

Introduction

Inverse probability of treatment weighting (IPTW) is a reference method for balancing covariates between treatment groups and estimating unbiased causal effects in observational studies.

However, the presence of large or even extreme weights may lead to imprecise effect estimates and excessive variance, particularly at the tails of the propensity score (PS) distribution

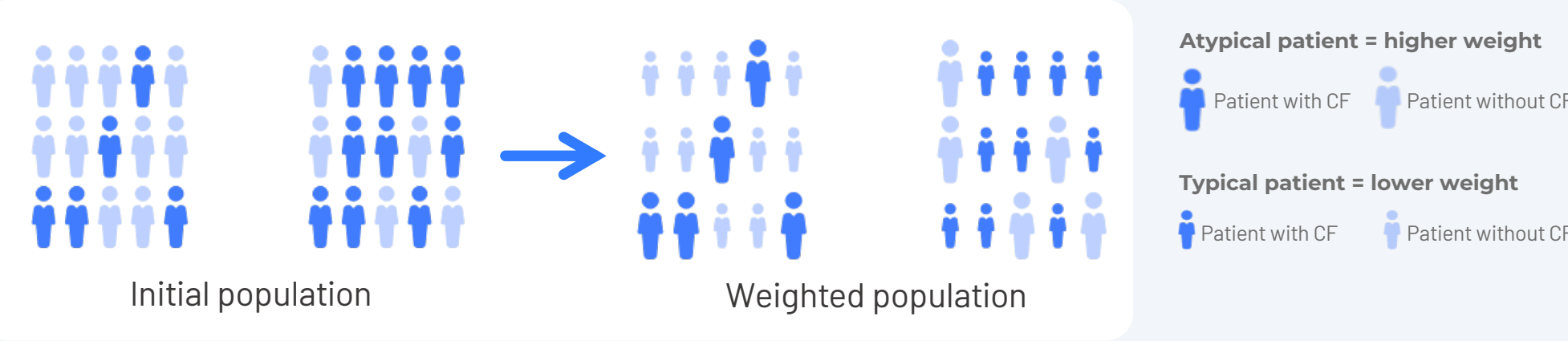
Objectives

This work examines different solutions for handling extreme IPTW weights depending on the targeted estimand.

