Received: 3 February 2021 | Revised: 18 May 2021 | Accepted: 9 August 2021

DOI: 10.1002/sim.9176

RESEARCH ARTICLE

Statistics in Medicine WILEY

Informing power and sample size calculations when using inverse probability of treatment weighting using the propensity score

Peter C. Austin[1,2,3]

[1] ICES, Toronto, Ontario, Canada

[2] Institute of Health Management, Policy and Evaluation, University of Toronto, Ontario, Canada

[3] Sunnybrook Research Institute, Toronto, Ontario, Canada

Correspondence
Peter C. Austin, ICES, G106, 2075 Bayview Avenue, Toronto, ON M4N 3M5, Canada.
Email: peter.austin@ices.on.ca

Funding information
Canadian Institutes of Health Research, Grant/Award Number: PJT 166161; Heart and Stroke Foundation of Canada; Ontario Ministry of Health and Long-Term Care

Propensity score weighting is increasingly being used in observational studies to estimate the effects of treatments. The use of such weights induces a within-person homogeneity in outcomes that must be accounted for when estimating the variance of the estimated treatment effect. Knowledge of the variance inflation factor (VIF), which describes the extent to which the effective sample size has been reduced by weighting, allows for conducting sample size and power calculations for observational studies that use propensity score weighting. However, estimation of the VIF requires knowledge of the weights, which are only known once the study has been conducted. We describe methods to estimate the VIF based on two characteristics of the observational study: the anticipated prevalence of treatment and the anticipated c-statistic of the propensity score model. We considered five different sets of weights: those for estimating the average treatment effect (ATE), the average treated effect in the treated (ATT), and three recently described sets of weights: overlap weights, matching weights, and entropy weights. The VIF was substantially smaller for the latter three sets of weights than for the first two sets of weights. Once the VIF has been estimated during the design phase of the study, sample size and power calculations can be done using calculations appropriate for a randomized controlled trial with similar prevalence of treatment and similar outcome variable, and then multiplying the requisite sample size by the estimated VIF. Implementation of these methods allows for improving the design and reporting of observational studies that use propensity score weighting.

KEYWORDS

inverse probability of treatment weighting, power, propensity score, sample size, study design

1 | INTRODUCTION

Dorn suggested that when conducting an observational study one ask “how would the study be conducted if it were possible to do it by controlled experimentation?[1] Rubin suggests that this question defines the objective of an observational study[2] (p. 16). Similarly, Hernan and Robins define the concept of the “target trial” as the randomized controlled experiment that one is trying to emulate using observational data.[3] Sample size and power calculations are an integral component of the design of controlled experiments (eg, randomized controlled trials [RCTs]). In RCTs, sample size calculations are conducted during the design phase of the study, to ensure that the sample size provides adequate power to detect clinically meaningful effects. Similarly, power calculations serve an important role in observational studies, as they provide reassurance that the study sample is sufficiently large to provide adequate power to detect as statistically significant a clinically meaningful effect size. Researchers using observational data are often tempted to provide post-hoc power calculations, despite this practice being criticized.[4]

In observational studies, unlike in RCTs, treatment selection is frequently confounded with subject characteristics, so that treated subjects often differ systematically from control subjects. Consequently, statistical methods must be used to reduce the effects of confounding. A popular way to do this is using methods based on the propensity score. The propensity score is the probability of treatment selection conditional on the subject's measured baseline covariates.[5,6] Weighting using the propensity score is one way of using the propensity score to estimate the effects of treatment.[7] Using propensity score-based weights results in a weighted sample in which the distribution of measured baseline covariates is similar between treated and control subjects. The use of weighting-based methods is becoming increasingly popular in the medical and epidemiological literature.

When using propensity score-based weighting, estimation of the point estimate of the treatment effect can mimic the analysis that would be done in an RCT with a similar outcome variable. However, estimation of the variance of the treatment effect must account for the within-subject homogeneity in outcomes that is induced by weighting.[8,9] Zhou and colleagues described a variance inflation factor (VIF) that describes the inflation in the sample size that has occurred due to the incorporation of weights.[10] The concept of the VIF, or equivalently the design effect (DE), arises from sample survey design and cluster randomized trials.[11-13] It represents the increase in the variance of a test statistic due to using a sample other than a simple random sample. A drawback to using the VIF for the prospective planning of studies and for the prospective evaluation of statistical power is that calculation of the VIF requires knowledge of the subject-specific weights, which are only known once the analyses have been conducted.

The objective of the current study is to examine the relationship between study characteristics and the VIF (or DE). We examine the relationship between the c-statistic (equivalent to the area under the receiver operating characteristic curve) of the propensity score model, prevalence of treatment and the VIF (or DE). This will allow study investigators to conduct power calculations before the weights have been estimated. The paper is structured as follows. In Section 2, we provide background on propensity score weighting and different sets of propensity score weights. In Section 3, we describe a set of Monte Carlo simulations to describe the relationship between the c-statistic of the propensity score model, prevalence of treatment and the VIF (or DE). In Section 4, we report the results of these simulations. In Section 5, we provide a brief case study illustrating the application of these results. Finally, in Section 6, we summarize our findings and place them in the context of the existing literature.

2 | PROPENSITY SCORE-BASED WEIGHTS AND VIFS

Let $\mathbf{X}$ denote a vector of observed baseline covariates and let Z denote a binary treatment variable (Z = 0 for control vs Z = 1 for active treatment). The propensity score is defined as $e(\mathbf{X}) = \Pr(Z = 1|\mathbf{X})$. In practice, the propensity score is often estimated using a logistic regression model in which treatment selection is regressed on measured baseline covariates.

