Statistical primer: propensity scores used as overlap weights provide exact covariate balance

Alexander M Zajichek[1], Gary L Grunkemeier[2]

[1] Department of Quantitative Health Sciences, Cleveland Clinic, Cleveland, OH, USA

[2] Department of Cardiothoracic Surgery, Oregon Health & Science University, Portland, OR, USA

Corresponding Author:

Alexander M Zajichek

9500 Euclid Avenue, JJN3-01

Cleveland, OH 44196

United States of America

(216) 444-0489

zajicha2@ccf.org

Abstract

Overlap weighting (OW), using weights defined as the probability of receiving the opposite treatment, is a relatively new, alternative propensity score (PS)-based weighting technique used to adjust for confounding when estimating causal treatment effects. It has preferable properties compared to inverse probability of treatment weighting (IPTW) such as exact covariate balance, safeguards against extreme weights, and emphasis on medical equipoise, where treatment decisions are most uncertain. In this article we introduce the OW methodology, compare it to IPTW, and provide some strategies for assessing weighting impact, through an applied example of hospital mortality. When the PS distributions have large separation, IPTW has been shown to produce biased and less efficient estimates of the treatment effect, making OW a preferred method in such cases.

Keywords: statistics, propensity score, weighting, covariate balance, causal inference

Abbreviations and acronyms: propensity scoring, PS; inverse probability of treatment weighting, IPTW; overlap weighting, OW; odds ratio, OR; average treatment effect, ATE; average treatment effect in the overlap population, ATO; confidence interval, CI; coronary artery bypass grafting, CABG; standardized mean difference, SMD;

INTRODUCTION

Causal interpretation of treatment effects in observational studies are often stymied by the real-world consequences of such designs being loaded with confounding. While randomized trials are considered the gold standard for alleviating these issues, they are often not feasible for a plethora of reasons [1]. Thus, the field of causal inference emerged with a focus on developing methods to better isolate and estimate treatment effects in the observational setting with a causal interpretation [2]. One such approach is using propensity scores (PS), which is defined as the probability of a patient receiving treatment given their individual characteristics [3]. Benedetto et al. provided an intuitive overview on using these scores through four different strategies to address confounding: matching, stratification, covariate adjustment, and weighting [4]. We turn our attention to the last category.

In the aforementioned article, inverse probability of treatment weighting (IPTW) was the method described, where subjects are weighted by the inverse of the probability of receiving their observed treatment [5]. However, the goal of the weighting strategy can be stated more generally: to identify the appropriate weight to place on each subject in order to isolate differences in the outcome of interest attributable to the treatment by canceling out confounding information. It turns out that IPTW is just one way to do that. This strives to emulate a randomized trial, where the properties of randomization allow subjects to be weighted equally while maintaining unbiased estimates. Despite these advances, IPTW has been shown to yield biased and nonoptimal estimates of the treatment effect when PS distributions have extreme separation [6].

To address issues from extreme propensity scores, an alternative weighting system for achieving balance was proposed in 2018, called overlap weighting (OW) [7]. In the dichotomous treatment setting, this method uses weights defined as the probability of receiving the opposite treatment from what was observed. It has preferable properties to IPTW, such as exact covariate balance, safeguards against extreme weights, and emphasis on medical equipoise [8]. For example, a collection of PS 0.1, 0.05 and 0.01 for treated patients would yield case weights of 0.90, 0.95 and 0.99 with OW, but 10, 20 and 100 with IPTW, respectively. In a simulation study, OW yielded more efficient estimation and consistent confidence interval (CI) coverage while maintaining unbiased estimates of the treatment effect across a wide range of scenarios, suggestive of more overall robustness [9]. Specifically, it is recommended that OW be used in favor of IPTW in the case of extreme propensity score distributions for more accurate inference and targeted relevance for clinical decision making.

This article provides an introduction to OW, a comparison to IPTW, and suggested strategies for analyzing weighting impact, through an applied example in hospital mortality with highly separated propensity score distributions. Analysis was performed using the R statistical programming language (version 4.2.1) [10]. The code underlying this article is available on GitHub, at https://github.com/ClevelandClinicQHS/pubsource/tree/main/OW_Tutorial.

EXAMPLE DATASET

Benedetto et al have previously described the dataset from the Bristol Heart Institute (UK) that compared hospital mortality between on and off-pump treatment for patients undergoing isolated first-time coronary artery bypass grafting (CABG) [4]. The current dataset was simulated using the summary statistics from the study to inform distributional attributes in terms of sample size, number of covariates, and covariate distributions. However, more emphasis was placed on characterizing patients receiving off-pump as considerably sicker than those receiving on-pump (Table 1). Thus, although the observed mortality rate is higher in the

off-pump group (14%) compared to the on-pump group (4%), the true (hidden) effect reflects off-pump benefit such that the odds of hospital mortality is ~32% lower (Table 1).