Methods

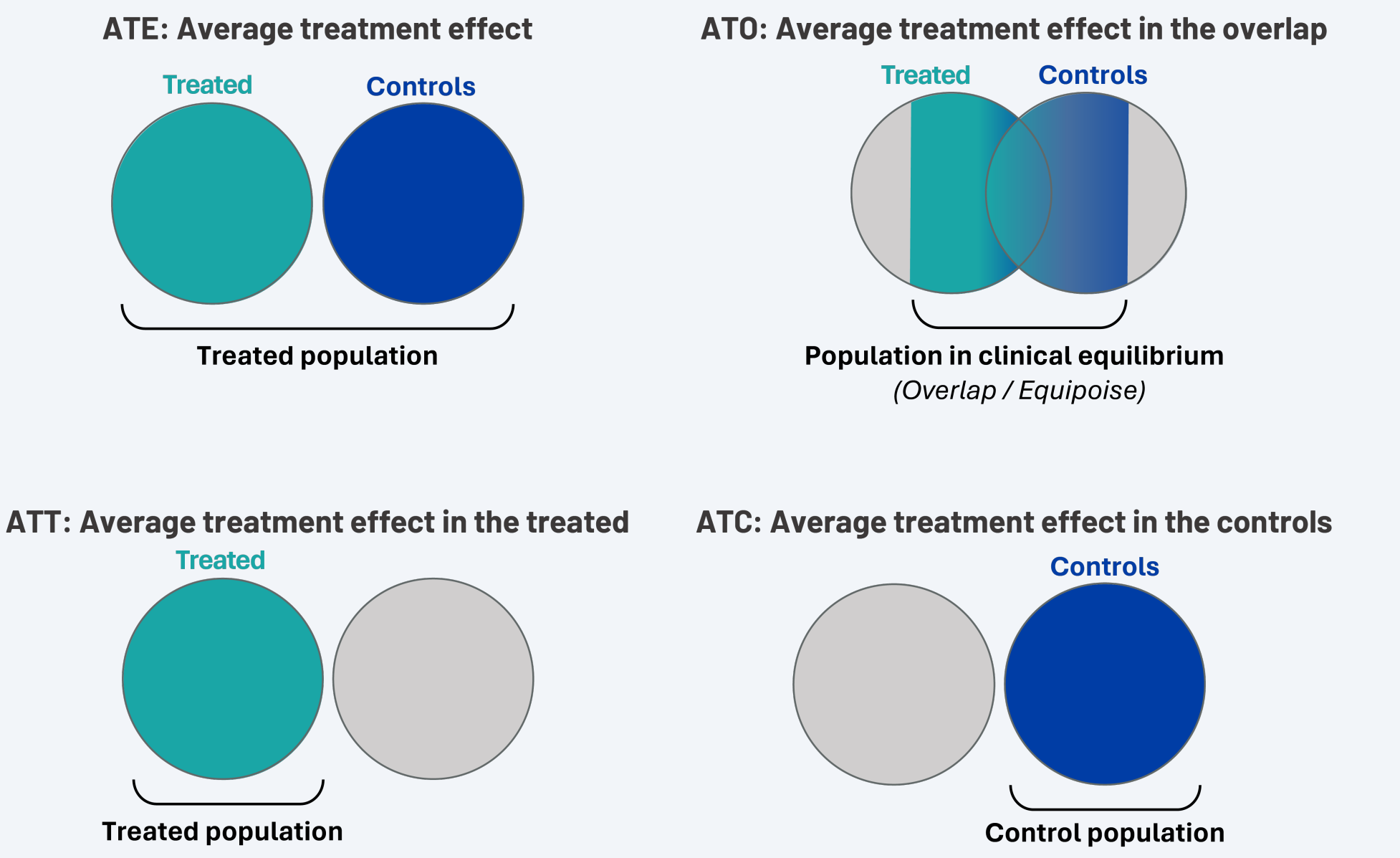
IPTW at a glance

- IPTW adjusts observational studies by balancing **confounders (CF)** between groups.
- It creates a **pseudo-population** where covariates have near-perfect balance.
- The **propensity score (PS)** is the probability of being exposed given the CF, estimated **using logistic regression**.
- Each individual is **weighted** by the inverse of their probability of actual exposure.
- Limitation:** only **measured CF** are accounted for, unlike in randomization.



Targeted estimand

It is essential to define the targeted estimand in order to select appropriate weights and strategies when dealing with extreme weights. Main estimands are:



- ATE:** Effect if everyone receives the treatment vs. no one
- ATO:** Effect in the population under clinical equilibrium (minimizes variance)
- ATT:** Effect among those who receive the treatment
- ATC:** Effect if controls (untreated individuals) had received the treatment

Research question and methodology

Using anonymized SNDS data comparing treatment A vs. B on a hospitalization outcome, we illustrate a case with a **PS imbalance**, leading to **extreme IPTW weights**. The benchmark analysis is the **PS matching** that estimates the ATT with RR = 0.73 (0.66; 0.80). Even though the common support is satisfactory, the PS distribution remains sufficiently heterogeneous to generate unstable weights in the weighting process, including for the **stabilised weights** used as a first resort. To tackle this, we suggest several approaches :

- Use **alternative weighting schemes** targeting the same estimand.
- Change the estimand** to adopt more appropriate weights.
- Apply **stabilization methods** such as truncation or trimming (on PS or on weights).

Estimates are obtained using a **weighted GLM**, and results are compared with the ones obtained with PS matching, which serve as methodological benchmarks.