Rubin developed a framework for causal inference that is referred to as Rubin's Causal Model.[14] Given an outcome Y, we define two potential outcomes: Y(0) and Y(1), which denote the outcomes under control and treatment, respectively, if the subject received either the control or treatment under identical circumstances. The effect of treatment is defined as Y(1) – Y(0). The average treatment effect (ATE) is defined as ATE = E[Y(1) − Y(0)], while the average treatment effect in the treated (ATT) is defined as ATT = E[Y(1) − Y(0)|Z = 1].[15] The ATE is the average effect of treatment in the entire population, while the ATT is the average effect of treatment in those subjects in the population who were treated.

In an RCT the population to which the estimated treatment effect pertains is defined by the study inclusion and exclusion criteria. Thus, in an RCT, the ATE is a well-defined estimand because the overall population is well defined due to the use of inclusion and exclusion criteria. Similarly, a well-designed observational study attempts to mimic the target trial, and, as such, will employ inclusion and exclusion criteria. Thus, in a well-designed observational study, the target population for the ATE is also well-defined. The target population of the ATT is defined as the subset of treated subjects in the ATE target population. While this population cannot be defined in terms of inclusion and exclusion criteria, it can be defined in terms of inclusion and exclusion criteria and receipt of treatment. However, the treated population may vary across jurisdictions, if there are regional characteristics (eg, insurance policy) that influence treatment assignment.

Different propensity-score based weights have been defined that allow one to balance measured baseline covariates between treated and control subjects. The original weights were defined as $w(\mathbf{X}) = (Z/e(\mathbf{X})) + (1 - Z/1 - e(\mathbf{X}))$.[7] We refer to these weights as IPTW-ATE weights (where IPTW stands for inverse probability of treatment weighting), as they permit estimation of the ATE. A second set of weights, $w(\mathbf{X}) = Z + (e(\mathbf{X})/1 - e(\mathbf{X}))$, permit estimation of the ATT.[15] We refer to these weights as IPTW-ATT weights. Use of IPTW-ATE weights implies that the ATE is the target estimands, while use of IPTW-ATT weights implies that the ATT is the target estimand.

Recently, alternative sets of weights have been proposed. These include overlap weights (OW), matching weights (MW), and entropy weights (EW).[10,16] These are defined as: $\text{IPTW-OW} = Z(1 - e(\mathbf{X})) + (1 - Z)(e(\mathbf{X}))$, $\text{IPTW-MW} = Z\min(e(\mathbf{X}), 1 - e(\mathbf{X}))/e(\mathbf{X}) + (1 - Z)\min(e(\mathbf{X}), 1 - e(\mathbf{X}))/1 - e(\mathbf{X})$, and $\text{IPTW-EW} = Z(-e(\mathbf{X})\log(e(\mathbf{X})) - (1 - e(\mathbf{X}))\log(1 - e(\mathbf{X})))/e(\mathbf{X}) + (1 - Z)(-e(\mathbf{X})\log(e(\mathbf{X})) - (1 - e(\mathbf{X}))\log(1 - e(\mathbf{X})))/(1 - e(\mathbf{X}))$, respectively. Use of these alternative weights targets inference at the subpopulation for whom there is the greatest clinical equipoise about treatment.[10] These weights have been shown to have desirable statistical properties.[10] However, it can be difficult to formally describe the population to whom the estimand applies, whereas this can easily be done for the ATE. While the target population for the ATE can be formally defined using the study inclusion and exclusion criteria, this cannot be done for the target population when using OWs, MWs, or EWs. While the target population cannot be formally defined, the characteristics of the target population can be described by reporting a table in which the weighted means and prevalences of baseline variables are reported.[17] In an RCT, an examination of such a table allows one to describe the study sample, but such a table does not allow one to define the population to which the estimated treatment effect pertains.

The use of weighting induces a within-subject correlation in outcomes as subjects can have weights that are unequal to one another.[8,9] This within-subject homogeneity in outcomes must be accounted for when estimating the variance of the estimated treatment effect. Thus, while the statistical analyses conducted in the weighted sample can reflect those conducted in a sample that is free from selection-bias (eg, data from an RCT), variance estimation must account for the within-subject homogeneity in outcomes induced by weighting.

Zhou et al describe a VIF quantifying the inflation in the sample size due to weighting: $\text{VIF} = (N_1(N - N_1)/N)\left[\left(\sum_{i=1}^{N} Z_i w(x_i)^2 \Big/ \left(\sum_{i=1}^{N} Z_i w(x_i)\right)^2\right) + \left(\sum_{i=1}^{N}(1 - Z_i)w(x_i)^2 \Big/ \left(\sum_{i=1}^{N}(1 - Z_i)w(x_i)\right)^2\right)\right]$, where N denotes the sample size and $N_1 = \sum_{i=1}^{N} Z_i$ denotes the number of treated subjects.[10] Computation of the VIF requires knowledge of the weights, which are only known after the analysis has been conducted, which complicates computations to determine statistical power or the necessary sample size prior to the sample being collected and the weights determined. The VIF (or DE) is related to the concept of the "effective sample" size discussed by Golinelli et al, which denotes the reduction in sample size due to the use of matching or the incorporation of weights.[18]

3 | EFFECTS OF STUDY CHARACTERISTICS ON THE VIF: METHODS

We conducted a set of Monte Carlo simulations to examine the effect of the proportion of subjects who were treated and the c-statistic of the propensity score model on the VIF for a given set of weights. The c-statistic is a measure of discrimination that assesses the ability of the estimated propensity score model to discriminate between treated and untreated subjects. It ranges from 0.5 to 1, which a value of 0.5 denoting discrimination no different than random choice, and 1 denoting perfect discrimination. The c-statistic can be computed by considering all possible pairs consisting of treated and untreated subjects and determining the proportion of these pairs in which the treated subject had a higher propensity score than the untreated subject. Perfect discrimination would occur if all treated subjects had a higher propensity score than all untreated subjects.

