Journal of Clinical Epidemiology 187 (2025) 111942

COMMENTARY

When and why to use overlap weighting: clarifying its role, assumptions, and estimand in real-world studies

John G. Rizk[a,b,*]

[a]Department of Practice, Sciences, and Health Outcomes Research, University of Maryland School of Pharmacy, Baltimore, MD, USA
[b]Department of Epidemiology and Public Health, University of Maryland School of Medicine, Baltimore, MD, USA

Accepted 17 August 2025; Published online 22 August 2025

Abstract

Objectives: To examine the strengths and limitations of overlap weighting in observational studies and to clarify when it is appropriate to use this method based on the target estimand.

Study Design and Setting: This is a narrative commentary that reviews recent methodological developments and real-world examples to highlight how overlap weighting operates, when it provides advantages over methods like inverse probability of treatment weighting, and the importance of aligning analytic methods with the causal question and estimand.

Results: Overlap weighting produces bounded, stable weights and achieves exact mean covariate balance in the subset of patients with overlapping treatment probabilities near 0.5—those considered to be in clinical equipoise. However, it targets the average treatment effect in the overlap population (ATO), a statistically defined subgroup that is difficult to characterize clinically. Use of this method without pre-specifying interest in the ATO may lead to misinterpretation of results. While overlap weighting improves statistical performance, it limits generalizability and interpretability. Study design and inclusion/exclusion criteria remain critical for addressing violations of positivity.

Conclusion: Overlap weighting is most appropriate when the research question explicitly targets the overlap population. It should not be adopted solely to resolve estimation issues with average treatment effect or average treatment effect in the treated methods. Researchers must define their target estimand before choosing a method and clearly report the characteristics of both the unweighted and overlap-weighted populations to ensure valid causal inference.

Keywords: Propensity scores; Confounding adjustment; Weighting methods; Real-world evidence; Pharmacoepidemiology

Plain Language Summary

Overlap weighting is a statistical method used in health research to compare treatments when people are not randomly assigned to different options. It focuses on patients who could realistically receive either treatment and helps improve the fairness and precision of comparisons. However, the results apply only to this specific group and not everyone in the study. Researchers should choose this method only when it fits the question they are asking.

1. Introduction

The use of overlap weighting is becoming increasingly popular and has grown substantially in recent years as researchers seek more robust alternatives to conventional confounding adjustment methods. The method was introduced by Li et al in 2016 and formally published in 2018

Funding: None.

* Corresponding author. Department of Practice, Sciences, and Health Outcomes Research, University of Maryland School of Pharmacy, Baltimore, MD, USA 21201.
E-mail address: john.rizk@umaryland.edu

https://doi.org/10.1016/j.jclinepi.2025.111942

What is new?

Key findings

- Overlap weighting emphasizes patients with propensity scores near 0.5.
- Bounded weights reduce variance and improve covariate balance without trimming.

What this adds to what is known?

- Overlap weighting is not a workaround for unstable treatment effect estimates.
- Choice of weighting must align with the study's causal question and estimand.

What is the implication and what should change now?

- Studies should report clearly who belongs to the overlap population.
- Transparent reporting will improve interpretation and reproducibility of findings.

[1]. To assess this trend, I searched PubMed using the query (as of May 3, 2025): ("overlap weight" [Title/Abstract] OR "overlap weighting" [Title/Abstract] OR "overlap-weighted" [Title/Abstract] OR "overlap weighted" [Title/Abstract]) AND propensity. This search returned over 300 results (this includes original studies, reviews, etc.), with a clear year-over-year increase in published studies using this method. As shown in the PubMed visualization, the number of studies employing overlap weighting peaked in 2024, with over 100 publications that year alone. This pattern reflects the method's rising influence in epidemiology, health outcomes research, and comparative effectiveness studies, especially in real-world evidence settings where covariate imbalance and weak overlap are common challenges. Overlap weighting is a propensity score (PS)-based method that improves upon earlier approaches like inverse probability of treatment weighting (IPTW) by using bounded, stable weights that emphasize individuals with a PS near 0.5 —those in clinical equipoise, who have approximately equal probabilities of receiving either treatment based on their baseline characteristics. By down-weighting individuals with extreme PS (near 0 or 1) and focusing on those closer to 0.5, overlap weighting achieves superior covariate balance and greater precision, especially in settings with limited overlap or relatively small sample sizes. These strengths make overlap weighting particularly well-suited for real-world data, where baseline imbalances between groups often occur.