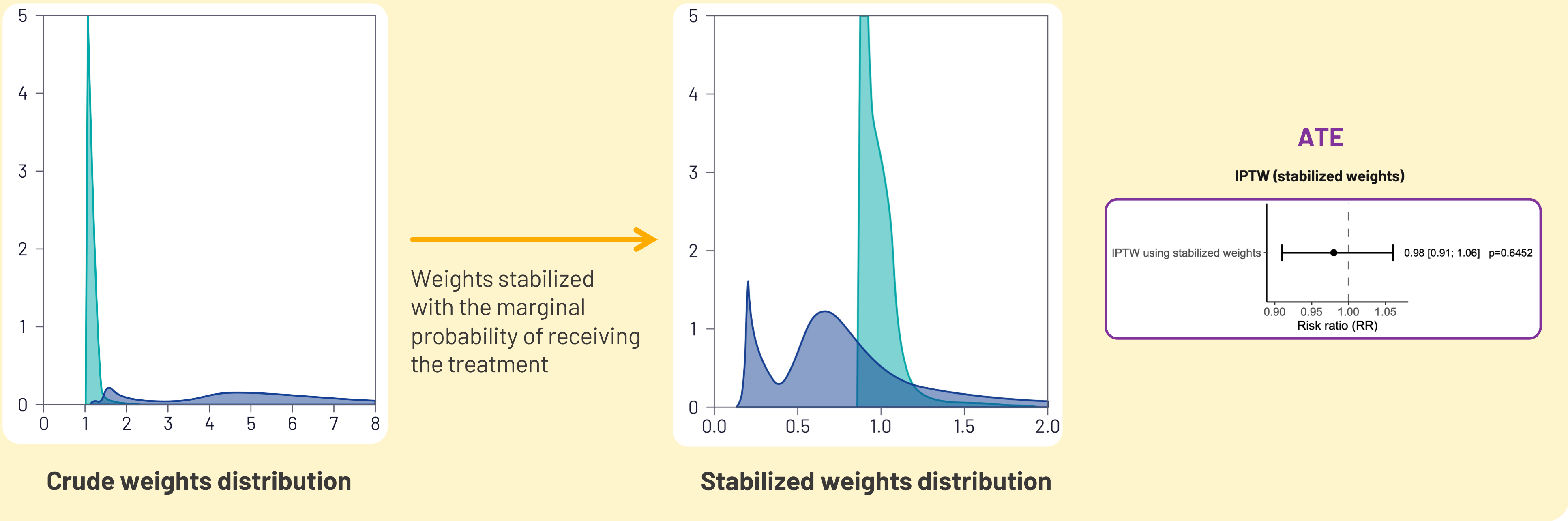
Weight	Treated	Untreated
IPTW with crude weights	$\frac{1}{PS}$	$\frac{1}{1 - PS}$
IPTW with stabilized weights	$\frac{P(X = 1)}{PS}$	$\frac{P(X = 0)}{1 - PS}$
Entropy	$\frac{-(PS * \log(PS) + (1 - PS) * \log(1 - PS))}{PS}$	$\frac{-(PS * \log(PS) + (1 - PS) * \log(1 - PS))}{1 - PS}$
Overlap weights	$1 - PS$	PS
Matching weights	$\frac{\min(PS, 1 - PS)}{PS}$	$\frac{\min(PS, 1 - PS)}{1 - PS}$

Conclusion

IPTW creates a balanced pseudo-population and can serve as a basis for more advanced methods incorporating repeated weighting over time. The choice of weighting method must be transparently reported and adapted to the study context. An appropriate choice improves treatment effect estimation and leads to more robust causal estimates.