We conducted four sets of simulations to examine the effect of study characteristics on the VIF. In the primary set of simulations, we assumed that the single predictor variable was normally distributed. We then conducted three sets of simulations as sensitivity analyses in which the single predictor variable followed either a Beta distribution, a Chi-squared distribution, or a log-normal distribution.

3.1 | Primary simulations: normally distribution predictor variable

For a given prevalence of treatment and c-statistic of the propensity score model, we simulated a continuous baseline covariate for each of 1 000 000 subjects: $x_i \sim N(0, 1)$. This covariate can be thought of as either a single covariate (eg, age) or as a linear predictor summarizing information on a set of covariates.

We specified a logistic model for treatment status as $\text{logit}(\Pr(Z = 1)) = \alpha_0 + \alpha_1 X$. Under the assumption that $\text{var}(X|Z = 0) = \text{var}(X|Z = 1) = \sigma^2$ the c-statistic of the logistic model is equal to $\Phi(\sigma\alpha_1/\sqrt{2})$, where Φ denotes the normal cumulative distribution function.[19] Thus, the c-statistic of the propensity score model is entirely determined by the log-odds ratio for the continuous variable and the variance of the continuous variable in the treated and control subjects (under the assumption that this variance is equal in treated and control subjects). Using the above relationship, we determined the value of α_1 that would result in a propensity score model with the desired c-statistic. We then used a bisection approach to determine the value of α_0 that would result in the desired prevalence of treatment in the overall sample. Once the values of α_0 and α_1 had been determined, we generated a binary treatment status for each subject using a Bernoulli distribution with a subject-specific probability determined from the propensity score model. This completed the construction of the simulated dataset of 1 000 000 subjects. The dataset consisted of two variables: the continuous predictor variable and the binary treatment status variable.

Once the simulated dataset had been constructed, we regressed the binary treatment status variable on the continuous predictor using a logistic regression model. We then determined the empirical c-statistic of the fitted propensity score model. This may differ slightly from the theoretical value because the assumption that $\text{var}(X|Z = 0) = \text{var}(X|Z = 1) = \sigma^2$ may be violated in the simulated sample. Using the estimated propensity score we computed the five different sets of weights described above (IPTW-ATE, IPTW-ATT, IPTW-OW, IPTW-MW, and IPTW-EW) and the corresponding VIFs.

We allowed two factors to vary in our simulations: the c-statistic of the propensity score model and the prevalence of treatment. We allowed the former to vary from 0.55 to 0.95 in increments of 0.025, while the latter varied from 0.10 to 0.90 in increments of 0.10. We thus considered 153 different scenarios (17 c-statistics × 9 prevalences of treatment).

Figure 1 reports the distribution of the propensity score separately in treated and control subjects when the target c-statistic ranged from 0.55 to 0.90 in increments of 0.05 and the prevalence of treatment was 0.5. There is one panel for

