begin{cases} (0,0,0) & \text{No unmeasured confounding} \\ (\log(2), \log(2), \log(2)) & \text{Moderate unmeasured confounding} \\ (\log(10), \log(10), \log(10)) & \text{Strong unmeasured confounding} \end{cases}$$

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

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) [3], matching weights (MW) [12, 13], and overlap weights (OW) [14, 15, 16]. The definitions were as follows.

$$IPTW_i = \frac{1}{\sum_{j=0}^{2} I(A_i = j) e_{ji}}$$

$$MW_i = \frac{\min(e_{0i}, e_{1i}, e_{2i})}{\sum_{j=0}^{2} I(A_i = j) e_{ji}}$$

$$OW_i = \frac{\frac{1}{\frac{1}{e_{0i}} + \frac{1}{e_{1i}} + \frac{1}{e_{2i}}}}{\sum_{j=0}^{2} I(A_i = j) e_{ji}}$$

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 [11]. 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 $1, ..., R$).

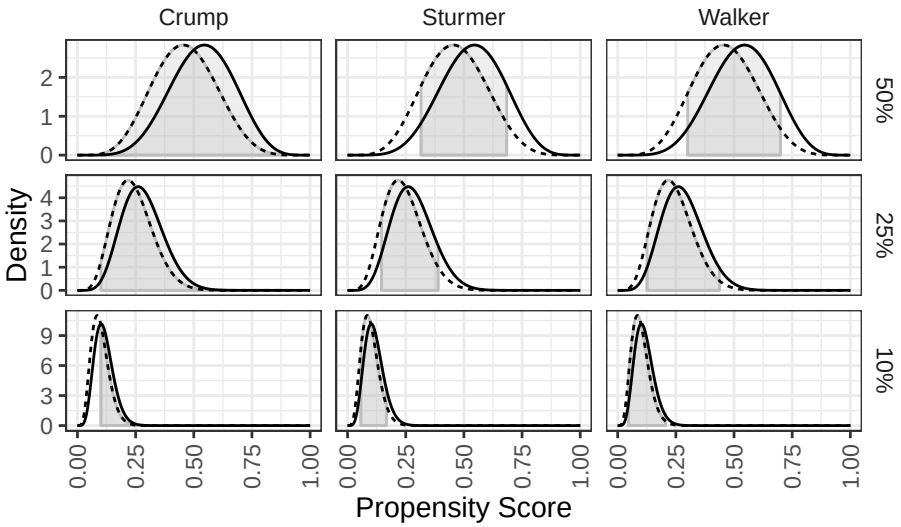
$$\text{Bias} = \left(\frac{1}{R} \sum_{r=1}^{R} \widehat{\theta}_r \right) - \theta$$

$$\text{SE} = \sqrt{\frac{1}{R-1} \sum_{r=1}^{R} \left(\widehat{\theta}_r - \left(\frac{1}{R} \sum_{r=1}^{R} \widehat{\theta}_r \right) \right)^2}$$

$$\text{MSE} = \text{SE}^2 + \text{Bias}^2$$

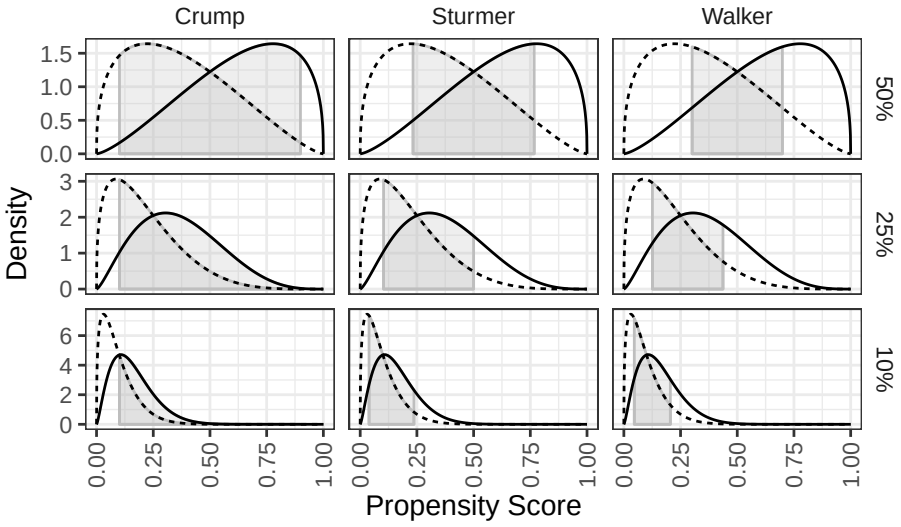
Bias of the estimators with respect to increasing trimming thresholds was the metric of most interest. Bias was calculated as the the average deviation of the estimate from the truth on the log rate ratio scale. The simulation SE was the variability (standard deviation) of estimates around their mean, whereas the MSE was the variability around the truth. MSE was used to examine the bias-variance trade off of increasing levels of trimming.

Web Figure 1. Visual comparison of methods (more similar treatment groups)
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.



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.

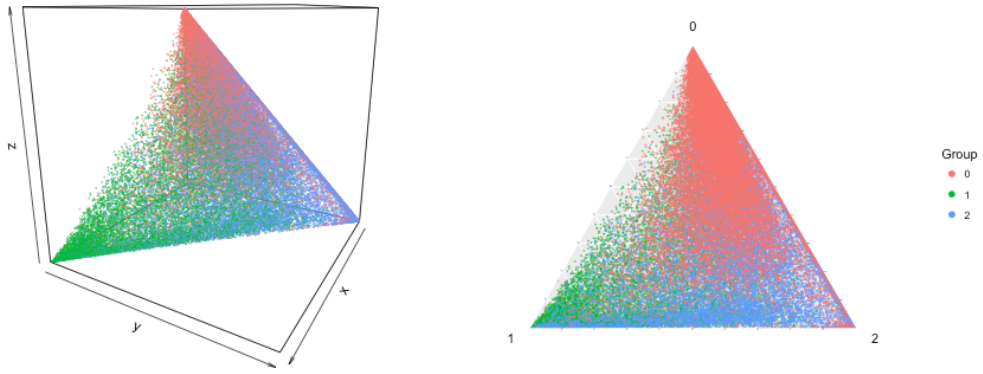
Web Figure 2. Visual comparison of methods (less similar treatment groups)



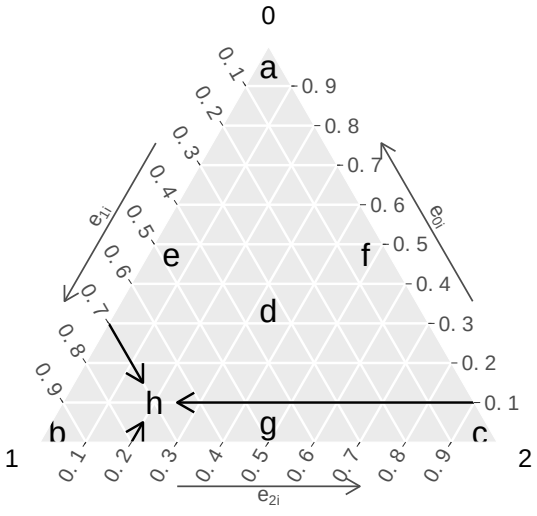
This example emulates a setting where covariates are *less* similar across treatment groups than the example in the main text, that is, covariates affect treatment assignment more strongly (more confounding). In this type of setting, Walker trimming tends to be more strict (narrower retention region) than Stürmer trimming.