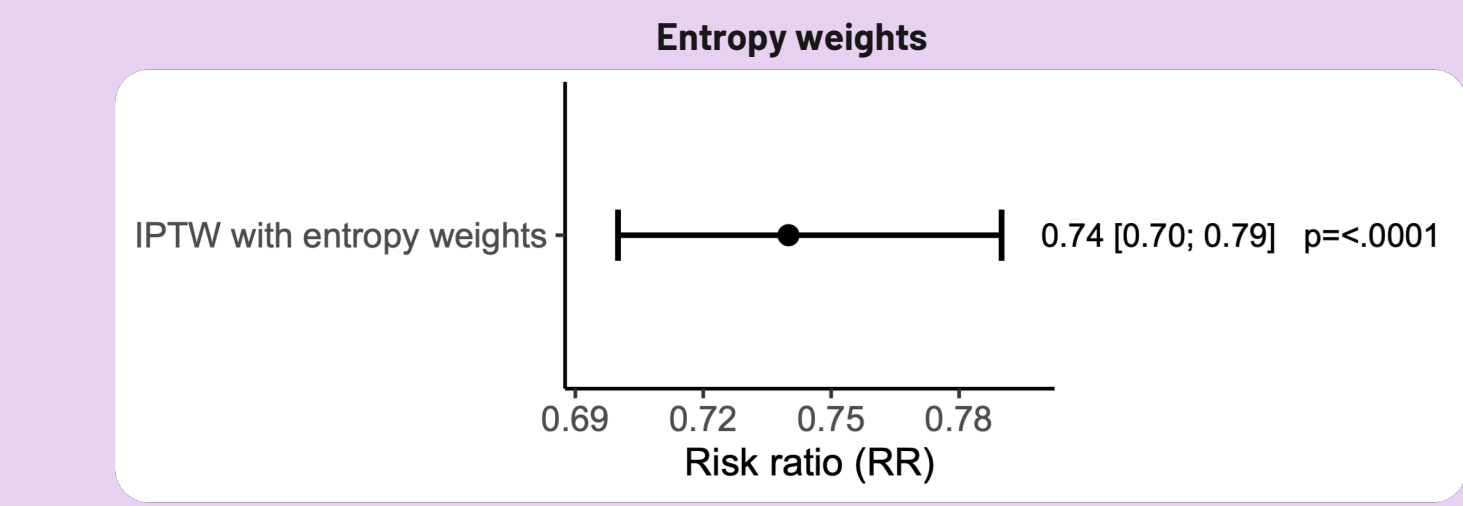
Results

Study with PS imbalance



1. Alternative weights targeting the same estimand (ATE)

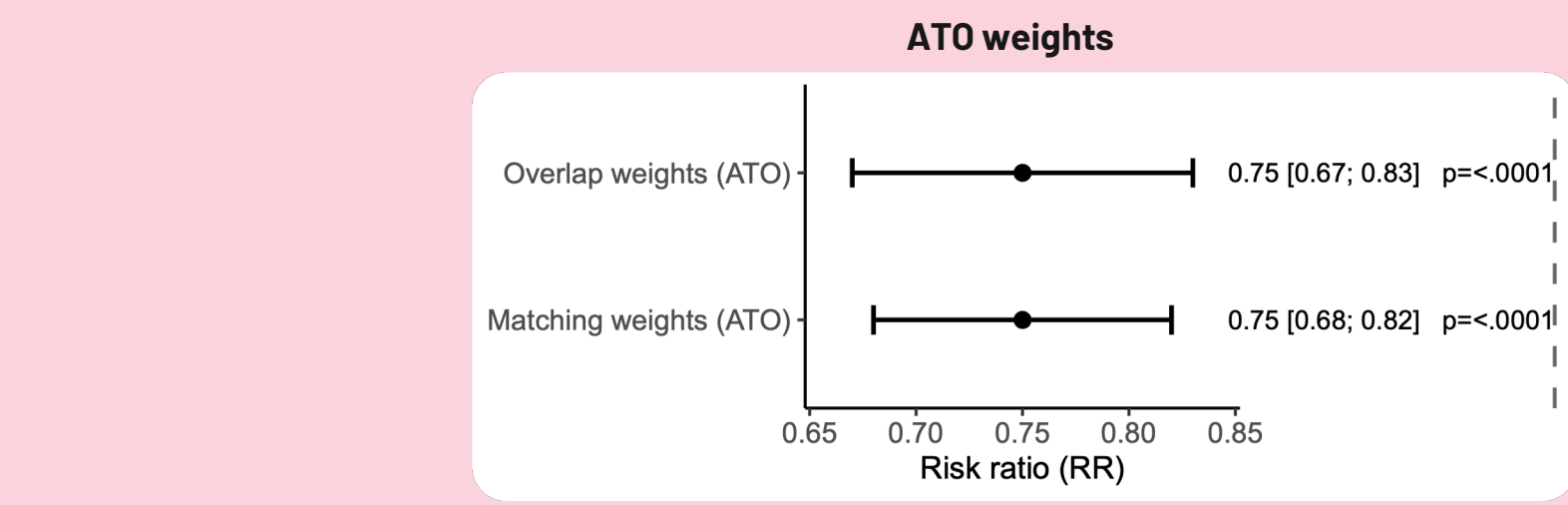
Entropy weighting was originally designed as an iterative process in order to improve covariate balance. It uses an entropy-based function of the PS to weight observations.