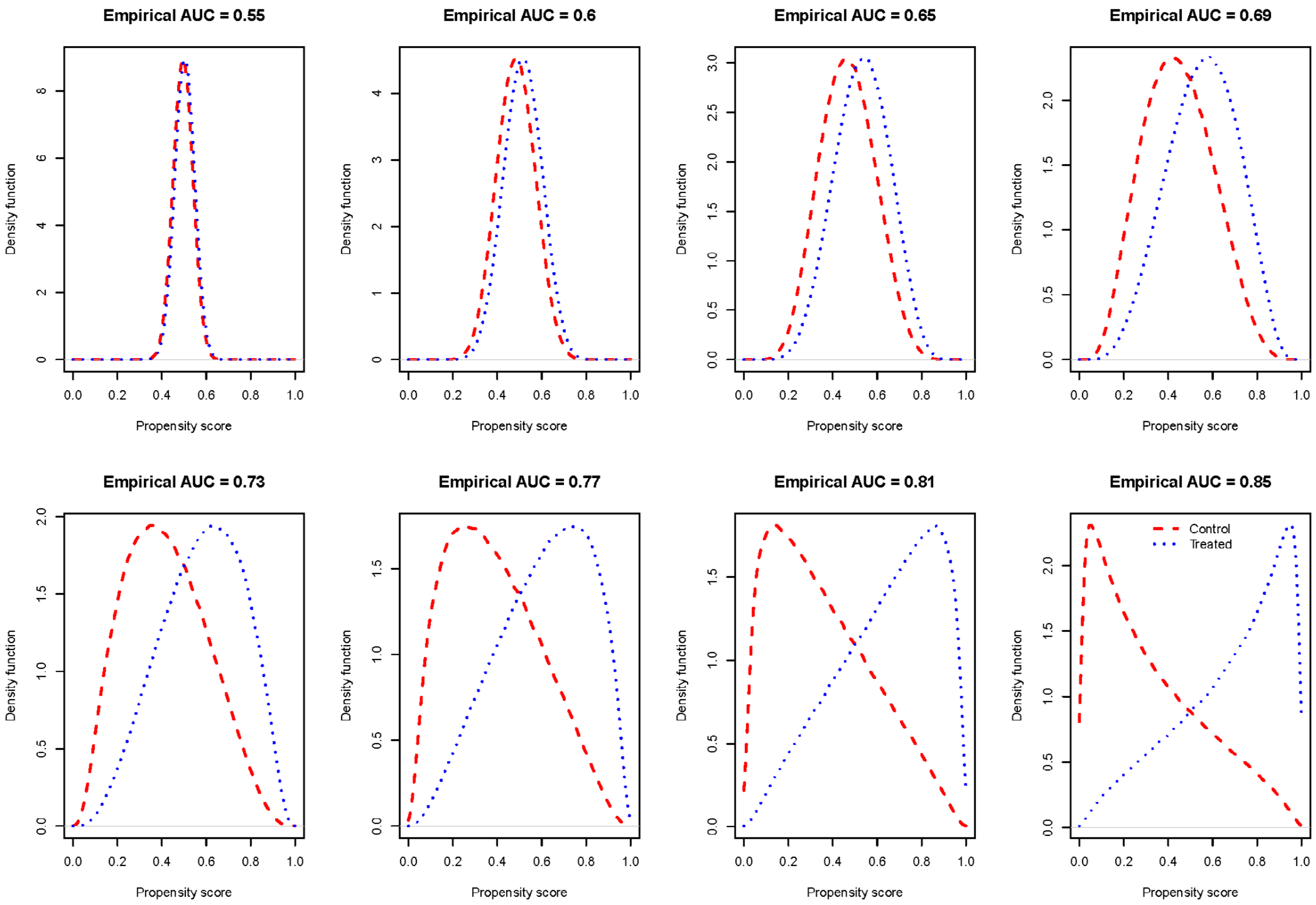

FIGURE 1 Distribution of the propensity score in treated and control subjects (primary simulations) [Colour figure can be viewed at wileyonlinelibrary.com]

each of the eight empirical c-statistics (as opposed to the target c-statistic). The overlap in the propensity score distribution between the treated and control subjects decreased as the empirical c-statistic increased. Thus, an increasing c-statistic is indicative of increasing dissimilarities between treated and control subjects. Note that despite the covariate being normally distributed, the distribution of the propensity can be non-normal in treated and control patients, particularly when the c-statistic is moderate to high. There were scenarios with a high overlap in the distribution of the propensity score between treated and control subjects and scenarios with relatively little overlap.

An R function is provided in the Appendix A that allows one to compute the VIF for each set of weights and for any value of the c-statistic of the propensity score model and prevalence of treatment. The function has two input parameters: the prevalence of treatment and the target c-statistic of the propensity score model. It assumes that the covariate follows a standard normal distribution and uses the methods described above to estimate the regression coefficients for the propensity score model. Along with outputting the VIFs for the different sets of weights, it also provides the empirical c-statistic that was associated with the simulations.

3.2 | Secondary simulations: non-normally distribution predictor variable

We repeated the above set of simulations, using three different distributions for the baseline covariate: a Beta distribution, a Chi-squared distribution, and a log-normal distribution. In the first set of secondary simulation, we simulated a continuous baseline covariate for each of 1 000 000 subjects: $x_i \sim \beta(1, 1)$. In the second set of secondary simulations, we assumed that the covariate followed Chi-squared distribution with 3 degrees of freedom. In the third set of secondary simulations, we assumed that the covariate followed a standard log-normal distribution (ie, the logarithm of the random variable had mean zero and variance one). In the latter two sets of simulations, once the baseline covariate was simulated, we standardized it to have mean zero and SD one.

In all three secondary simulations we specified a logistic model for treatment status as $\text{logit}(\Pr(Z = 1)) = \alpha_0 + \alpha_1 X$ and used an iterative bisection approach to determine values of α_0 and α_1 to induce the desired prevalence of treatment and c-statistic. Apart from these modifications, the simulations were conducted identically to those described above. Figures S1–S3 in the online supplemental material describe the distribution of the propensity score separately in treated and control subjects when the target c-statistic ranged from 0.55 to 0.90 in increments of 0.05 and the prevalence of treatment was 0.5.

4 | EFFECTS OF STUDY CHARACTERISTICS ON THE VIF: RESULTS

4.1 | Primary set of simulations

The relationship between the prevalence of treatment, the empirical c-statistic of the propensity score model and the VIF is reported in Figure 2 (there is one panel for each of the five sets of weights). We added a horizontal line denoting a VIF of 2 to two of the panels, indicating that the use of weights resulted in an effective sample size that was 50% smaller than the initial sample. Note that the scale of the vertical axis for the figures for the IPTW-ATE and IPTW-ATT weights is substantially different from that for the other three sets of weights.

Across all five sets of weights, the VIF increased as the empirical c-statistic of the propensity score model increased. When using IPTW-ATE weights, the VIF tended to be below 2 when the c-statistic of the propensity score model was less than or equal to 0.75. The VIF was very large when the c-statistic was very high and the prevalence of treatment was either very low or very high. Similar observations were made for the ITPW-ATT weights. With IPTW-OW, IPTW-MW, and IPTW-EW, the VIF was always less than 2, even when the c-statistic of the propensity score model was very high and treatment prevalence was very low or very high. When the empirical c-statistic of the propensity score model was modest ($\leq$0.75), then the VIF was always lower than 1.3 for these three latter sets of weights.