2. Study design, positivity, and the primacy of the estimand

In observational studies, the choice of causal estimand, whether the average treatment effect (ATE), the average treatment effect in the treated (ATT), or the average treatment effect in the overlap population (ATO), must precede the analytic method. As outlined in the causal inference roadmap [2], identifying the target population and the scientific question should guide the design and selection of PS techniques, not the other way around.

A critical assumption underlying all PS-based methods is positivity, the requirement that every individual in the target population has a nonzero probability of receiving each treatment option, conditional on observed covariates [3]. This assumption is central to valid causal inference, as it helps ensure there is overlap in the covariate distributions between treatment groups [4,5]. Without positivity, key subgroups will lack comparable counterparts, making the estimation of treatment effects for those individuals either unreliable or entirely infeasible.

Violations of the positivity assumption can be broadly categorized as either stochastic or structural [3,5]. Stochastic violations may result from small sample sizes or PS model misspecification, and can often be mitigated through model refinement or increased sample size. In contrast, structural violations arise when treatment assignment is strongly influenced by clinical characteristics, such that certain subgroups rarely, if ever, receive one of the treatment options. Structural violations must be handled at the design stage, through appropriate inclusion and exclusion criteria that define a clinically relevant and causally identifiable population.

The presence of extreme weights in methods like IPTW often signals a violation of the positivity assumption. While techniques like stabilized weights can reduce variance inflation caused by extreme weights [6], they do not resolve the underlying issue, namely, that some individuals in the study population are highly unlikely to receive one of the treatment options (ie, poor covariate overlap). No method, including overlap weighting, g-computation, or targeted maximum likelihood estimation (TMLE), can fully recover from structural positivity violations, as all rely on having sufficient treatment variation across levels of covariates. When the goal remains to estimate the ATE, and some extrapolation beyond observed data support is acceptable, model-based approaches such as g-computation or TMLE may offer more stable alternatives [7]. These approaches are less influenced by extreme weights but depend on stronger modeling assumptions, and their performance may deteriorate in regions of the covariate space where certain combinations of covariates are rare or underrepresented [8−10]. A more scientifically rigorous strategy is to address potential violations of positivity at the design stage, for example, by refining the inclusion and exclusion criteria

to define a population in which all individuals could plausibly receive either treatment [11]. This approach helps ensure sufficient overlap and supports valid causal inference.

It is important to emphasize that the ATO estimated using overlap weighting should not be used merely as a workaround when ATE or ATT estimation encounters challenges like extreme or unstable weights. The ATO is a fundamentally different estimand, targeting individuals whose PS are near 0.5—those who have a roughly equal probability of receiving either treatment based on observed covariates [1]. These individuals are considered to be in clinical equipoise [1]. Shifting to the ATO in response to estimation difficulties risks distorting the original scientific question and may compromise the interpretability or relevance of results. Researchers must recognize that each estimand addresses distinct causal questions and serves different inferential goals. The decision to adopt overlap weighting should therefore be driven by a clearly articulated interest in the overlap population. When a shift to the ATO is warranted, it must be accompanied by an explicit acknowledgment of the revised causal question and the corresponding trade-offs in generalizability, interpretability, and policy relevance.

3. Limitations of IPTW in real-world settings

With the conceptual groundwork established, we now turn to the limitations of IPTW and why overlap weighting offers a more robust alternative under certain conditions, particularly when the target estimand is the ATO.