Web Figure 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 [17].



Web Figure 4. Ternary plot coordinate system
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) $\mathbf{e}_i = (0, 1/2, 1/2)$.

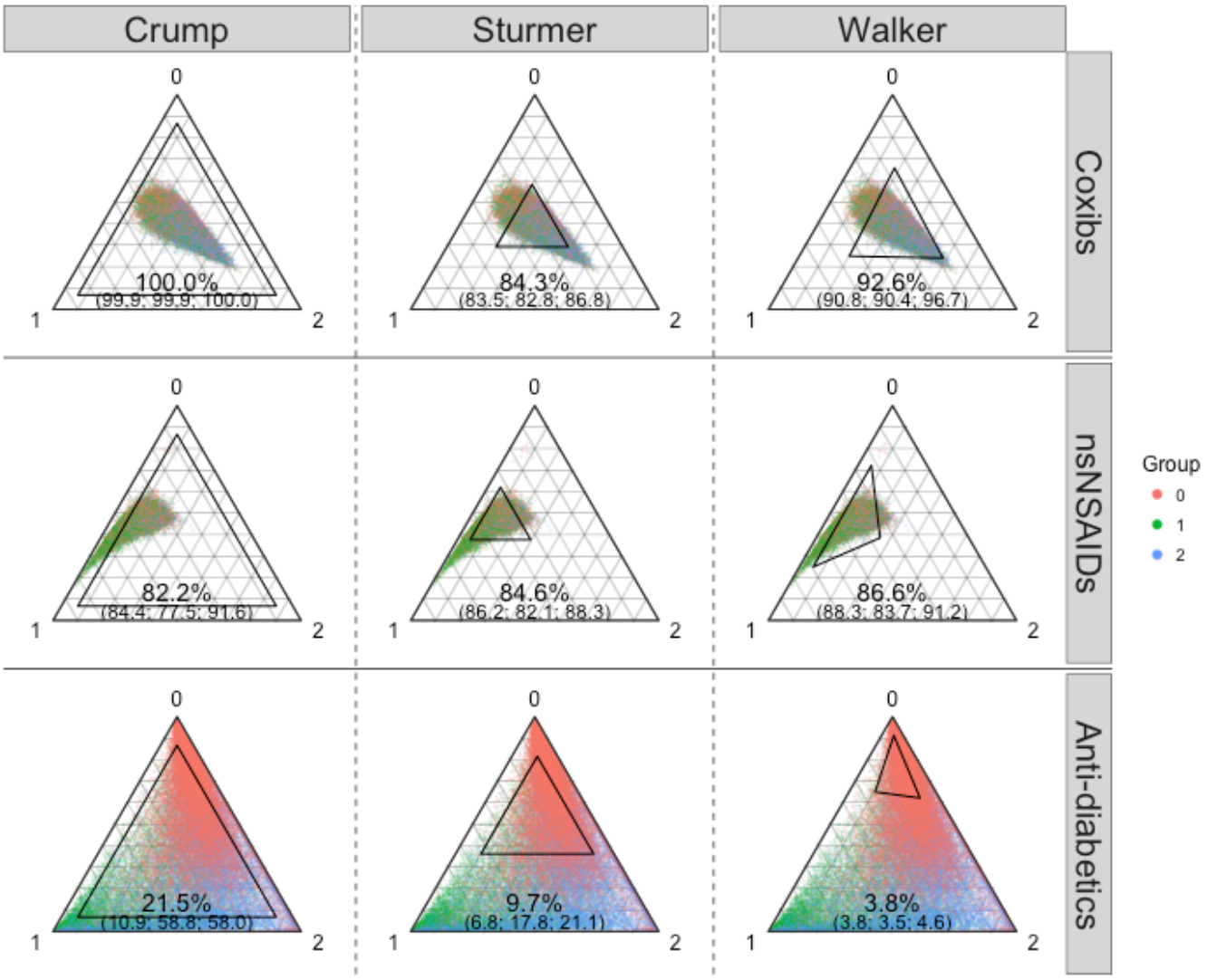


	e_{0i}	e_{1i}	e_{2i}
(a)	1	0	0
(b)	0	1	0
(c)	0	0	1
(d)	1/3	1/3	1/3
(e)	1/2	1/2	0
(f)	1/2	0	1/2
(g)	0	1/2	1/2
(h)	0.1	0.7	0.2

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.

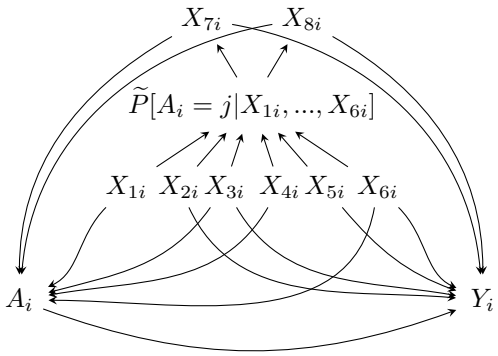
Web Figure 5. Empirical data examples



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.

Web Figure 6. Stürmer *et al*'s data generation mechanism

Stürmer *et al* [5] 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}.



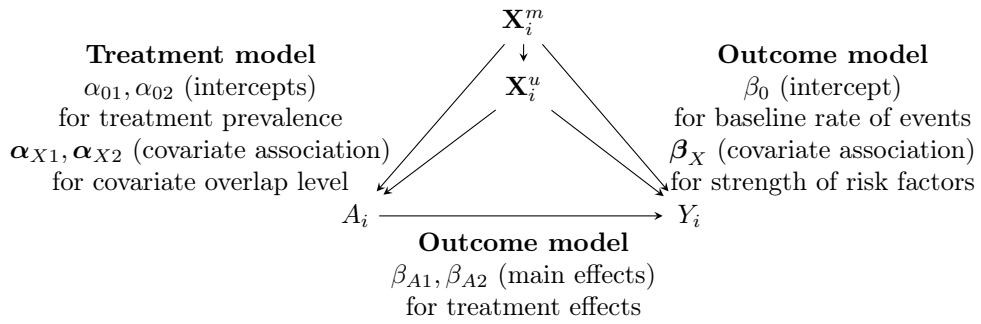
Web Figure 7. Overview of our data generation mechanism
Let $i \in 1, ..., n$ index individuals.