ESTIMATING THE WEIGHTS

Propensity scores

The first step in the estimation process via weighting is to estimate the PS based on a set of patient characteristics. In general, characteristics means the set of attributes confounding the relationship between the treatment and the outcome, primarily specified through domain knowledge and subject matter expertise. Though other methods can be implemented [11], a common approach is to use logistic regression. As we'll see, there are advantages to doing so in the context of OW. To estimate the PS in the current dataset, we regress treatment on an additive logistic regression model with a logit link function containing the set of confounders in Table 1 as linear covariates, and obtain the fitted values transformed to the probability scale. This resulted in highly separated PS distributions between patients receiving off-pump and on-pump treatment (Figure 1).

Inverse probability of treatment weights

IPTW uses weights defined as the inverse of the probability of receiving the observed treatment. Using these weights in the outcome model produces an estimate of the average treatment effect (ATE) in the overall population. To compute them, the estimated PS is entered into the following formula: IPTW = T/PS + (1-T)/(1-PS), where T is 1 for a patient receiving off-pump and 0 for a patient receiving on-pump (see [4] for further details). In the current dataset, this yielded a heavily right-skewed distribution of weights (Figure 2a). This illustrates one of the potential problems with IPTW in the case of extreme propensity score distributions in that they are unbounded and can take on extreme values, leading to the estimated treatment effect possibly being dominated by very few subjects. Trimming, which amounts to setting boundaries

on the range of PS to include, can be used as a possible remedy for this at the expense of losing subjects from the sample [12]. Additionally, the chosen thresholds can be arbitrary, and when the treatment distribution is imbalanced, can disproportionately exclude more subjects in one group (when we did this in the current dataset using the rule-of-thumb PS inclusion range for symmetric trimming of 0.1 to 0.9 [9], we excluded 580 (31.5%) patients receiving off-pump and 410 (25.9%) patients receiving on-pump).

Overlap weights

OW uses weights defined as the probability of receiving the opposite treatment from what was observed. Using these weights in the outcome model produces an estimate of the average treatment effect in the overlap population (ATO) [9]. To compute them, the estimated PS is entered into the following formula: OW = T x (1-PS) + (1-T) x PS, where T is 1 for a patient receiving off-pump and 0 for a patient receiving on-pump. In the current dataset, the patients in the extremes got down-weighted and considerable weight was allocated to patients with the most amount of overlap in characteristics with the opposing treatment group, representing medical equipoise since their PS was near 0.5 (Figure 2b). While it is still right-skewed due to the extreme PS distributions, it is a flatter distribution of weights compared to IPTW and they are contained between 0 and 1. Additionally, the standardized mean difference (SMD) between all confounders in the OW-adjusted sample was exactly zero (Table 1).

ESTIMATING THE TREATMENT EFFECT

Treatment effects on the outcome are estimated using the derived weights. It is common practice to first normalize them within treatment groups, so they contribute equally in aggregate to the subsequent estimation [9]. This is done by dividing each weight by the sum within the respective group. Additionally, robust (sandwich) estimators or bootstrapping must be used for

accurately quantifying estimation uncertainty for valid inference [9]. We have implemented both for comparison (Table 1) but will focus on the results from the robust estimates.

Differences of inpatient mortality rates

The unadjusted difference of inpatient mortality rates between patients receiving off-pump and on-pump was 9.7% (95% CI: 7.9% to 11.5%). After weighting adjustment, the estimates were -5.8% (95% CI: -12.8% to 1.3%) and -1.4% (95% CI: -3.6% to 0.7%) for IPTW and OW, respectively, in favor of off-pump. As shown, using IPTW resulted in a considerably larger point estimate of off-pump benefit and a wider confidence interval than OW.

Odds ratios

The unadjusted odds of inpatient mortality were 3.78 (95% CI: 3.03 to 5.55) times higher for patients receiving off-pump compared to patients receiving on-pump. After weighting adjustment, the estimates were 0.57 (95% CI: 0.32 to 1.01) and 0.78 (95% CI: 0.55 to 1.11) for IPTW and OW, respectively, in favor of off-pump. Again, IPTW displayed a larger point estimate of off-pump benefit and a wider confidence interval than OW. Additionally, the IPTW was overly optimistic compared to the actual off-pump benefit set during simulation (an odds ratio (OR) of 0.78), of which OW was able to adequately recover. This exemplifies the bias that can be induced from IPTW under extreme propensity score distributions.

ASSESSMENT OF WEIGHTING IMPACT

Weights from differing methodologies are allocated to subpopulations differently when estimating the treatment effect. We illustrate some diagnostic strategies for contrasting the areas of focus of IPTW and OW in the current dataset to better understand their impact and implications. In general, these are useful exploratory tools to consider in any weighted analysis.

Predictors of the propensity score

A good starting point is to interrogate the shape and magnitude of the confounder effects in the PS model. When that is based on logistic regression, as it was here, a summary of OR may be sufficient. In the current dataset, age, COPD, NYHA, LMD, and IDDM were the top five characteristics with the strongest association to treatment allocation (Figure 3), suggestive of areas where the treatment groups are most imbalanced. When the PS model is more complex, containing interactions, non-linear terms, or a combination of both, directly plotting the estimated PS by treatment across the range of confounder levels and subgroups may be more informative.