IPTW is a widely used PS method for estimating ATE, but it has several well-documented limitations. A major concern is the occurrence of extreme weights, particularly for individuals with very high or very low PS [4,12,13]. In this approach, each individual is assigned a weight based on the inverse of the probability of receiving the treatment they actually received [14]. Specifically, treated individuals receive a weight of 1 divided by their PS (1/PS), while untreated individuals are weighted by 1 divided by one minus their PS (1/(1−PS)) [14]. This becomes especially problematic when an individual's actual treatment does not align with what their score would predict; for example, when someone who was unlikely to be treated based on their covariates (ie, low PS) is treated, or when someone likely to be treated (ie, high PS) is not. These rare cases receive disproportionately large weights, which can inflate variance, destabilize estimates, and allow a small number of individuals to exert undue influence on the analysis [15,16].

To address this, researchers often apply trimming (removing individuals with extreme PS values), truncation (capping large weights), or use stabilized weights [17]. However, each of these strategies come with trade-offs: trimming can reduce sample size and shift the population under study, potentially affecting generalizability; truncation may be arbitrary and still leave residual imbalance [17]; and while stabilized weights reduce variance, they may not fully eliminate extreme weights, sometimes necessitating additional truncation [18]. Moreover, although trimming may remove the most extreme weights, the choice of cutoff threshold can substantially influence covariate balance and does not guarantee that treated and untreated groups will be comparable [19].

IPTW is most robust under specific conditions: when the treated and untreated cohorts are relatively similar in baseline characteristics (ie, there is sufficient overlap), when extreme outliers are absent, and when the population size is adequate. Under these circumstances, the estimated PS tends not to produce extreme weights, and IPTW performs more reliably [4].

In observational data, where initial differences between treatment groups may be substantial (ie, low PS overlap), IPTW methods can inadvertently alter the target population (by disproportionately amplifying the influence of a small subset of individuals with extreme weights), fail to achieve adequate covariate balance, or significantly reduce estimation precision [1,20]. Low overlap itself indicates that the two groups are very different at baseline.

3.1. Example: IPTW challenges in comparing direct oral anticoagulants (DOACs) vs Warfarin

This example illustrates how IPTW can lead to instability and bias in real-world comparisons when treatment groups differ substantially at baseline, especially when the cohort is small. In a study comparing DOACs vs. warfarin for atrial fibrillation, patients prescribed DOACs are often younger and healthier, while those on warfarin tend to be older with more comorbidities (eg, advanced chronic kidney disease or end-stage renal disease); this prescribing pattern reflects clinical caution and the fact that DOACs are sometimes avoided in patients with advanced kidney dysfunction, leading to a sicker, older population in the warfarin group [21,22]. These differences create poor overlap in PS: DOAC users can have scores near 1, and warfarin users near 0 (PS reflects the probability of receiving a DOAC). Under IPTW, patients whose treatment aligns with their predicted probability, such as a DOAC user with a high PS or a warfarin user with a low PS, receive moderate weights (eg, $1/0.9 \approx 1.1$ or $1/0.1 \approx 1.1$), contributing proportionally to the analysis. However, serious issues arise when treatment contradicts expectations. In such cases, individuals with extreme PS receive very large weights. For example, a DOAC user with a PS of 0.05 (indicating they were unlikely to be treated with a DOAC) would receive a weight of 1/0.05, which equals 20. Similarly, a warfarin user with a PS of 0.95 (indicating they were highly likely to be treated with a DOAC)

would also receive a weight of 1/(1−0.95), which equals 20. Even more extreme cases include a DOAC user with a PS of 0.02 or a warfarin user with a PS of 0.98, each would receive a weight of 50.

Now let us consider a sample of 1000 patients. If just five patients have extreme weights of 20−50, each one contributes the equivalent of 20−50 patients in the weighted analysis, meaning five individuals could account for the influence of 100 to 250 patients. When multiple such cases exist, as is common in real-world data, they can dominate the treatment effect estimate, inflate standard errors, and undermine both precision and generalizability [4,12,13].

As discussed, researchers may address this issue by trimming patients with extreme PS or truncating their weights; however, these strategies involve trade-offs. Trimming can reduce sample size and exclude clinically relevant patients, potentially distorting the study population, while truncation may mask covariate imbalance and introduce bias. Both approaches can undermine the interpretability and validity of IPTW, particularly in real-world comparative studies like this one.