Measured covariates

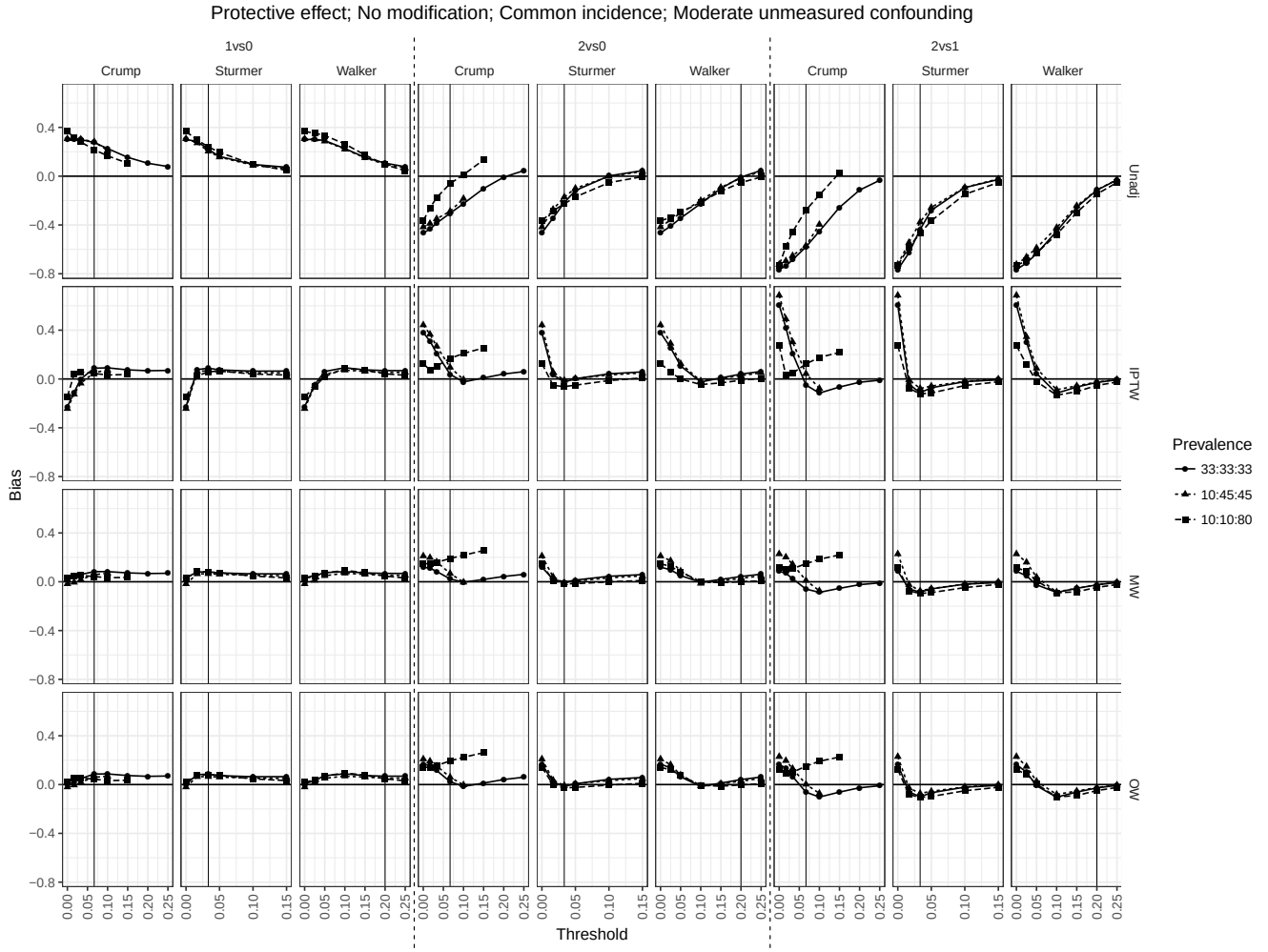
$$\mathbf{X}_i^m = \begin{bmatrix} X_{1i} & X_{2i} & X_{3i} & X_{4i} & X_{5i} & X_{6i} \end{bmatrix}^T$$

Unmeasured covariates

$$\mathbf{X}_i^u = \begin{bmatrix} X_{7i} & X_{8i} & X_{9i} \end{bmatrix}^T$$



Web Figure 8. Bias in log rate ratio estimates (moderate unmeasured confounding)