Cumulative weight distribution

In an unweighted analysis, each patient contributes equally to the subsequent treatment effect estimation. Weight allocation changes this by shifting the aggregated contributions to a disproportionate share of patients in the original sample to better reflect the target population of interest: the *overall* population for IPTW, and the *overlap* population for OW [9]. The reweighted sample is referred to as a pseudo-population, which represents the hypothetical target population where confounding is balanced across treatment groups, and the only remaining difference is the treatment itself [13].

Since IPTW and OW seek to create different pseudo-populations, we evaluated the amount of cumulative model weight allocated by each method versus the share of unique patients to better understand how concentrated the treatment effect estimates were on a subset of the original sample (Figure 2c-d). The 25% of patients with the lowest IPTW and OW accounted for only 12-14% and 3-4% of the model weight in the treatment effect estimation, respectively. Because of the extreme PS distributions, OW put much less focus on patients with a clinically deterministic treatment assignment to elicit more influence from patients in the middle (the overlap population), where the tougher clinical decisions are to be made. IPTW

attempted to more uniformly assign weights as to yield treatment effect estimates representative of the entire population.

Weight shift within confounders

To further assess specific differences in population subgroups that each method focused on, we evaluated the difference in the amount of weight that shifted from the unweighted to the weighted analysis between IPTW and OW within strata of the confounders (Figure 4). Using age as an intuitive example (top left panel), OW allocated a ~3-5 percentage point lower model weight to the extremes of the age distribution (<65 years and 80+ years) than IPTW. Age was previously established as a top predictor of treatment allocation, solidifying that this middle area represents where the off-pump and on-pump groups most overlap in age, or where the most medical equipoise is, and hence who the estimated treatment effect is most targeted for.

DISCUSSION

OW is a relatively new weighting technique based on the PS that has preferable properties to IPTW such as: (1) exact covariate balance when logistic regression is used; (2) limiting of extreme weights by being bound between 0 and 1; and (3) emphasis on quantifying treatment impact where there is more medical equipoise. In this study, we have introduced the OW methodology in comparison to IPTW and demonstrated strategies for analyzing the weights to understand their impact and implications on estimation of treatment effects. Under heavily separated PS distributions between the treatment groups, IPTW was shown to produce biased treatment effect estimates [6]—and we have observed evidence of that in the current study as well.

It is recommended that OW is used in favor of IPTW in the context of extreme PS distributions for more robust statistical inference and relevance in clinical decision making. The IPTW helps estimate the treatment effect averaged over the entire population, but if most of the

sample consists of cases in which the treatment is clinically deterministic (because they have extreme PS), the resulting treatment effect estimate will be less representative of those patients in the middle who have the tougher treatment decision to make, and likely for whom the estimate is most practically useful for. In contrast, this is precisely where the OW shifts its focus to, those patients whose choice in treatment is most uncertain. It does this in a smooth and proportionate way, preventing individual patients from taking on arbitrarily large influence without the need to discard samples through trimming or other ad-hoc procedures, since it is, by definition, bound between 0 and 1. Additionally, the properties of OW ensure that when logistic regression is used to estimate the PS, the weighted-mean difference between treatment groups for all covariates included in the PS model will be exactly zero [9]. With IPTW, the differences may be reduced, but by an arbitrary amount depending on the modeling context at hand, forcing the practitioner to accept it as being satisfactory without the ability to untangle the impact of the remaining differences on subsequent estimation.

However, OW does not go without limitation. First, no statistical method is perfect, so it still may be subject to its own biases and statistical inefficiencies depending on the modeling context. Second, it is unintuitive to understand exactly what an “overlap” population means practically. Conceptually it is clear, but in terms of precise statistical interpretation, it remains rather ambiguous where the line is drawn, making it difficult to understand which patients an estimated treatment effect truly applies to when attempting to use it for practical, day to day clinical decision making. Finally, ad-hoc procedures such as trimming can be used to adequately correct IPTW based treatment effect estimates, so although OW does bode well as a generally more robust method in the context of extreme PS distributions, it’s not a guarantee that it will be more performant. The practitioner must evaluate the context of the modeling problem at hand to choose the methodology best suited to answer the research question regardless of what the PS distributions look like.

CONFLICT OF INTEREST

None declared.

DATA AVAILABILITY STATEMENT

The data underlying this article are available on GitHub at https://github.com/ClevelandClinicQHS/pubsource/tree/main/OW_Tutorial.

FUNDING STATEMENT/FINANCIAL DISCLOSURE

This work is supported by the Department of Quantitative Health Sciences, Cleveland Clinic, Cleveland, OH, USA.