For each set of weights we used quantile regression to regress the logarithm of the VIF on the empirical c-statistic of the propensity score, the square of the empirical c-statistic of the propensity score, and the prevalence of treatment, with the latter being treated as a categorical variable with nine levels.[20,21] The estimated regression coefficients for each of the five models are reported in Table 1. For each of the 153 scenarios and each set of weights, we computed the absolute difference between the true VIF and the VIF estimated using the regression coefficients. The median absolute difference across the 153 scenarios was 0.12 (IPTW-ATE), 0.17 (IPTW-ATT), 0.01 (IPTW-OW), 0.01 (IPTW-MW),

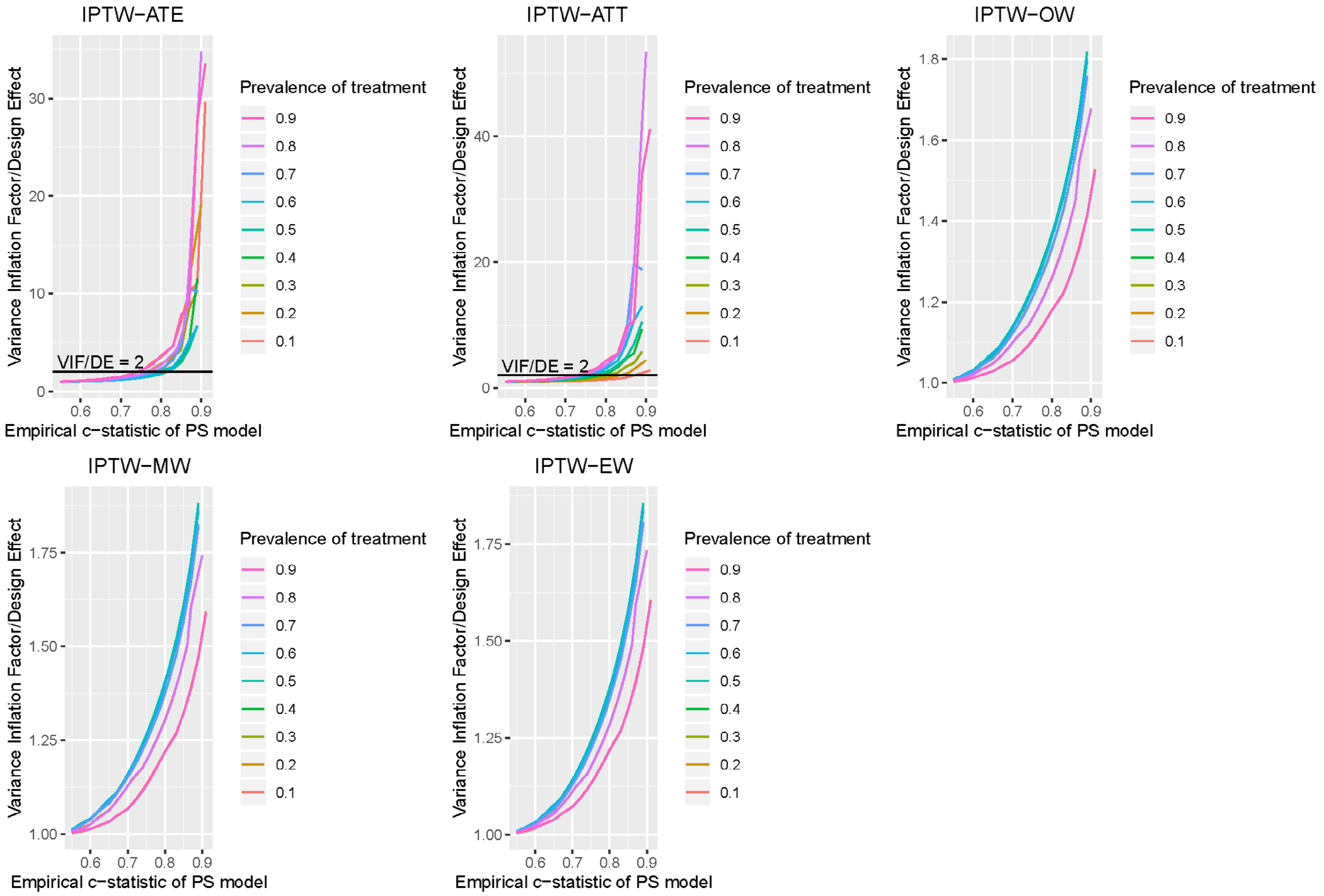

FIGURE 2 Variance inflation factors/design effects for main simulations [Colour figure can be viewed at wileyonlinelibrary.com]

TABLE 1 Quantile regression results for VIF analysis, where the outcome is log(VIF)