2. Change estimand to adopt more appropriate weights (ATO)

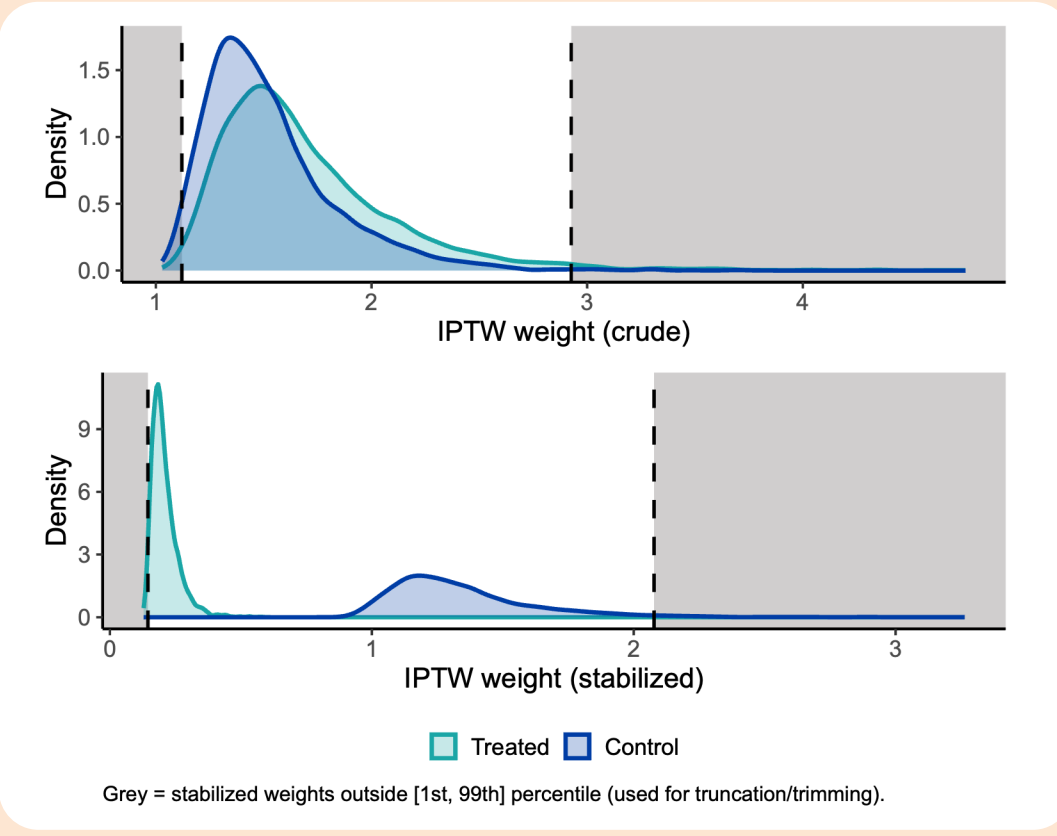
Matching weighting mimics 1:1 matching without replacement while accounting for treatment prevalence.

Overlap weighting weights patients according to their PS, prioritizing those in equipoise (probability ≈ 0.5) and minimizing the influence of extreme cases.



3. Apply stabilization methods on PS or weights while keeping the same estimand (ATE)

Weight-based methods (crude or stabilized)