REFERENCES

1. Rajadhyaksha V. Conducting feasibilities in clinical trials: an investment to ensure a good study. Perspect Clin Res. 2010;1(3):106-109.
2. Olier I, Zhan Y, Liang X, Volovici V. Causal inference and observational data. BMC Med Res Methodol. 2023;23(1):227. Published 2023 Oct 11. doi:10.1186/s12874-023-02058-5
3. Rosenbaum PR, Rubin DB. The Central Role of the Propensity Score in Observational Studies for Causal Effects. Biometrika. 1983; 70 (1): 41–55.
4. Benedetto U, Head SJ, Angelini GD, Blackstone EH. Statistical primer: propensity score matching and its alternatives. Eur J Cardiothorac Surg. 2018;53(6):1112-1117. doi:10.1093/ejcts/ezy167
5. Austin PC, Stuart EA. Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. Stat Med. 2015;34(28):3661-3679. doi:10.1002/sim.6607
6. Lee BK, Lessler J, Stuart EA (2011) Weight Trimming and Propensity Score Weighting. PLoS ONE 6(3): e18174. https://doi.org/10.1371/journal.pone.0018174
7. Fan L, Morgan KL, Zaslavsky AM. Balancing Covariates via Propensity Score Weighting, Journal of the American Statistical Association. 2018; 113:521, 390-400, DOI: 10.1080/01621459.2016.1260466
8. Braakhekke M, Mol F, Mastenbroek S, Mol BW, van der Veen F. Equipoise and the RCT. Hum Reprod. 2017;32(2):257-260. doi:10.1093/humrep/dew286
9. Li F, Thomas LE, Li F. Addressing Extreme Propensity Scores via the Overlap Weights [published correction appears in Am J Epidemiol. 2021 Jan 4;190(1):189-190]. Am J Epidemiol. 2019;188(1):250-257. doi:10.1093/aje/kwy201
10. R Core Team (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

11. Lee BK, Lessler J, Stuart EA. Improving propensity score weighting using machine learning. Stat Med. 2010;29(3):337-346. doi:10.1002/sim.3782

12. Imbens G, Crump R, Hotz JV, Mitnik O. Dealing with Limited Overlap in Estimation of Average Treatment Effects. Biometrika. 2009;96 (1) :187-199.

13. 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

Table 1: Preoperative confounder distributions, observed in-hospital mortality rates, and estimated treatment effects in the original, IPTW-adjusted, and OW-adjusted samples from the simulated dataset.

	Original sample			IPTW-adjusted			OW-adjusted		
	Off-pump	On-pump	SMD	Off-pump	On-pump	SMD	Off-pump	On-pump	SMD
Set of confounders	1841	1583		1841	1583		1841	1583	
Age (years), mean (SD)	76 (6)	70 (6)	0.878	74 (6)	74 (6)	0.027	73 (6)	73 (6)	0
Female, n (%)	592 (32.2)	430 (27.2)	0.109	498 (27.1)	471 (29.7)	0.059	531 (28.9)	457 (28.9)	0
NYHA III/IV, n (%)	931 (50.6)	505 (31.9)	0.386	739 (40.2)	672 (42.5)	0.047	746 (40.5)	642 (40.5)	0
MI within 30 days, n (%)	888 (48.2)	471 (29.8)	0.386	688 (37.4)	584 (36.9)	0.009	703 (38.2)	605 (38.2)	0
Prior PCI, n (%)	123 (6.7)	54 (3.4)	0.150	96 (5.2)	69 (4.4)	0.040	92 (5.0)	79 (5.0)	0
IDDM, n (%)	234 (12.7)	96 (6.1)	0.229	183 (9.9)	138 (8.7)	0.042	148 (8.0)	127 (8.0)	0
Smoking, n (%)	208 (11.3)	114 (7.2)	0.142	197 (10.7)	159 (!0.0)	0.023	167 (9.1)	144 (9.1)	0
Creatine > 200 mmol/l, n (%)	128 (7.0)	58 (3.7)	0.147	130 (7.1)	123 (7.8)	0.026	92 (5.0)	79 (5.0)	0
COPD, n (%)	361 (19.6)	158 (10.0)	0.274	263 (14.3)	226 (14.3)	0.001	236 (12.8)	203 (12.8)	0
CVA, n (%)	158 (8.6)	84 (5.3)	0.129	133 (7.2)	100 (6.3)	0.036	130 (7.1)	112 (7.1)	0
PVD, n (%)	566 (30.7)	305 (19.3)	0.267	511 (27.8)	450 (28.4)	0.014	447 (24.3)	384 (24.3)	0
NVD, n (%) [a]			0.346			0.061			0
1	31 (1.7)	119 (7.5)		86 (4.7)	82 (5.2)		53 (2.9)	60 (3.8)	
2	321 (17.4)	388 (24.5)		408 (22.2)	282 (17.8)		397 (21.6)	314 (19.8)	
3	1489 (80.9)	1076 (68.0)		1347 (73.2)	1220 (77.0)		1390 (75.5)	1209 (76.4)	
LMD, n (%)	743 (40.4)	371 (23.4)	0.369	581 (31.6)	534 (33.7)	0.046	553 (30.0)	475 (30.0)	0
LVEF <30%, n (%)	369 (20.0)	207 (13.1)	0.188	317 (17.2)	336 (21.2)	0.101	319 (17.3)	275 (17.3)	0
Cardiogenic shock, n (%)	47 (2.6)	23 (1.5)	0.079	36 (2.0)	22 (1.4)	0.045	32 (1.7)	28 (1.7)	0
Preoperative IABP, n (%)	72 (3.9)	51 (3.2)	0.037	75 (4.1)	44 (2.8)	0.072	63 (3.4)	54 (3.4)	0
Emergency, n (%)	135 (7.3)	64 (4.0)	0.142	106 (5.8)	121 (7.7)	0.075	103 (5.6)	89 (5.6)	0
BMI, mean (SD)	27 (5)	27 (5)	0.006	27 (5)	27 (5)	0.089	27 (5)	27 (5)	0
YOP, mean (SD)	2005 (5)	2006 (5)	0.044	2005 (5)	2005 (5)	0.015	2006 (5)	2006 (5)	0
Performed by trainee, n (%)	500 (27.2)	374 (23.6)	0.081	482 (26.2)	377 (23.8)	0.054	474 (25.7)	407 (25.7)	0
Estimation of treatment effect									
In-hospital mortality, n (%)	253 (13.7)	64 (4.0)	0.346	164 (8.9)	232 (14.7)	0.179	103 (5.6)	111 (7.0)	0.059
Robust SE, Diff % (95% CI)	9.7 (7.9, 11.5)			-5.8 (-12.8, 1.3)			-1.4 (-3.6, 0.7)		
[b] Bootstrap SE, Diff % (95% CI)	9.7 (8.2, 12.0)			-5.8 (13.2, 0.9)			-1.4 (-3.5, 0.4)		
Robust SE, OR (95% CI)	3.78 (2.85, 5.02)			0.57 (0.32, 1.01)			0.78 (0.55, 1.11)		
[b] Bootstrap SE, OR (95% CI)	3.78 (3.03, 5.55)			0.57 (0.33, 1.12)			0.78 (0.55, 1.09)		
True treatment effect, OR	0.78			0.78			0.78		