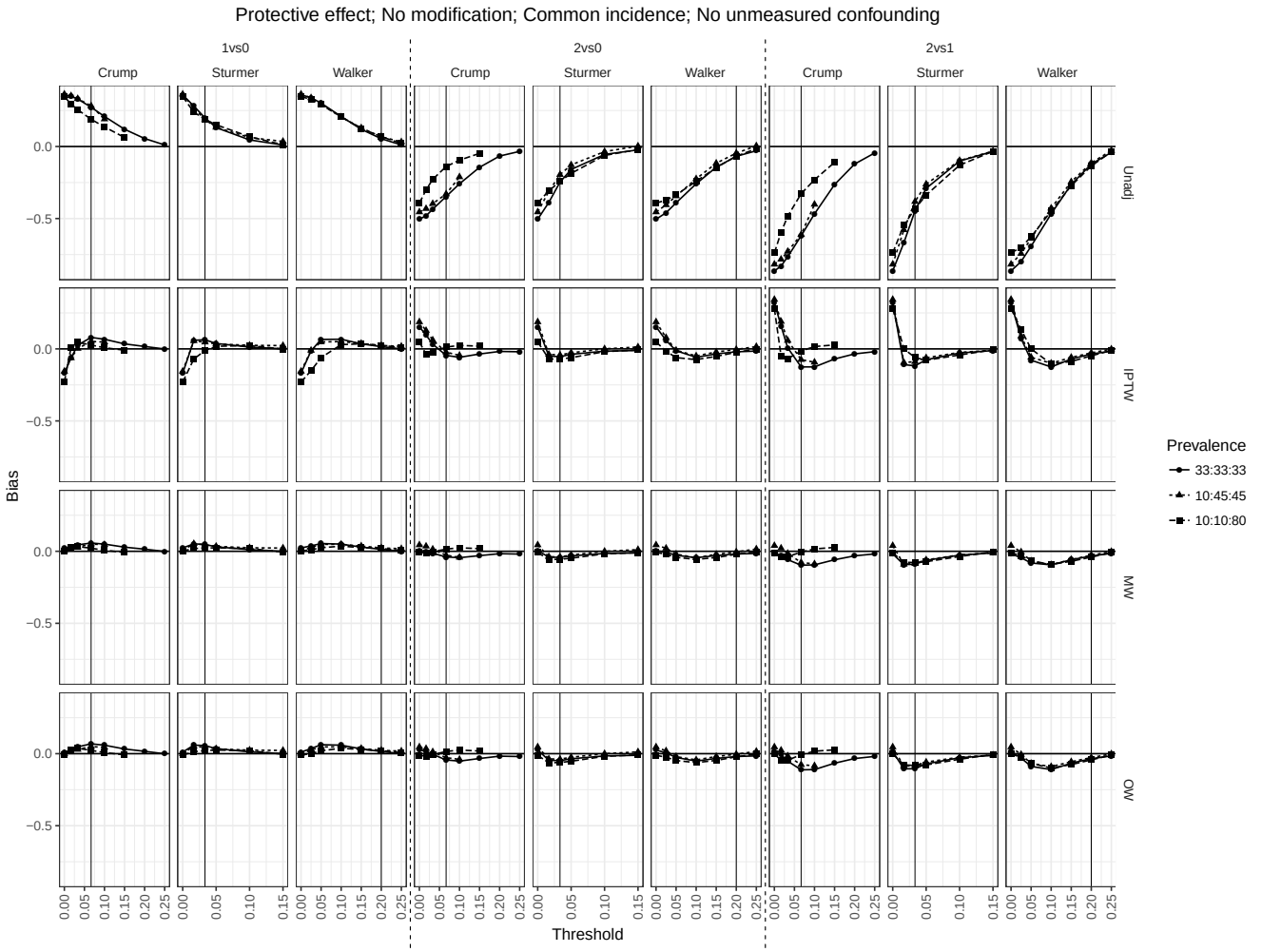


Panel layout: The rows of panels represent confounding adjustment methods: unadjusted, IPTW, MW, and OW. The columns of panels represent the group contrast and then trimming methods. Within each panel, the X axis represents progressive increase in trimming threshold (more observations are trimmed off). 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.

See text for explanation.

Web Figure 9. Bias in log rate ratio estimates (no unmeasured confounding)

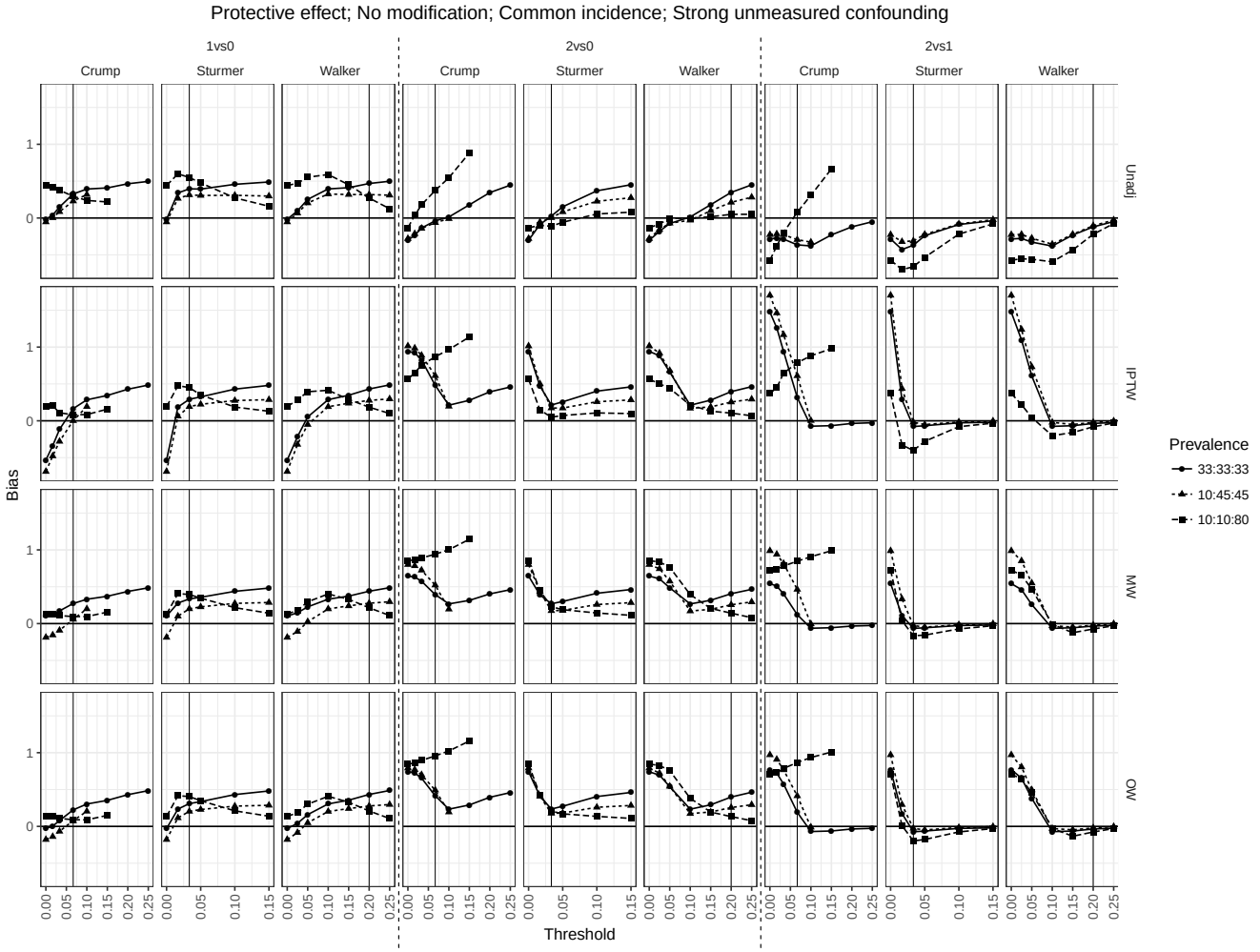


Panel layout: The rows of panels represent confounding adjustment methods: unadjusted, IPTW, MW, and OW. The columns of panels represent the group contrast and then trimming methods. Within each panel, the X axis represents progressive increase in trimming threshold (more observations are trimmed off). 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.

Web Figure 10. Bias in log rate ratio estimates (strong unmeasured confounding)