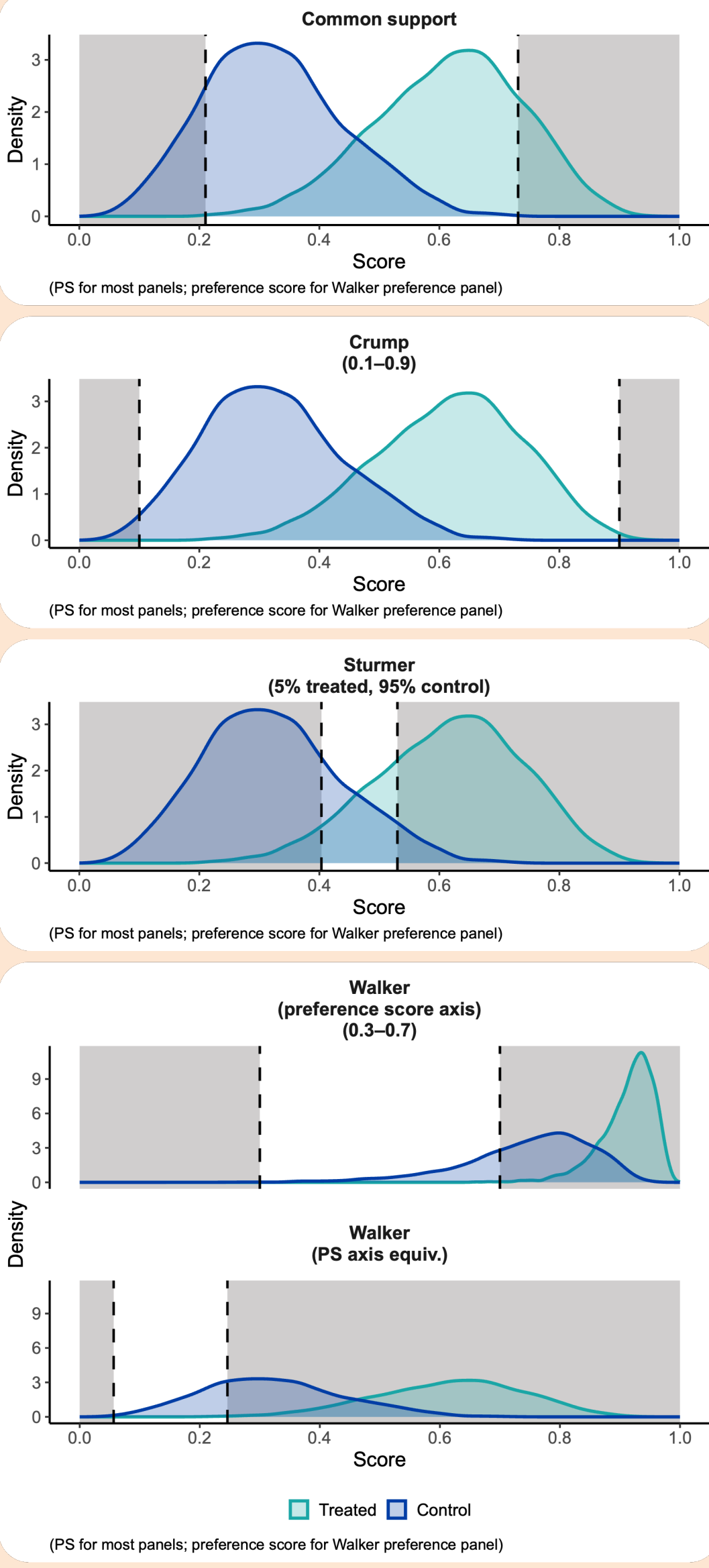


Truncation or winsorization replaces extreme IPTW weights with predefined thresholds (e.g., 1st and 99th percentiles), reducing their influence while keeping all the individuals in the analysis.

Trimming permanently excludes from the analysis individuals whose weights exceed predefined thresholds (e.g., below 1st and beyond 99th percentiles of the weight).

Both of these methods can be applied to crude or stabilized weights.