4. How overlap weighting works

Overlap weighting is gaining traction as a preferred method for confounding adjustment in observational studies, particularly in settings where traditional PS techniques like IPTW may perform poorly. As introduced earlier, overlap weighting centers analysis on patients in clinical equipoise, which is defined as those with a PS around 0.5 [1,16], meaning they have approximately equal probabilities of receiving either treatment based on their baseline characteristics. By assigning greater weight to individuals in this middle range and minimizing the influence of those with near-certain treatment assignment (PS close to 0 or 1), overlap weighting creates a stable, bounded, and efficient weighted sample that mimics a randomized trial within the region of maximum overlap [1]. When comparator groups are initially very different (ie, low PS overlap), such as in real-world studies with strong baseline imbalances, overlap weighting offers clear advantages over IPTW [16]. Although both methods tend to yield similar results when baseline differences are modest [16], overlap weighting performs better when covariate overlap is poor or outliers are present. IPTW, by design, assigns the largest weights to individuals whose actual treatment contradicts what their PS would predict, such as DOAC users with very low PS or warfarin users with very high scores, while those whose treatment aligns with expectations receive moderate weights. Overlap weighting assigns weights proportional to the probability of receiving the *opposite* treatment [23]. That is, individuals who receive Treatment A are weighted by 1 minus their PS (1−PS), and those who receive Treatment B are weighted by their PS. This emphasizes individuals whose PS are close to 0.5, representing the region of greatest comparability (ie, weighted sample remains centered on the target population in clinical equipoise) [16,20,24]. Overlap weighting minimizes the undue influence of individuals with extreme scores (ie, outliers) by down-weighting individuals who are nearly always treated (PS $\approx$ 1) or never treated (PS $\approx$ 0) [15,16]. A key benefit of overlap weighting is that the weights are naturally bounded between 0 and 1, eliminating the extreme weights seen in IPTW. As a result, no single individual dominates the analysis. These properties yield a more stable weight distribution, increase precision (producing narrower confidence intervals), and eliminate the need for trimming or truncation [19,25]. By preserving all individuals in the dataset while emphasizing those most comparable, overlap weighting also maximizes the use of available data and enhances statistical efficiency [26].

Returning to the DOAC vs. warfarin example, if the PS reflects the probability of receiving a DOAC, DOAC users are weighted by 1 − PS, and warfarin users are weighted by PS. This means that individuals with extreme PS values, those almost certain to receive DOACs (PS near 1) or warfarin (PS near 0), are down-weighted, while those with PS near 0.5 contribute more, reflecting their greater comparability across treatment groups. For example, a patient with PS = 0.5, regardless of treatment, receives a weight of 0.5, while a DOAC user with PS = 0.98 or a warfarin user with PS = 0.02 would be down-weighted to 0.02. A DOAC user with PS = 0.05 receives a weight of 0.95, and a warfarin user with PS = 0.95 also receives 0.95, though such cases are relatively uncommon.

Overlap weighting also ensures exact balance on the mean of every measured covariate (eg, age, comorbidities) between treatment groups, meaning the weighted averages for all covariates are identical across groups [1,20,23]. Achieving this type of balance, where patient characteristics are similar across treatment arms, is essential for minimizing bias in observational studies [27].

Because overlap weighting targets a specific subpopulation, those in clinical equipoise, it is important to clearly characterize this group. Best practice includes reporting both unweighted and overlap-weighted baseline characteristics [23]. The unweighted table helps readers understand the original study sample and its generalizability, while the weighted table describes the characteristics of the target population for inference under the ATO estimand. This dual presentation improves transparency and ensures that findings are appropriately interpreted in clinical and policy contexts.