IPTW: Inverse probability of treatment weighting; OW: Overlap weighting; SMD: standardized mean difference; OR, Odds ratio; Diff %, 100 X difference in proportions

[a] Treated as numeric for propensity score estimation and SMD calculation

[b] Based on 100 bootstrap resamples

Figure 1: Mirrored histogram showing the PS distribution by treatment group. PS: propensity score.

Figure 2: Top: Mirrored histograms showing the weighting distributions by treatment group using IPTW (a) and OW (b). Bottom: Cumulative share of model weight contribution (x-axis) versus individual patients (y-axis) by treatment group using IPTW (c) and OW (d). IPTW: inverse probability of treatment weighting; OW: overlap weighting.

Figure 3: Forest plot showing the OR and 95% CI for each confounder in the PS model for off-pump versus on-pump treatment allocation, ordered by effect magnitude. OR: odds ratio; CI, confidence interval; PS, propensity score.

Figure 4: Percentage point difference between OW and IPTW in the change of the share of model weight attributed compared to the unweighted sample stratified by levels of the top six confounding factors in the PS model: age (years; upper left), COPD (upper middle), IDDM (upper right), LMD (lower left), MI (lower middle), and NYHA (lower right) within each treatment group. OW: overlap weighting; IPTW: inverse probability of treatment weighting; PS: propensity score

Central Image: A comparison of the (absolute) SMD of confounders in the PS model between treatment groups for the unadjusted, IPTW adjusted, and OW adjusted samples. SMD: standardized mean difference; PS: propensity score; IPTW: inverse probability of treatment weighting; OW: overlap weighting