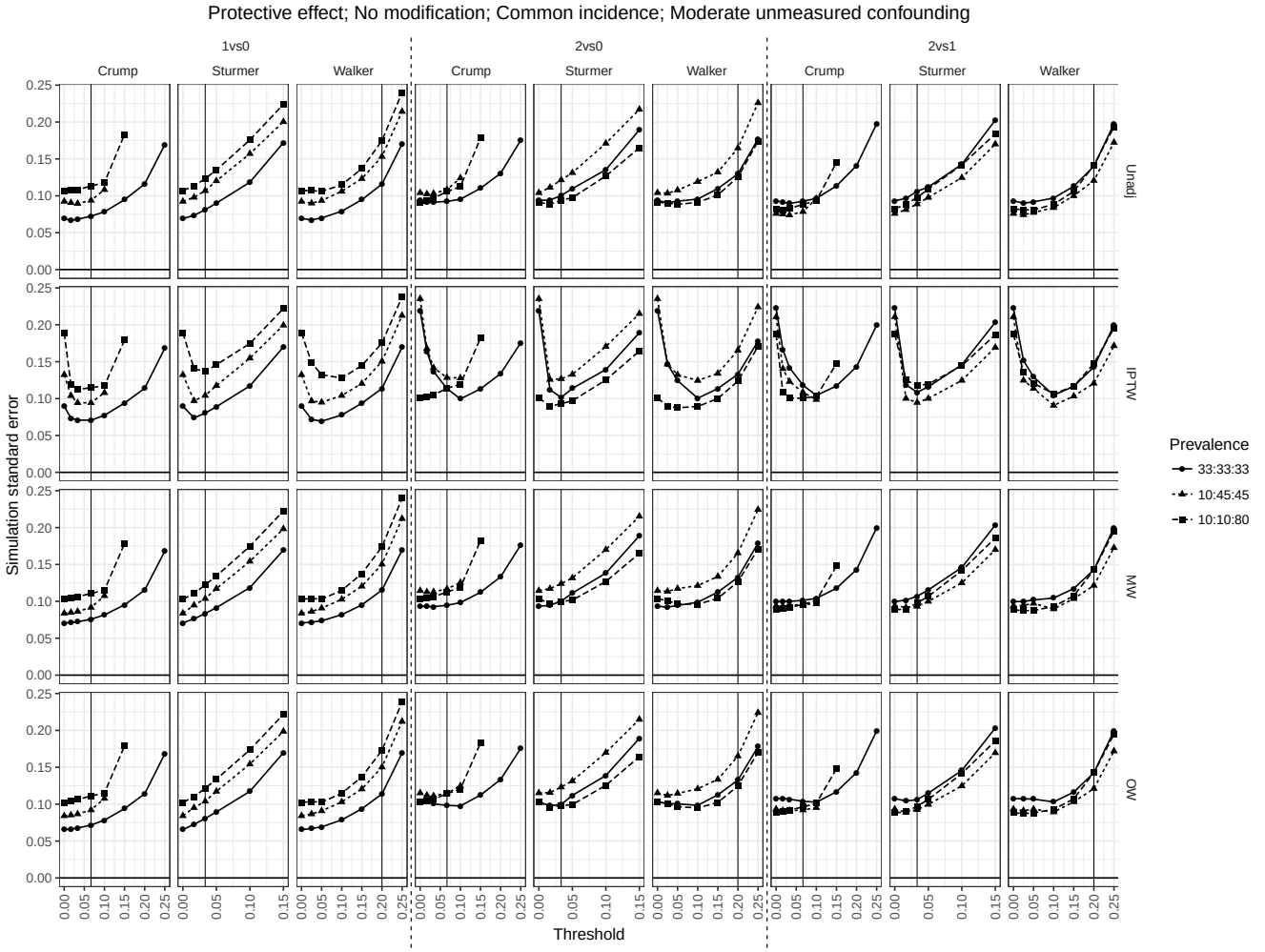


Panel layout: The rows of panels represent confounding adjustment methods: unadjusted, IPTW, MW, and OW. The columns of panels represent the group contrast and then trimming methods. Within each panel, the X axis represents progressive increase in trimming threshold (more observations are trimmed off). 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.

Web Figure 11. Variance of log rate ratio estimates (moderate unmeasured confounding)

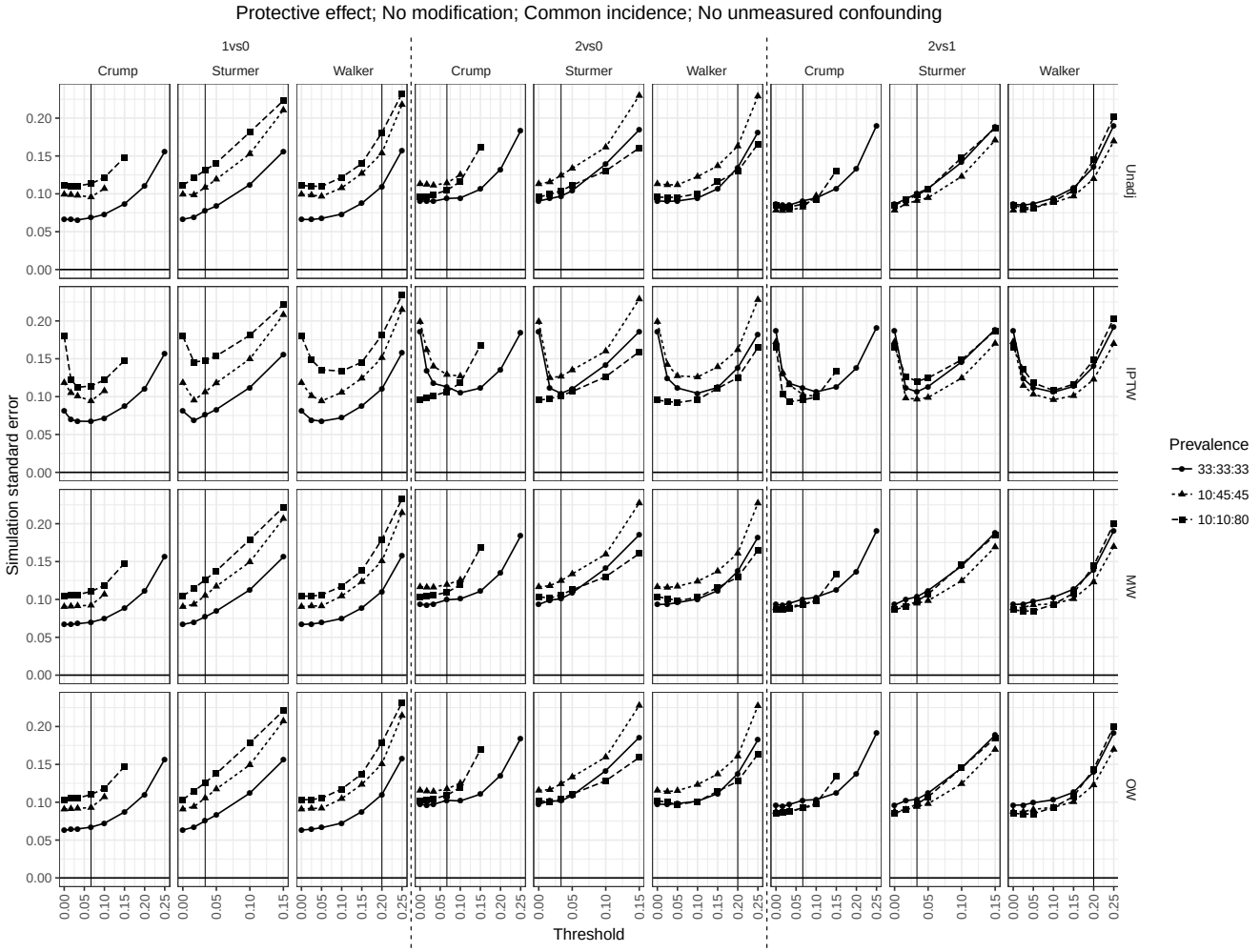


Panel layout: The rows of panels represent confounding adjustment methods: unadjusted, IPTW, MW, and OW. The columns of panels represent the group contrast and then trimming methods. Within each panel, the X axis represents progressive increase in trimming threshold (more observations are trimmed off). 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.