4.1. Examples from the literature that used overlap weighting

Table 1 highlights five recent studies across diverse clinical areas that implemented overlap weighting to address confounding, improve precision, and estimate treatment effects in subgroups with clinical equipoise. In all five studies, overlap weighting served as the primary analytic approach. However, none explicitly defined ATO as their

Table 1. Examples of how overlap weighting was applied in recent observational studies

First author, year [reference number]	Description of the use of overlap weighting in the study
Wei, 2023 [28]	Compared patients with gout and type 2 diabetes who initiated SGLT2 inhibitors to those starting GLP-1 receptor agonists or DPP-4 inhibitors. Overlap weighting was used to balance baseline characteristics across groups and reduce the influence of extreme propensity scores. Patients with a propensity score near 0.5 contribute most to the effect estimate, while those near 0 or 1 contribute least. This approach reduces the influence of extreme scores without excluding any patients, avoiding bias from outliers. This approach also allowed estimation of treatment effects in a subgroup with clinical equipoise and improved the validity of comparisons for recurrent gout flares and all-cause mortality outcomes.
Krishnapura, 2025 [29]	Compared pregnant individuals with opioid use disorder who received buprenorphine during pregnancy to those who did not. The author selected overlap weights over IPTW because they provide better covariate balance between treatment groups and are more robust to extreme propensity score values. Weighted logistic regression models were then used to estimate the association between buprenorphine treatment and key maternal and infant outcomes, such as severe maternal morbidity and preterm birth.
Quinlan, 2025 [30]	Used overlap weighting to compare bleeding and stroke outcomes among older patients with HIV and atrial fibrillation initiating apixaban, warfarin, or rivaroxaban. Overlap weighting was applied separately in three pairwise comparisons to balance 45 baseline covariates and reduce bias from treatment selection.
Robinson, 2025 [31]	Emulated a target trial comparing cyclophosphamide vs calcineurin inhibitors in children with nephrotic syndrome using overlap weighting to balance extensive baseline demographic and clinical covariates. The authors mention that this method was chosen over IPTW to improve precision, reduce bias from extreme scores, and estimate the treatment effect in a subgroup with clinical equipoise. Overlap weighting allowed for a more stable and generalizable comparison of relapse rates, hospitalization risk, and kidney function outcomes in a diverse pediatric cohort.
Cipolletta, 2025 [32]	Compared cardiovascular outcomes among gout patients initiating urate-lowering therapy with vs without colchicine for flare prophylaxis. Overlap weighting was used to balance baseline characteristics, such as age, kidney function, gout severity, and cardiovascular risk, between groups and minimize confounding from treatment selection. Overlap weighting was selected to ensure comparability between exposed and unexposed groups, and enabled robust estimation of the association between colchicine prophylaxis and cardiovascular events using real-world linked United Kingdom healthcare data.

IPTW, inverse probability of treatment weighting.

estimand. This underscores a common practice in observational research where the methodological advantages of overlap weighting are leveraged without a corresponding shift in the stated causal question. As discussed in this article, it would be best practice for future studies to clarify the target estimand when using overlap weighting.

4.2. *Limitations of overlap weighting*

While overlap weighting offers important statistical advantages, such as improved covariate balance, bounded weights, and greater precision, its interpretive and inferential limitations deserve careful attention. As discussed, overlap weighting targets the ATO, which represents individuals in clinical equipoise, typically those with PS near 0.5 [20]. Unlike the ATE, the ATO does not correspond to a population easily defined by clinical or demographic characteristics, making it difficult to identify in practice or generalize to broader populations. As a result, treatment effects estimated via overlap weighting may apply to a narrow or poorly characterized group, which may limit its clinical or policy relevance. Without clear reporting of both unweighted and overlap-weighted baseline characteristics, readers may mistakenly assume that the effect estimate generalizes to the full study population. However, the ATO remains highly relevant in many applied settings. It is particularly useful for comparative effectiveness questions involving patients whose treatment decisions vary due to uncertainty or lack of clear guidelines [33]. The ATO is also valuable in policy and health equity research, where interest often centers on treatment effects among individuals whose care decisions are not strongly constrained by clinical indications or structural barriers [34].

Second, as with all PS-based methods, overlap weighting cannot address unmeasured confounding [16], and its validity hinges on the correct specification of the PS model. Third, although overlap weighting balances the means of measured covariates between groups, this may not ensure complete adjustment for confounding on those variables [16].