Figure 1

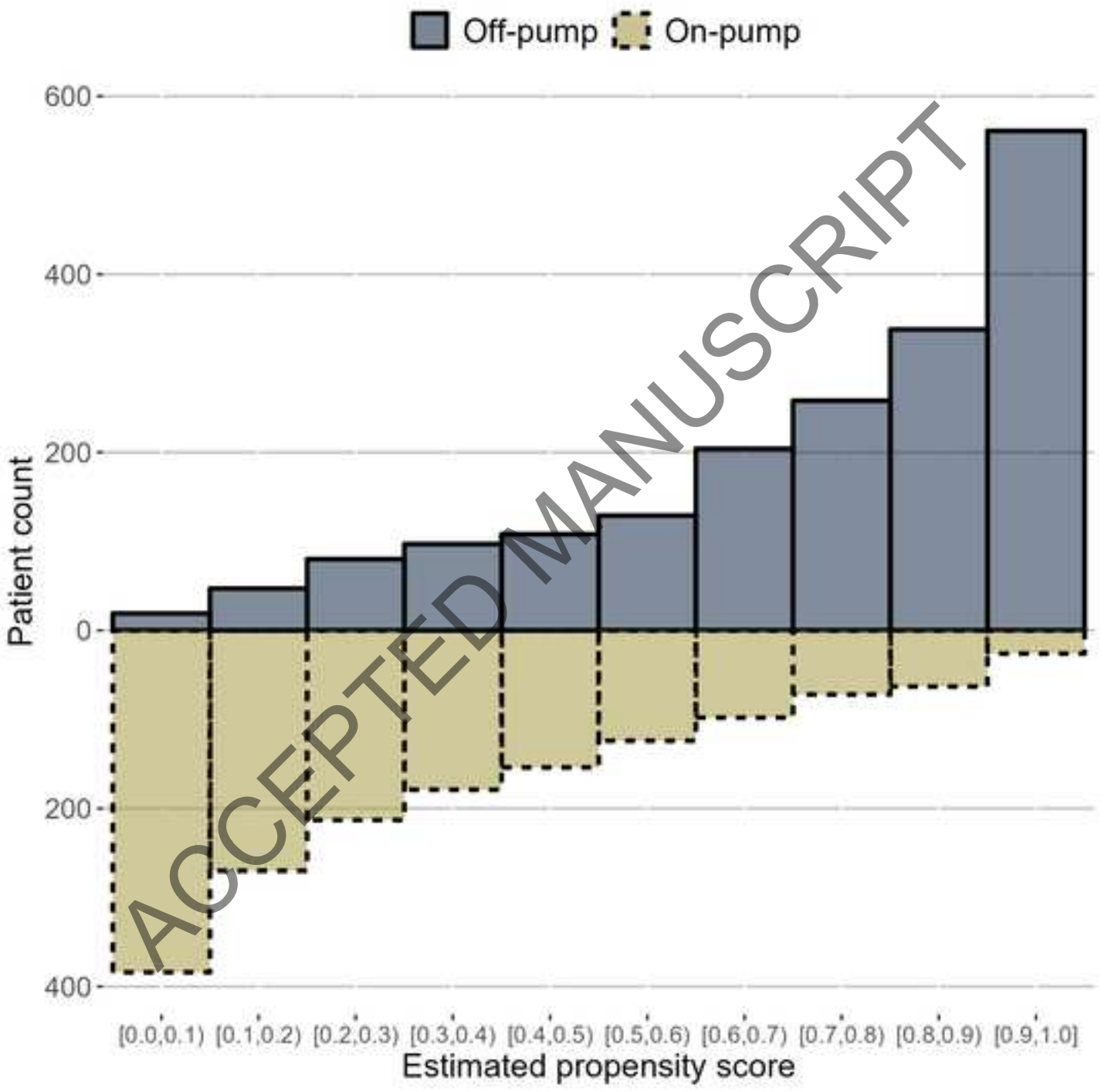


Figure 2

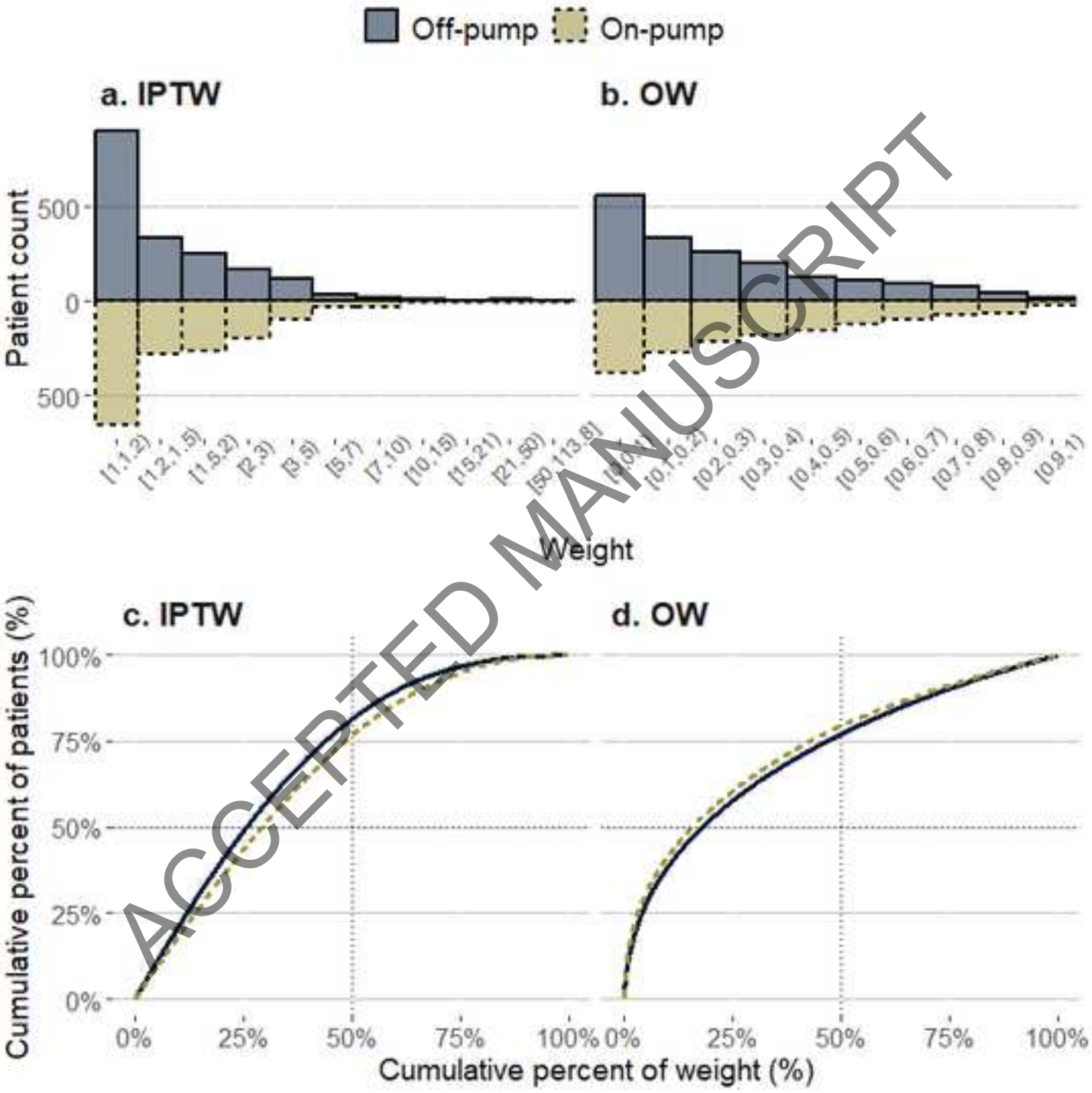