See text for explanation.

Web Figure 12. Variance of log rate ratio estimates (no unmeasured confounding)

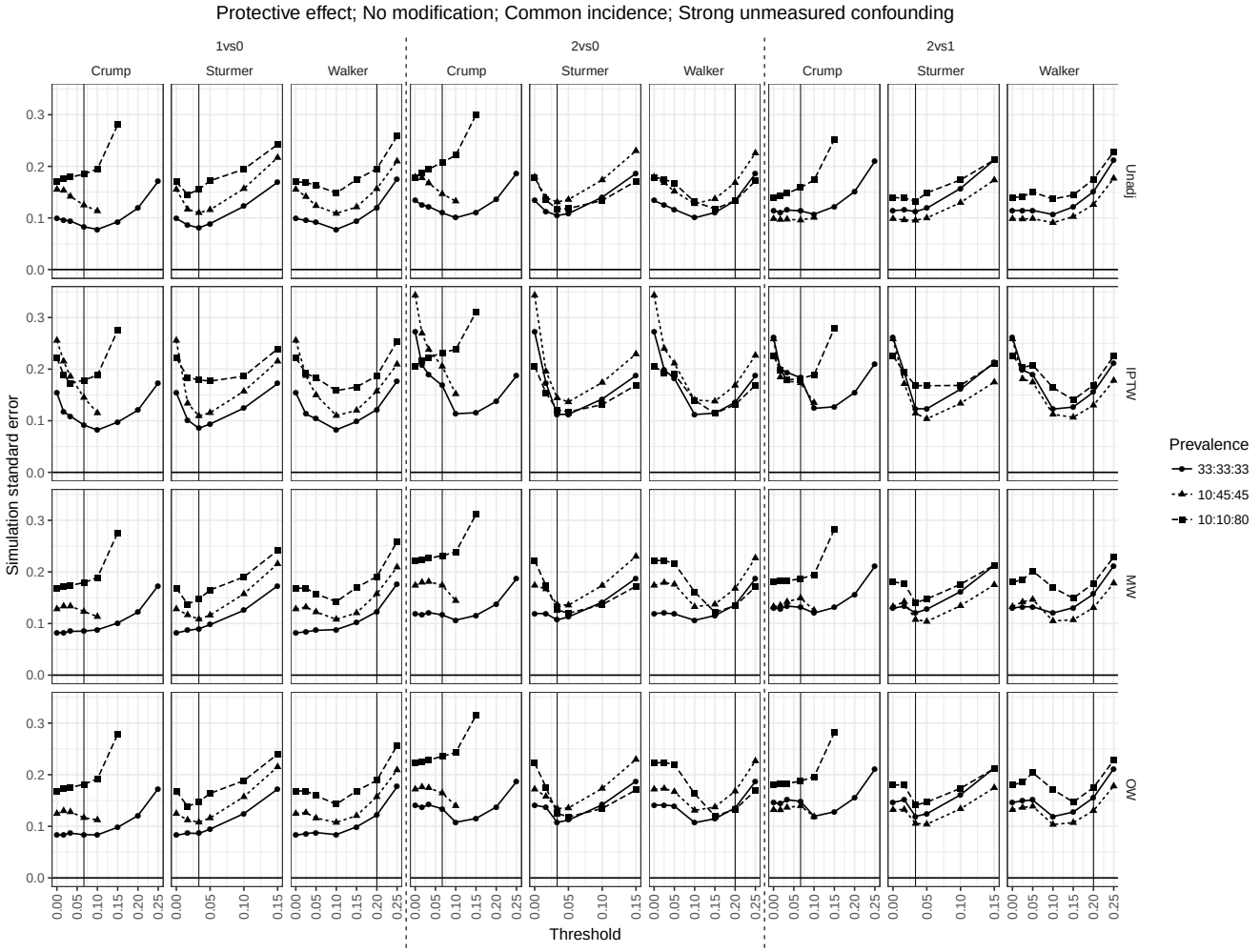


Panel layout: The rows of panels represent confounding adjustment methods: unadjusted, IPTW, MW, and OW. The columns of panels represent the group contrast and then trimming methods. Within each panel, the X axis represents progressive increase in trimming threshold (more observations are trimmed off). 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).

Web Figure 13. Variance of log rate ratio estimates (strong unmeasured confounding)