5. Concluding remarks

The growing use of overlap weighting reflects its strength and value in addressing confounding in observational studies, especially when groups differ significantly at baseline. Its ability to provide stable, balanced, and interpretable estimates has made it increasingly popular in real-world research. However, these statistical advantages come

with important trade-offs. The ATO estimand targets a subgroup defined statistically rather than clinically, which can limit interpretability and generalizability. As such, overlap weighting should not be used solely to resolve estimation issues with other methods like, IPTW. Instead, its use must be grounded in a clearly defined research question centered on individuals in clinical equipoise; for example, when the goal is to determine which treatment is more effective among patients for whom no clear prescribing guidelines exist. As this method continues to gain momentum, it is essential for researchers to apply it thoughtfully, aligning estimand, method, and scientific question, and to report results transparently to support valid and interpretable causal inference.

Declaration of competing interest

There are no competing interests for any author.

Data availability

No data was used for the research described in the article.

References

[1] Li F, Thomas LE, Li F. Addressing extreme propensity scores via the overlap weights. Am J Epidemiol 2019;188:250−7.
[2] Dang LE, Gruber S, Lee H, Dahabreh IJ, Stuart EA, Williamson BD, et al. A causal roadmap for generating high-quality real-world evidence. J Clin Transl Sci 2023;7:e212.
[3] Zhu Y, Hubbard RA, Chubak J, Roy J, Mitra N. Core concepts in pharmacoepidemiology: violations of the positivity assumption in the causal analysis of observational data: consequences and statistical approaches. Pharmacoepidemiol Drug Saf 2021;30:1471−85.
[4] Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika 1983;70(1):41−55.
[5] Matsouaka RA, Zhou Y. Causal inference in the absence of positivity: the role of overlap weights. Biom J 2024;66(4):e2300156. https://doi.org/10.1002/bimj.202300156.
[6] 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.
[7] Chatton A, Le Borgne F, Leyrat C, Gillaizeau F, Rousseau C, Barbin L, et al. G-computation, propensity score-based methods, and targeted maximum likelihood estimator for causal inference with different covariates sets: a comparative simulation study. Sci Rep 2020;10:9219.
[8] Schuler MS, Rose S. Targeted maximum likelihood estimation for causal inference in observational studies. Am J Epidemiol 2017;185:65−73.
[9] Kreif N, Gruber S, Radice R, Grieve R, Sekhon JS. Evaluating treatment effectiveness under model misspecification: a comparison of targeted maximum likelihood estimation with bias-corrected matching. Stat Methods Med Res 2016;25:2315−36.
[10] Léger M, Chatton A, Le Borgne F, Pirracchio R, Lasocki S, Foucher Y. Causal inference in case of near-violation of positivity: comparison of methods. Biometrical J 2022;64:1389−403.
[11] Rubin DB. The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials. Stat Med 2007;26:20−36.
[12] Austin PC. The relative ability of different propensity score methods to balance measured covariates between treated and untreated subjects in observational studies. Med Decis Making 2009;29:661−77.
[13] Stuart EA. Matching methods for causal inference: a review and a look forward. Stat Sci 2010;25:1−21.
[14] Chesnaye NC, Stel VS, Tripepi G, Dekker FW, Fu EL, Zoccali C, Jager KJ. An introduction to inverse probability of treatment weighting in observational research. Clin Kidney 2022;J:14−20.
[15] Yoshida K, Hernández-Díaz S, Solomon DH, Jackson JW, Gagne JJ, Glynn RJ, Franklin JM. Matching weights to simultaneously compare three treatment groups comparison to three-way matching. Epidemiology 2017;28:387−95.