PS-based methods



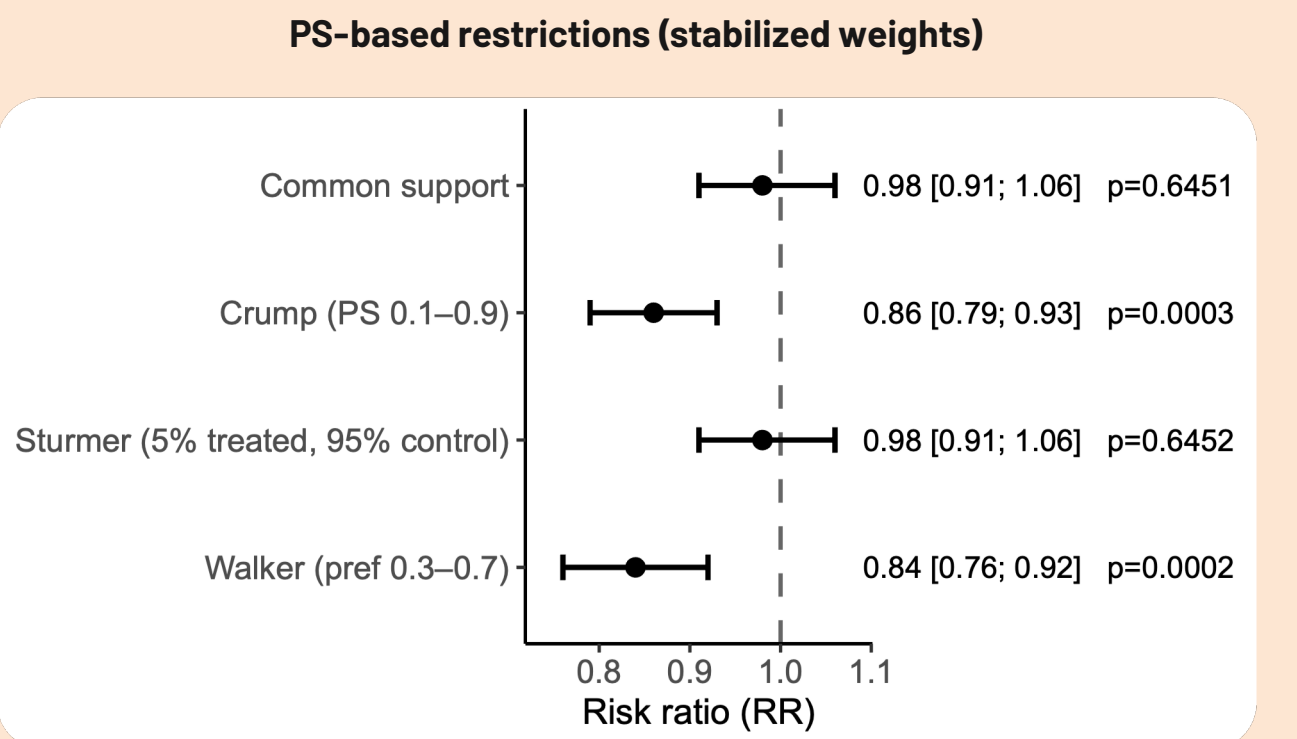
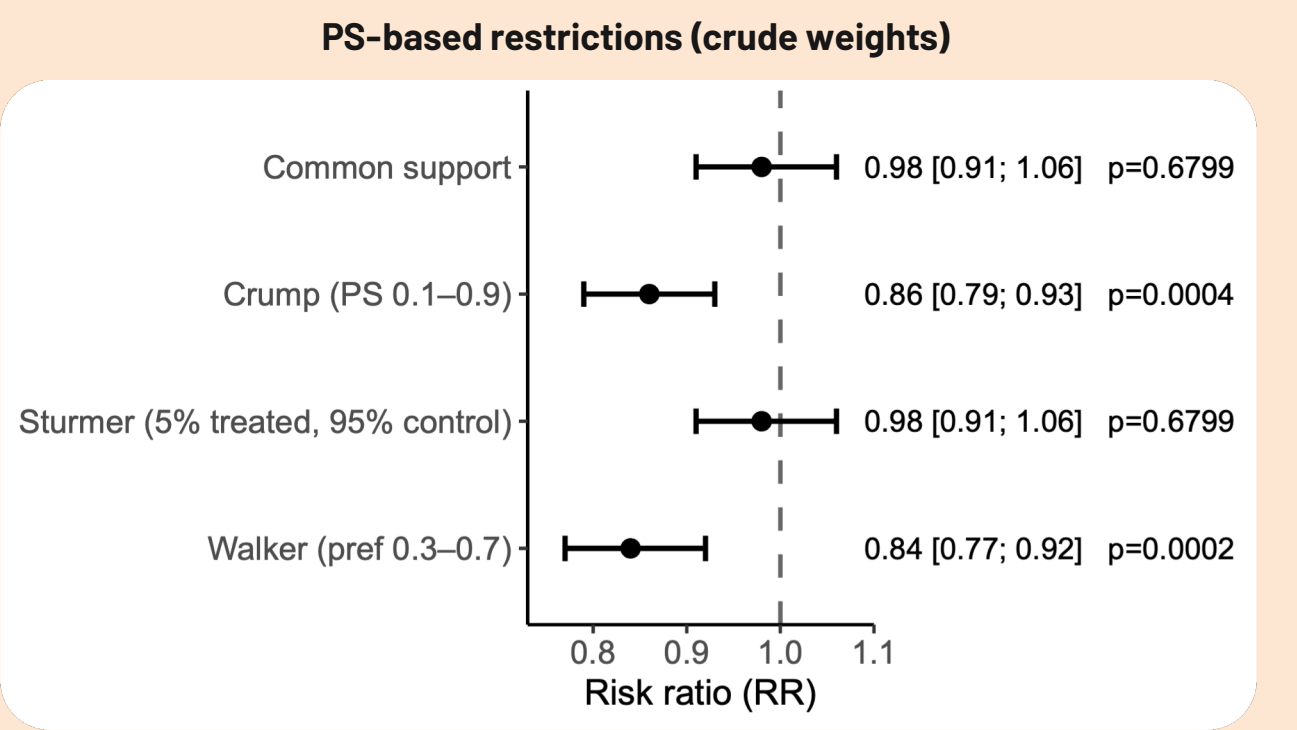
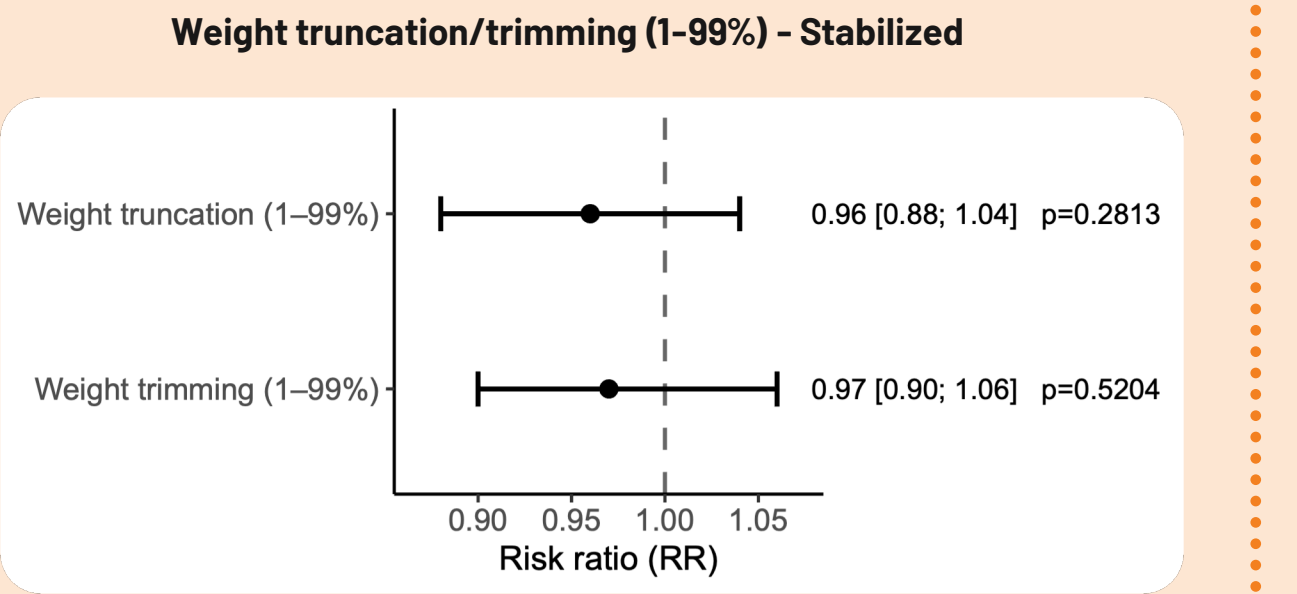
The **common support trimming** excludes individuals whose PS is lower than the minimum observed among treated individuals or higher than the maximum observed among controls, ensuring complete overlap between groups.