Parameter	ATE	ATT	OW	MW	EW
Intercept	10.88	8.65	1.18	1	1.27
c-Statistic	−34.53	−29.9	−4.59	−4.11	−4.85
c-Statistic^2	28.03	24.84	4.21	3.94	4.44
Prevalence of treatment = 0.2	−0.16	0.09	0.06	0.07	0.05
Prevalence of treatment = 0.3	−0.25	0.24	0.09	0.09	0.07
Prevalence of treatment = 0.4	−0.34	0.36	0.1	0.1	0.08
Prevalence of treatment = 0.5	−0.36	0.39	0.1	0.1	0.09
Prevalence of treatment = 0.6	−0.36	0.48	0.1	0.1	0.08
Prevalence of treatment = 0.7	−0.21	0.55	0.09	0.09	0.07
Prevalence of treatment = 0.8	−0.18	0.61	0.06	0.07	0.05
Prevalence of treatment = 0.9	0	0.66	0	0	0
Median absolute prediction error					
Main simulations	0.12	0.17	0.01	0.01	0.01
Beta distribution	0.19	0.28	0.02	0.03	0.02
Chi-squared distribution	0.87	2.10	0.04	0.04	0.04
Log-normal distribution	2.81	18.72	0.07	0.06	0.06

Note: The quantile regression models are for log(VIF). The linear predictor must be exponentiated to obtain estimated VIF.

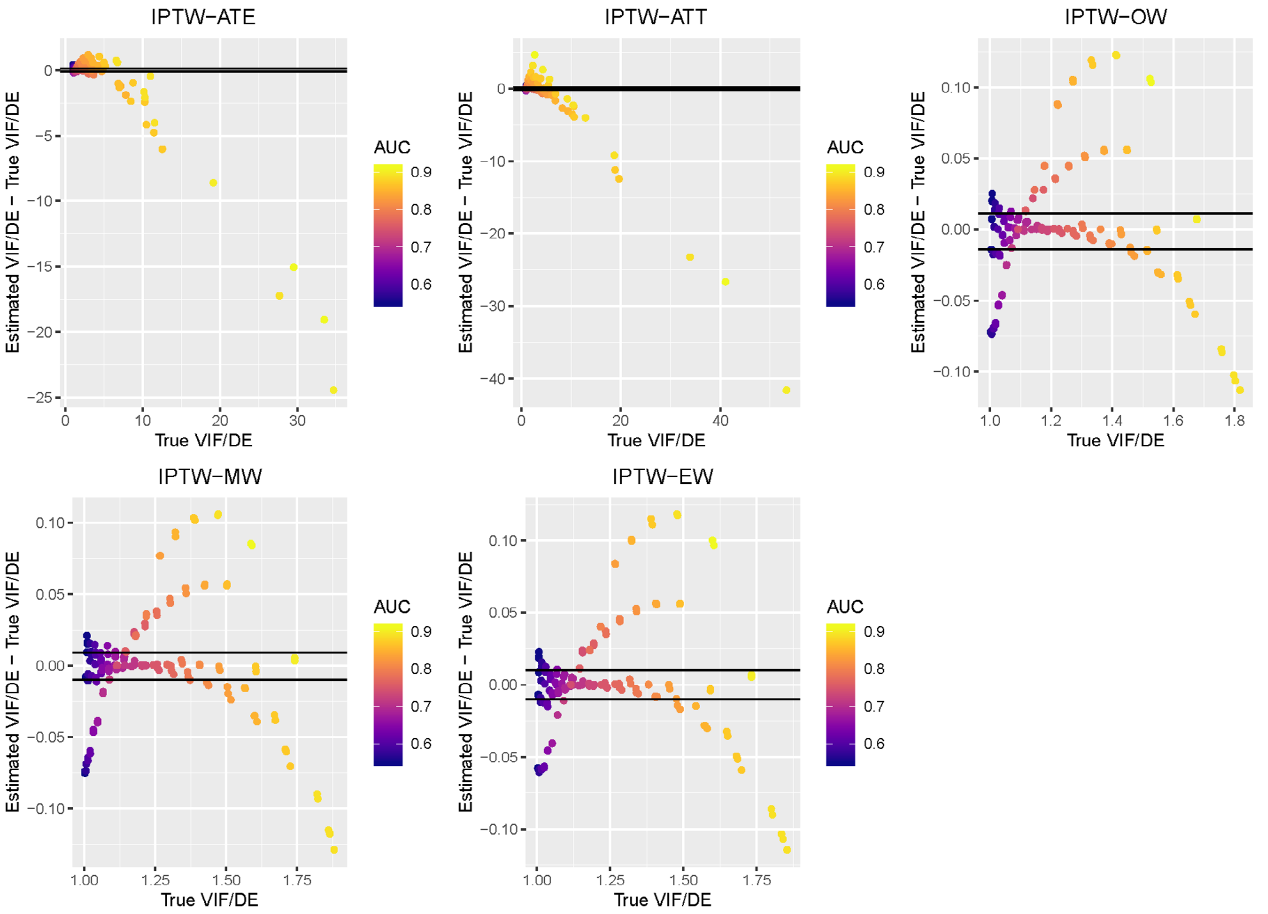

FIGURE 3 Comparing estimated and true variance inflation factor/design effect: normal distribution [Colour figure can be viewed at wileyonlinelibrary.com]

and 0.01 (IPTW-EW) (Table 1). Modified Bland-Altman plots comparing estimated vs true VIFs across the 153 scenarios are reported in Figure 3.[22] The horizontal axis denotes the true VIF, while the vertical axis denotes the difference between true and estimated VIF. There is one panel for each set of weights. The points in each panel are color coded to indicate the empirical c-statistic associated with that point. On each panel we have superimposed two horizontal lines denoting the 25th and 75th percentiles of the absolute difference between the true and estimated VIFs across the 153 scenarios. With the IPTW-ATE and IPTW-ATT weights, prediction error was highest when the c-statistic was very high. In general, prediction was accurate when the c-statistic of the propensity score model was not very high.