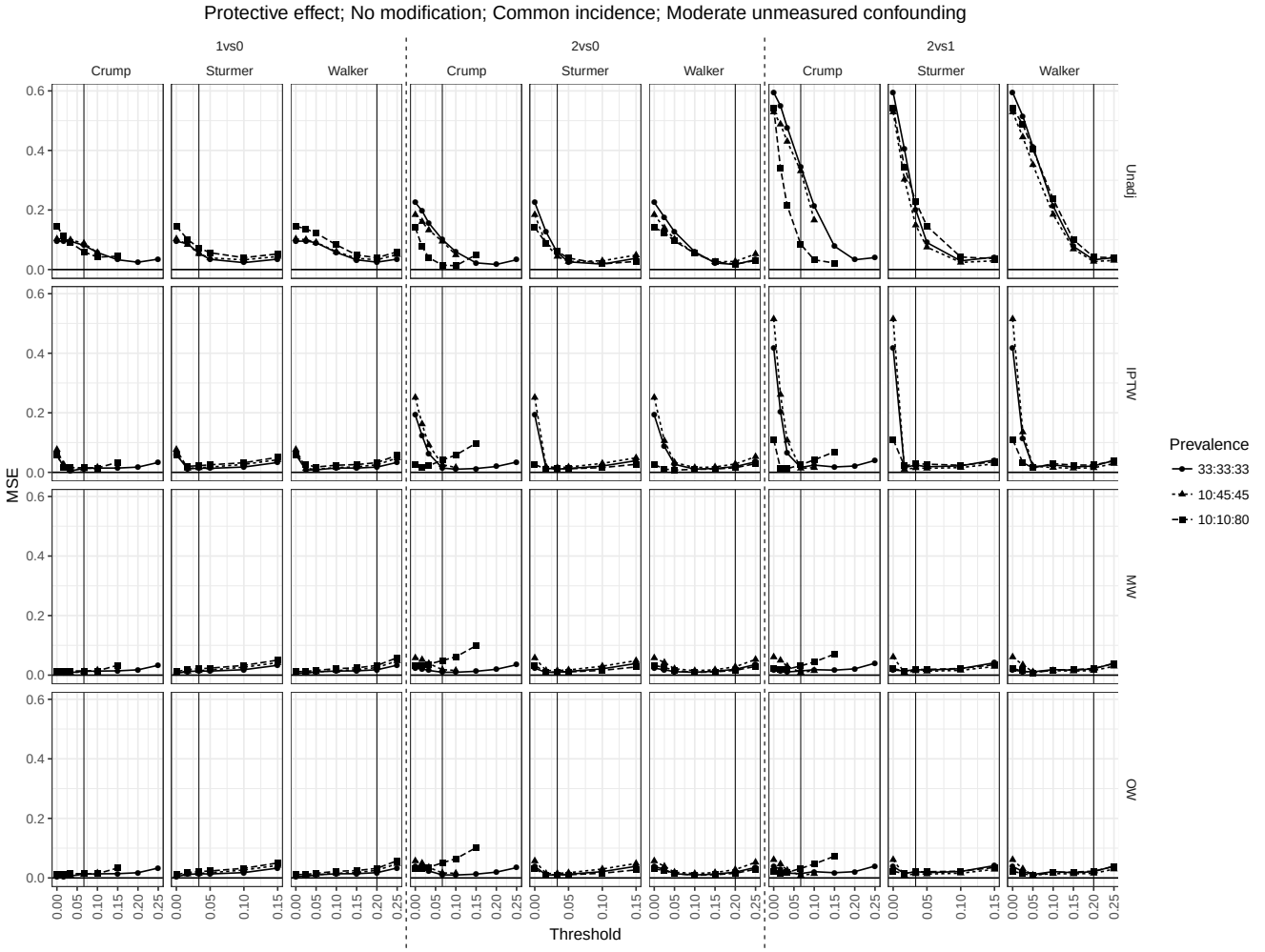


Panel layout: The rows of panels represent confounding adjustment methods: unadjusted, IPTW, MW, and OW. The columns of panels represent the group contrast and then trimming methods. Within each panel, the X axis represents progressive increase in trimming threshold (more observations are trimmed off). 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).

Web Figure 14. MSE of log rate ratio estimates (moderate unmeasured confounding)

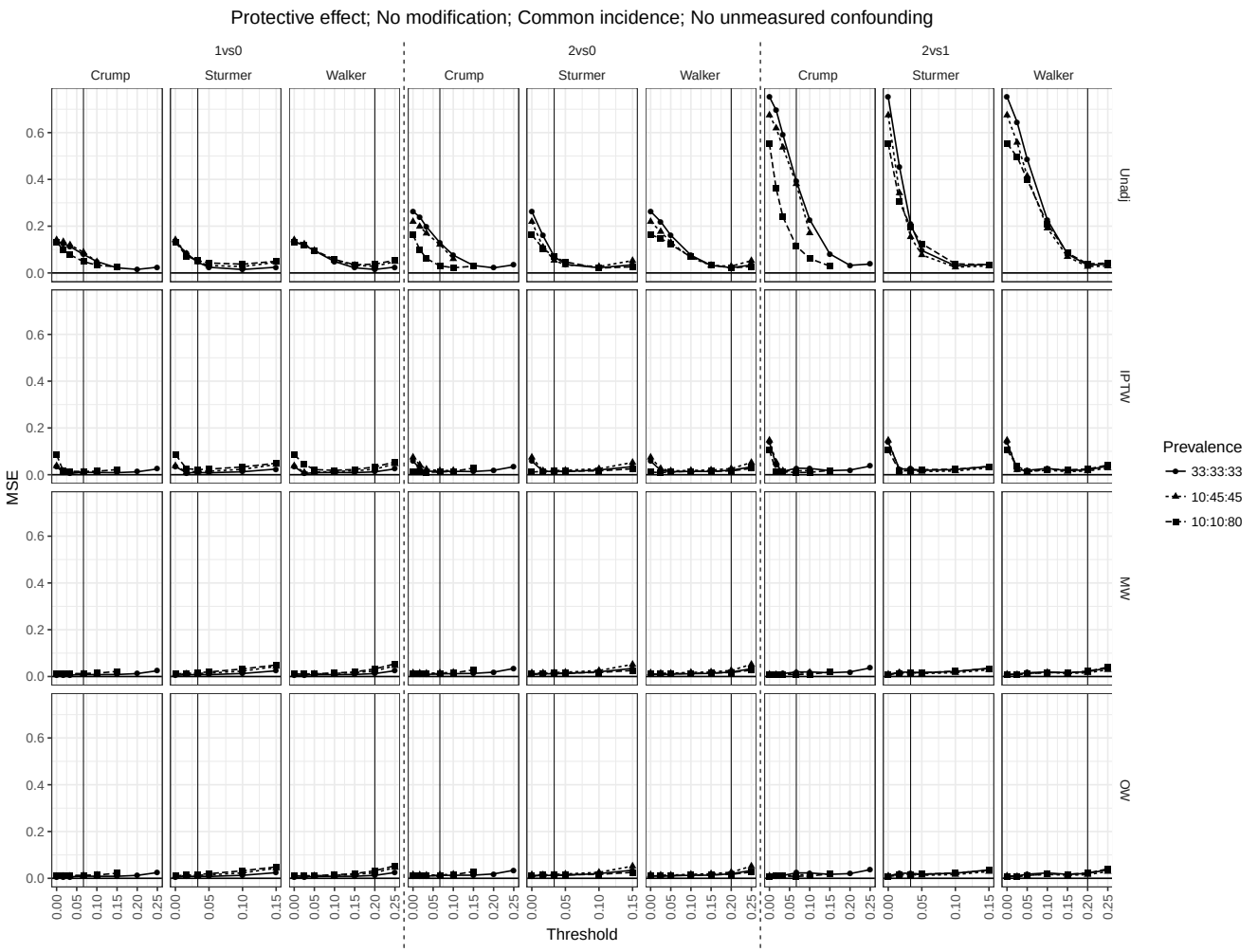


Panel layout: The rows of panels represent confounding adjustment methods: unadjusted, IPTW, MW, and OW. The columns of panels represent the group contrast and then trimming methods. Within each panel, the X axis represents progressive increase in trimming threshold (more observations are trimmed off). 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.

See text for explanation.