Figure 3

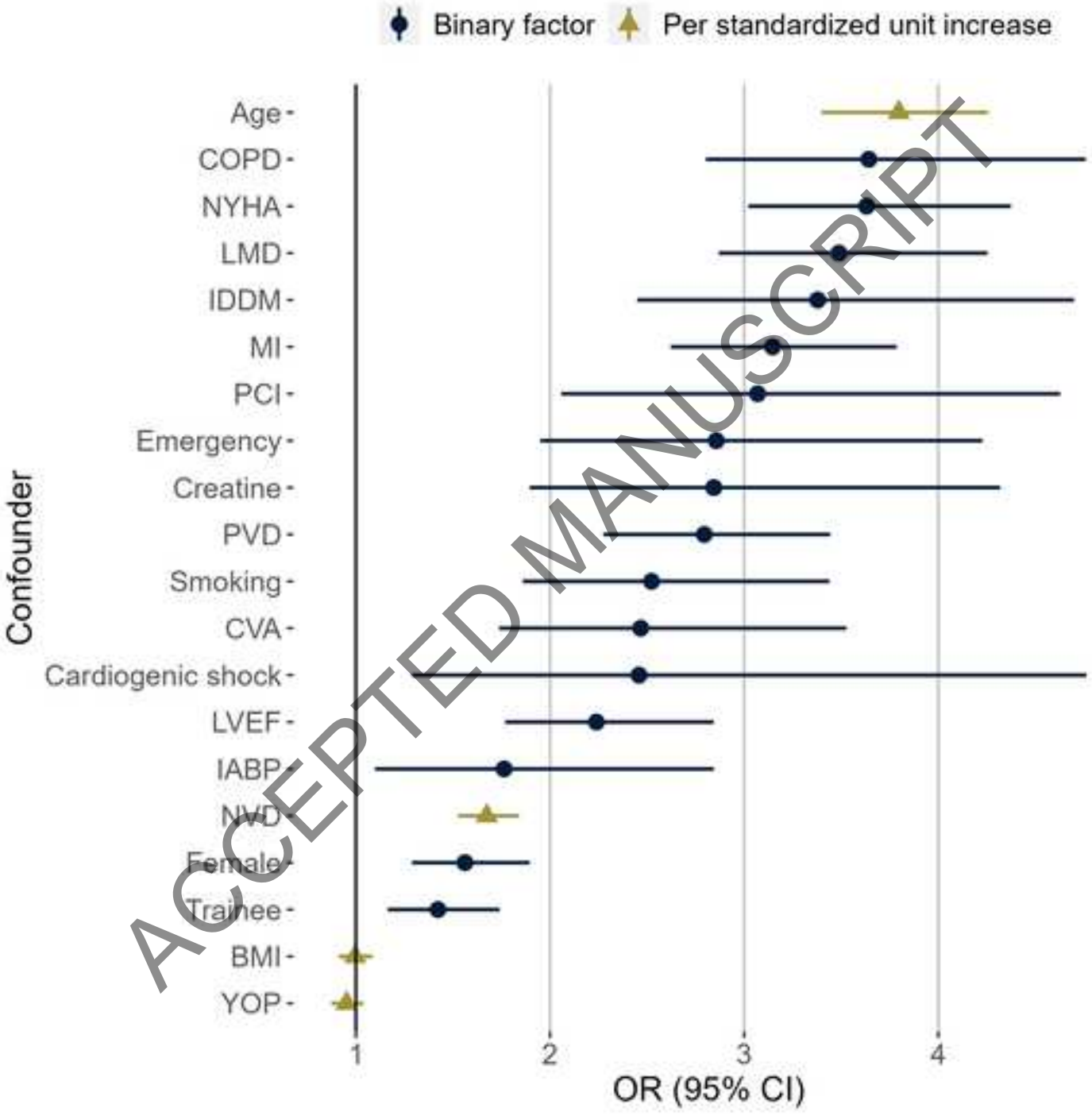
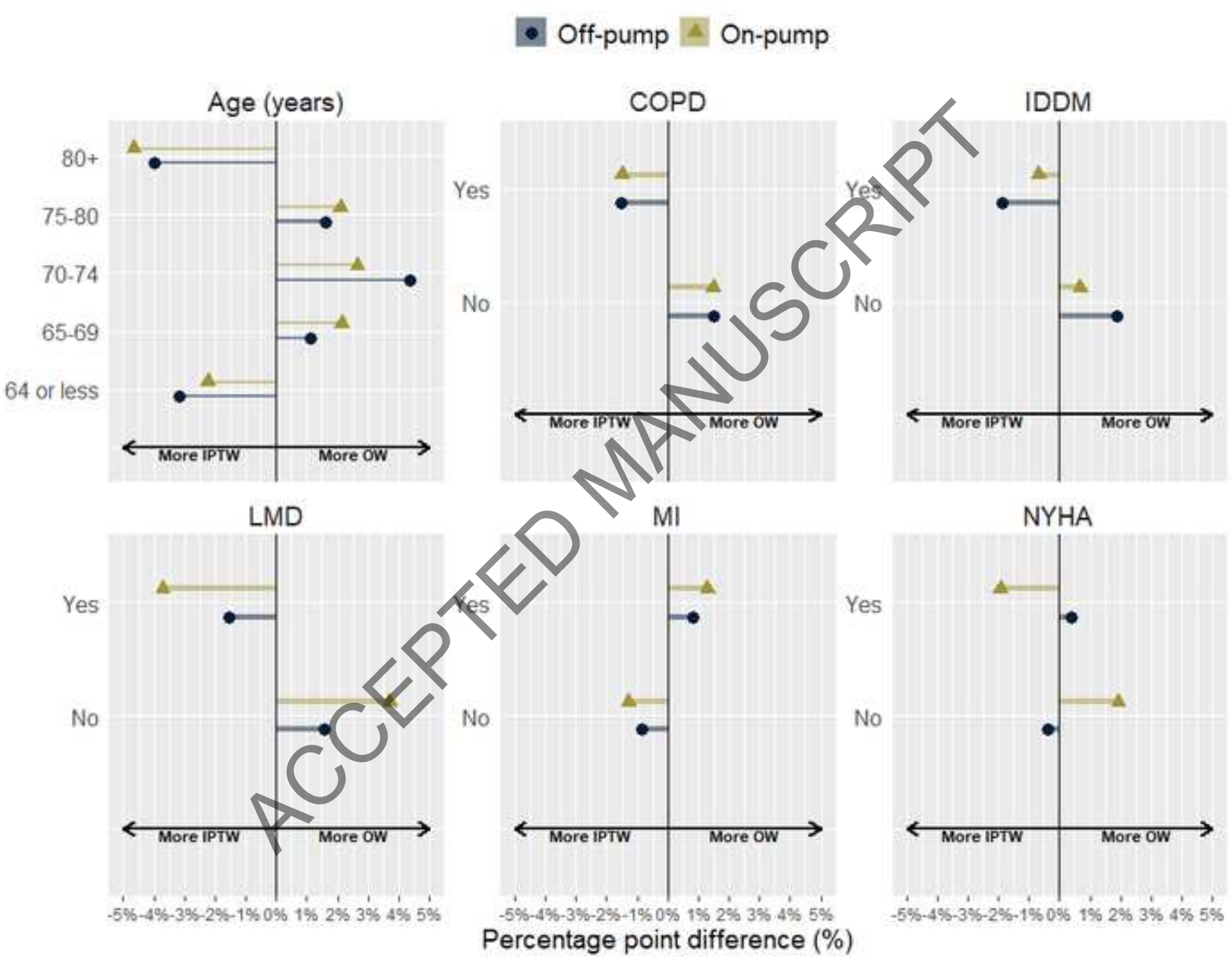