4.2 | Results of secondary simulations

In each of the three sets of secondary simulations, the empirical prevalence of treatment was within 0.01 of the target prevalence of treatment across the 153 scenarios. Due to these minimal differences, we use the target prevalence of treatment in the following analyses. The relationship between the empirical c-statistic of the propensity score model and the VIF across different prevalences of treatment is described in Figures S4–S6 in the supplemental online material. In general, for a given prevalence of treatment, the VIF increased as the c-statistic of the propensity score model increased. There were scenarios in which very large VIFs were observed for the IPTW-ATE and IPTW-ATT weights.

For each set of secondary simulations, we applied the quantile regression model estimated in the previous section (whose coefficients are reported in Table 1) when the baseline covariate was normally distributed. The median absolute difference between estimated and true VIFs are reported in Table 1. Modified Bland-Altman plots comparing estimated

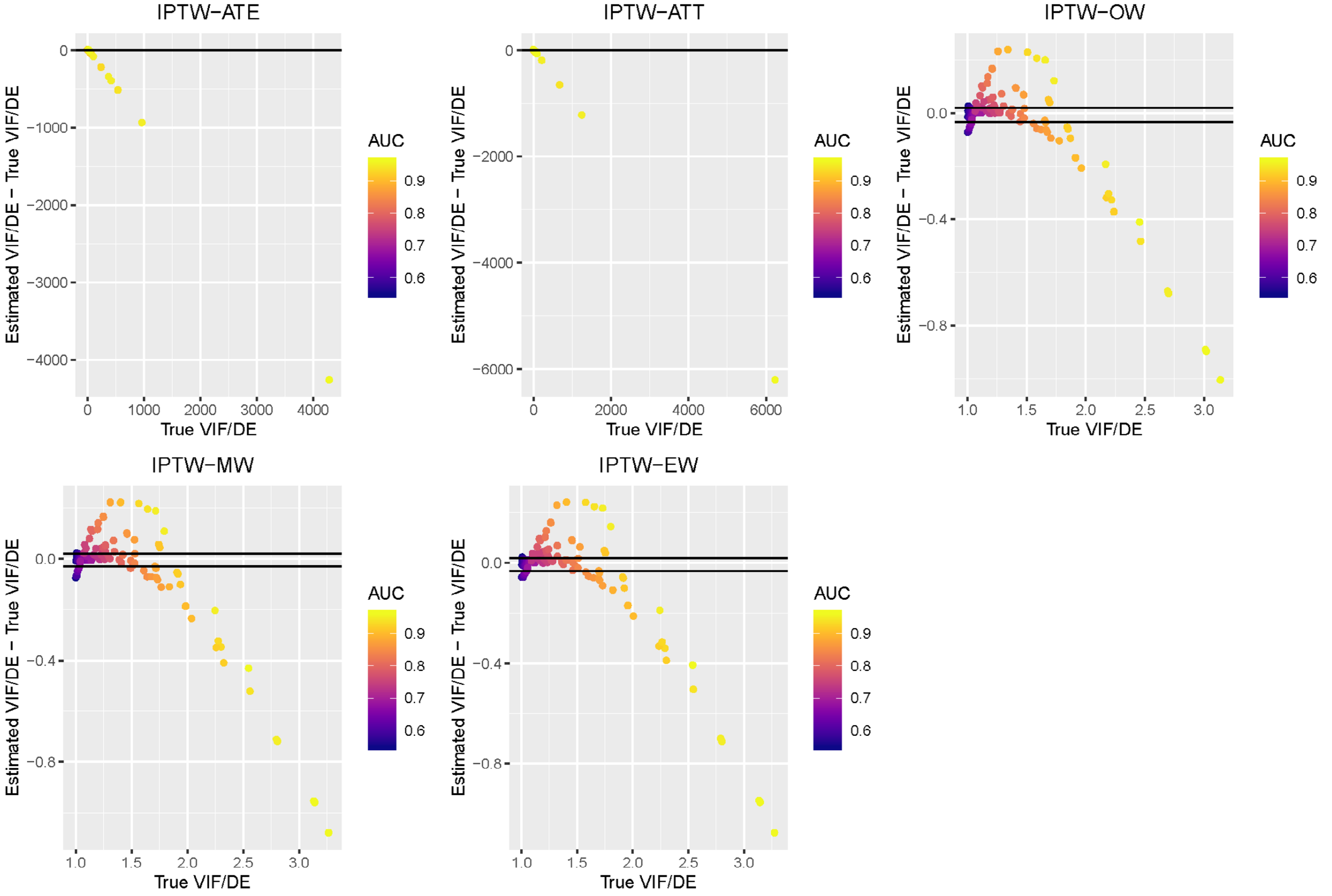

FIGURE 4 Comparing estimated and true VIF/DE: Beta distribution [Colour figure can be viewed at wileyonlinelibrary.com]

and true VIFs across the 153 scenarios for each of the three sets of secondary simulations are described in Figures 4–6. As in the main set of simulations, prediction error tended to be minimal except when the c-statistic of the propensity score model was very high.