Web Figure 15. MSE of log rate ratio estimates (no unmeasured confounding)

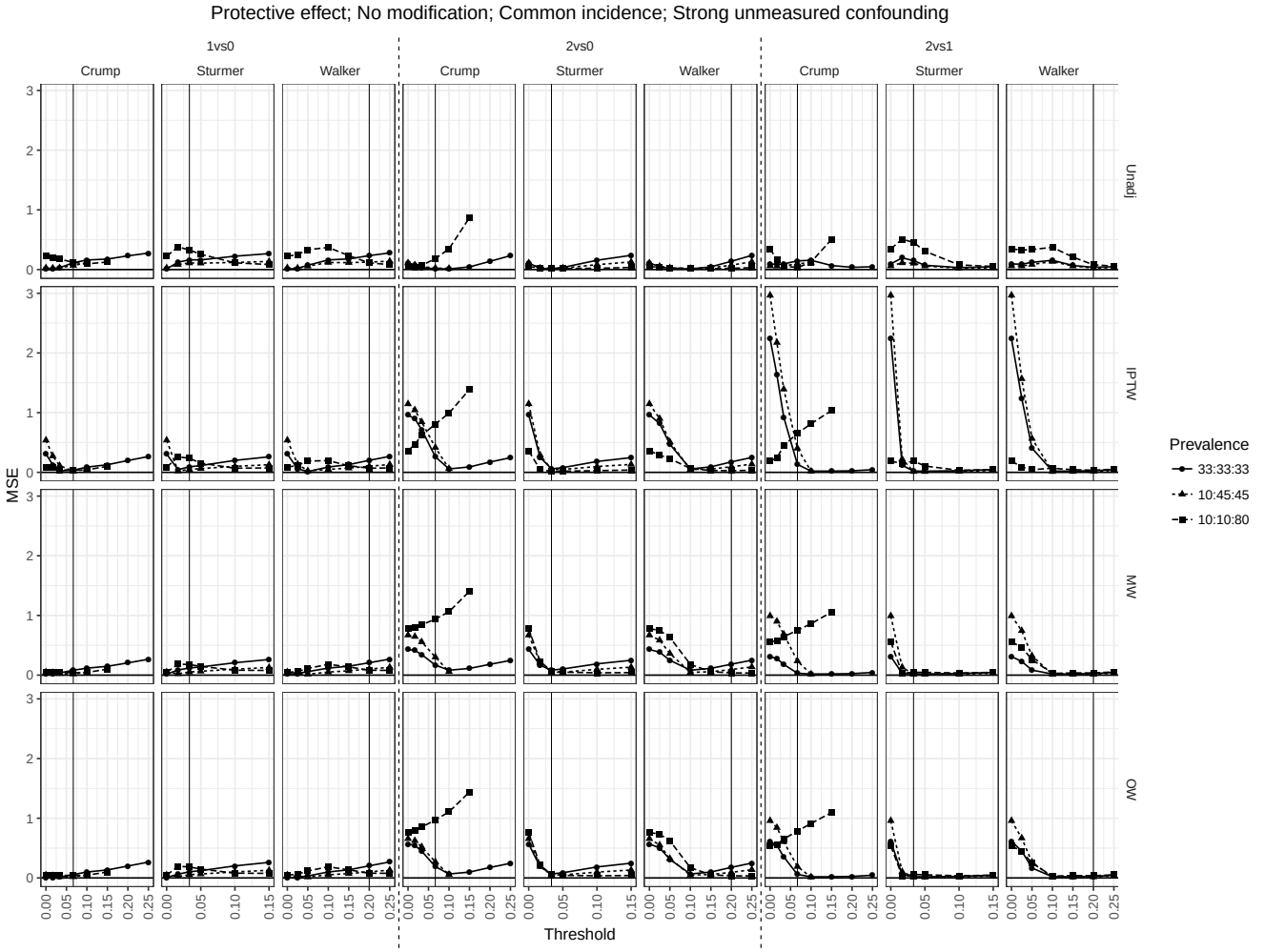


Panel layout: The rows of panels represent confounding adjustment methods: unadjusted, IPTW, MW, and OW. The columns of panels represent the group contrast and then trimming methods. Within each panel, the X axis represents progressive increase in trimming threshold (more observations are trimmed off). 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.

Web Figure 16. MSE of log rate ratio estimates (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.

Bibliography

[1] A. M. Walker, A. R. Patrick, M. S. Lauer, M. C. Hornbrook, M. G. Marin, R. Platt, V. L. Roger, P. Stang, and S. Schneeweiss, "A tool for assessing the feasibility of comparative effectiveness research," *Comp Eff Res*, vol. 2013, pp. 11–20, Jan. 2013.

[2] R. K. Crump, V. J. Hotz, G. W. Imbens, and O. A. Mitnik, "Dealing with limited overlap in estimation of average treatment effects," *Biometrika*, vol. 96, pp. 187–199, Jan. 2009.

[3] J. M. Robins, M. A. Hernán, and B. Brumback, "Marginal structural models and causal inference in epidemiology," *Epidemiology*, vol. 11, pp. 550–560, Sept. 2000.

[4] M. L. Petersen, K. E. Porter, S. Gruber, Y. Wang, and M. J. van der Laan, "Diagnosing and responding to violations in the positivity assumption," *Stat Methods Med Res*, vol. 21, pp. 31–54, Feb. 2012.

[5] T. Stürmer, K. J. Rothman, J. Avorn, and R. J. Glynn, "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.*, vol. 172, pp. 843–854, Oct. 2010.

[6] B. Freedman, "Equipoise and the ethics of clinical research," *N. Engl. J. Med.*, vol. 317, pp. 141–145, July 1987.

[7] D. H. Solomon, J. A. Rassen, R. J. Glynn, J. Lee, R. Levin, and S. Schneeweiss, "The comparative safety of analgesics in older adults with arthritis," *Arch. Intern. Med.*, vol. 170, pp. 1968–1976, Dec. 2010.

[8] E. Patorno, B. M. Everett, A. B. Goldfine, R. J. Glynn, J. Liu, C. Gopalakrishnan, and S. C. Kim, "Comparative cardiovascular safety of glucagon-like peptide-1 receptor agonists versus other antidiabetic drugs in routine care: A cohort study," *Diabetes Obes Metab*, vol. 18, pp. 755–765, Aug. 2016.

[9] T. W. Yee, "The VGAM Package for Categorical Data Analysis," *J Stat Softw*, vol. 29, pp. 1427–1445, Jan. 2010.

[10] T. P. Morris, I. R. White, and M. J. Crowther, "Using simulation studies to evaluate statistical methods," *arXiv:1712.03198 [stat]*, Dec. 2017.

[11] S. Greenland, J. M. Robins, and J. Pearl, "Confounding and Collapsibility in Causal Inference," *Statistical Science*, vol. 14, no. 1, pp. 29–46, 1999.

[12] L. Li and T. Greene, "A weighting analogue to pair matching in propensity score analysis," *Int J Biostat*, vol. 9, no. 2, pp. 215–234, 2013.

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

[14] F. Li, K. L. Morgan, and A. M. Zaslavsky, "Balancing Covariates via Propensity Score Weighting," *Journal of the American Statistical Association*, vol. 0, pp. 1–11, Dec. 2016.

[15] F. Li, L. E. Thomas, and F. Li, "Addressing Extreme Propensity Scores via the Overlap Weights [available online ahead of print September 5, 2018]," *Am. J. Epidemiol.*, vol. doi: 10.1093/aje/kwy201, Sept. 2018.

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

[17] N. Hamilton, "Ggtern: An Extension to 'ggplot2', for the Creation of Ternary Diagrams." https://CRAN.R-project.org/package=ggtern, Nov. 2018.