[16] Thomas LE, Li F, Pencina MJ. Overlap weighting: a propensity score method that mimics attributes of a randomized clinical trial. JAMA 2020;323:2417−8.
[17] Hashimoto Y, Yasunaga H. Theory and practice of propensity score analysis. Ann Clin Epidemiol 2022;4:101−9.
[18] Desai RJ, Franklin JM. Alternative approaches for confounding adjustment in observational studies using weighting based on the propensity score: a primer for practitioners. BMJ 2019;367:l5657.
[19] Lee BK, Lessler J, Stuart EA. Weight trimming and propensity score weighting. PLoS One 2011;6:e18174.
[20] Thomas L, Li F, Pencina M. Using propensity score methods to create target populations in observational clinical research. JAMA 2020;323:466−7.
[21] Marzec LN, Wang J, Shah ND, Chan PS, Ting HH, Gosch KL, et al. Influence of direct oral anticoagulants on rates of oral anticoagulation for atrial fibrillation. J Am Coll Cardiol 2017;69:2475−84.
[22] Tham D, Zhao L, Yu W, Kherani J, Kou R, Li A, et al. Safety and efficacy of direct oral anticoagulants in chronic kidney disease: a meta-analysis. Res Pract Thromb Haemost 2024;8(7):102584. https://doi.org/10.1016/j.rpth.2024.102584.
[23] Li F, Morgan KL, Zaslavsky AM. Balancing covariates via propensity score weighting. J Am Stat Assoc 2018;113(521):390−400.
[24] Rizk JG, Wenziger C, Tran D, Hashemi L, Moradi H, Streja E, Ahluwalia A. Angiotensin-converting enzyme inhibitor and angiotensin receptor blocker use associated with reduced mortality and other disease outcomes in US veterans with COVID-19. Drugs 2022;82:43−54.
[25] Cole SR, Hernán MA. Adjusted survival curves with inverse probability weights. Comput Methods Programs Biomed 2004;75:45−9.
[26] Mlcoch T, Hrnciarova T, Tuzil J, Zadak J, Marian M, Dolezal T. Propensity score weighting using overlap weights: a new method applied to regorafenib clinical data and a cost-effectiveness analysis. Value in Health 2019;22:1370−7.
[27] Zubizarreta JR. Stable weights that balance covariates for estimation with incomplete outcome data. J Am Stat Assoc 2015;110(511):910−22.
[28] Wei J, Choi HK, Dalbeth N, Li X, Li C, Zeng C, et al. Gout flares and mortality after sodium-glucose Cotransporter-2 inhibitor treatment for gout and type 2 diabetes. JAMA Netw Open 2023;6:e2330885.
[29] Krishnapura SR, McNeer E, Loch SF, Reese T, Dudley J, Phillippi JC, et al. Buprenorphine treatment in pregnancy and maternal-infant outcomes. JAMA Health Forum 2025;6(4.11):e251814. https://doi.org/10.1001/jamahealthforum.2025.1814.
[30] Quinlan CM, Avorn J, Kesselheim AS, Singer DE, Zhang Y, Cervone A, Lin KJ. Comparative bleeding risk in older patients with HIV and atrial fibrillation receiving oral anticoagulants. JAMA Intern Med 2025;185(4):431−9. https://doi.org/10.1001/jamainternmed.2024.8335.
[31] Robinson CH, Aman N, Banh T, Brooke J, Chanchlani R, Cuthbertson BH, et al. Comparative efficacy of nonsteroid immunosuppressive medications in childhood nephrotic syndrome. JAMA Pediatr 2025;179(3):321−31. https://doi.org/10.1001/jamapediatrics.2024.5286.

[32] Cipolletta E, Nakafero G, McCormick N, Yokose C, Avery AJ, Mamas MA, et al. Cardiovascular events in patients with gout initiating urate-lowering therapy with or without colchicine for flare prophylaxis: a retrospective new-user cohort study using linked primary care, hospitalisation, and mortality data. Lancet Rheumatol 2025;7(3):e197−207. https://doi.org/10.1016/S2665-9913(24)00248-0.

[33] Zajichek AM, Grunkemeier GL. Propensity scores used as overlap weights provide exact covariate balance. Eur J Cardiothorac Surg 2024;66(3):ezae318. https://doi.org/10.1093/ejcts/ezae318.

[34] Li FAN, Li FAN. Propensity score weighting for causal inference with multiple treatments. Ann Appl Stat 2019;13(4):2389−415.