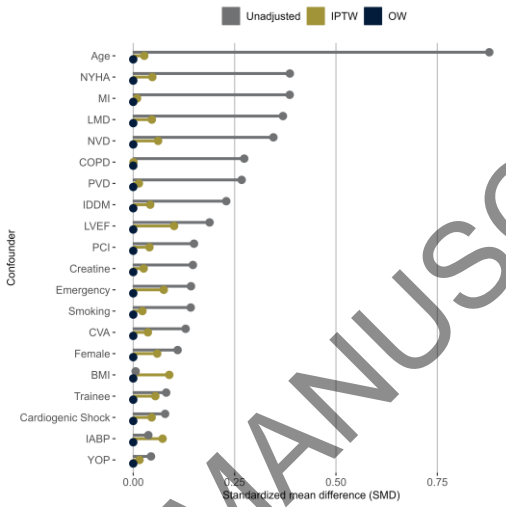
Figure 4



Statistical Primer: propensity scores used as overlap weights provide exact covariate balance

Summary

OW is an alternative PS-based weighting technique to adjust for confounding when estimating treatment effects. It has been shown to be more robust than IPTW in the context of extreme PS distributions, with preferable properties such as exact covariate balance, safeguards against extreme weights, and emphasis on medical equipoise, where treatment decisions are most uncertain.



Legend: OW, overlap weighting; PS, propensity score; IPTW, inverse probabilty of treatment weighting