The **Crump** method excludes individuals with extremes PS values (PS ≤ 0.1 or PS ≥ 0.9), thus minimizing the asymptotic variance of the causal effect estimate.

The **Stürmer** method excludes individuals below the 5th percentile of the PS among treated individuals or above the 95th percentile among untreated individuals i.e. controls, thus removing observations outside of the overlap region.

The **Walker** method uses a preference score derived from the PS and excludes individuals with a preference score ≤ 0.3 (treated) or ≥ 0.7 (controls), targeting the region of clinical equipoise.

$$\text{logit(Preference score)}$$
$$= \text{logit(PS)} - \text{logit(Treatment prevalence)}$$



Key findings

Using SNDS data, PS matching (the gold standard) yielded a risk ratio (RR) of 0.73 for the ATT. Stabilized IPTW resulted in an RR of 0.98, truncation methods gave RRs ranging from 0.96 to 0.98, and trimming methods yielded RRs between 0.81 and 0.99. For the ATO, overlap weighting and matching weights both resulted in an RR of 0.75, while entropy weighting produced an RR of 0.74. Overall, changing the weighting method or the estimand (ATE → ATO) led to results closer to the gold standard. Restriction or stabilization methods were generally less effective, except for the Walker method. Other weighting approaches targeting different estimands (ATT, ATC) could also be explored.

Acknowledgement

The authors would like to thank the following teams for data extraction: Direction de la stratégie, des études et des statistiques, du Département Accès, Traitement et Analyse de la Donnée et des Demandes Externes de la Caisse nationale de l'assurance maladie (CNAH).

Bibliography

Stürmer T, Webster-Clark M, Lund JL, Wyss R, Ellis AR, Lunt M, Rothman KJ, Glynn RJ. Propensity Score Weighting and Trimming Strategies for Reducing Variance and Bias of Treatment Effect Estimates: A Simulation Study. Am J Epidemiol. 2021 Aug 1;190(8):1659-1670. doi: 10.1093/aje/kwab041. Chesnaye NC, Stel VS, Tripepi G, et al. An introduction to inverse probability of treatment weighting in observational research. Clin Kidney J. 2021;15(1):14-20. Published 2021 Aug 26. doi:10.1093/ckj/sfab158. Kostouraki A, Hajage D, Rachet B, et al. On variance estimation of the inverse probability-of-treatment weighting estimator: A tutorial for different types of propensity score weights. Stat Med. 2024;43(13):2672-2694. doi:10.1002/sim.10078