5 | CASE STUDY

Assume that an investigator wanted to design an observational study to compare the effectiveness of coronary artery bypass graft (CABG) surgery with that of percutaneous coronary interventions with drug eluting stents in patients with unprotected left main coronary artery disease. In a previous study on this topic, Zheng and colleagues reported a propensity score model with a c-statistic of 0.831 and that, in their single-center study, approximately two thirds of patients were treated with CABG surgery.[23]

We assumed that the prevalence of CABG surgery in the setting in which the new study is being planned was 0.67 and that the c-statistic of the propensity score model was 0.83 (ie, we assumed that both the prevalence of treatment and the c-statistic in the new study would be the same as in the study of Zheng and colleagues). We estimated the VIF for the five different sets of weights using both the regression equations in Table 1 and the R function provided in the Appendix A (to do so, we rounded the prevalence of treatment to 0.70, as this factor is categorical in the regression model). When using the regression equations in Table 1, the estimated VIFs were 3.75 (IPTW-ATE), 4.42 (IPTW-ATT), 1.43 (IPTW-OW), 1.47 (IPTW-MW), and 1.45 (IPTW-EW). When using the R function, the estimated VIFs were 1.71 (IPTW-ATE), 2.41 (IPTW-ATT), 1.35 (IPTW-OW), 1.37 (IPTW-MW), and 1.35 (IPTW-EW). The function indicated that the empirical c-statistic upon which this was based was 0.80. Using an input c-statistic of 0.88 produced an empirical c-statistic of 0.83. The VIFs associated with this value of the c-statistic were 3.11 (IPTW-ATE), 5.28 (IPTW-ATT), 1.46 (IPTW-OW), 1.52 (IPTW-MW), and 1.42 (IPTW-EW).

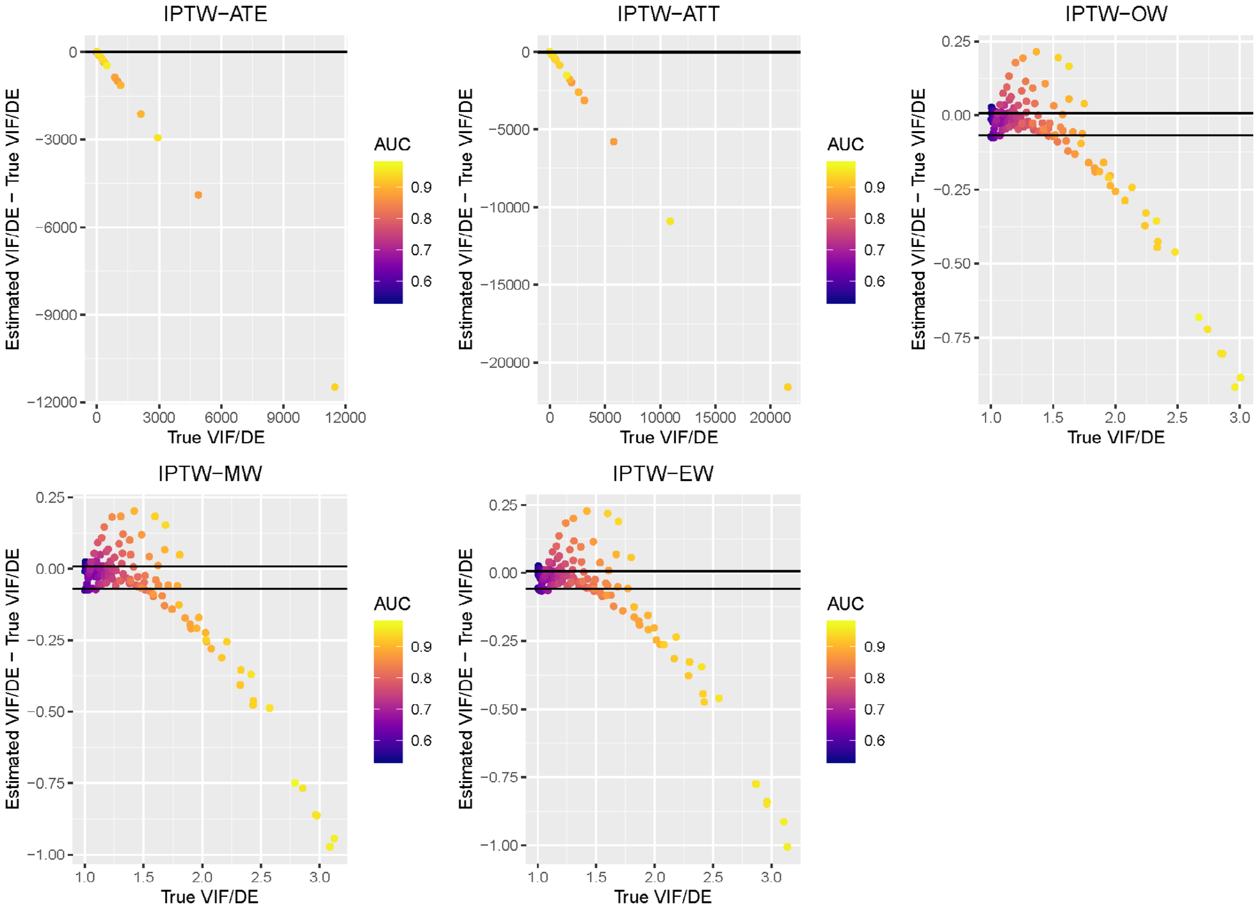

FIGURE 5 Comparing estimated and true VIF/DE: Chi-squared distribution [Colour figure can be viewed at wileyonlinelibrary.com]

In an RCT in which the probability of assignment to the experimental intervention was 0.67, we would require 865 subjects under the assumption that the event was observed for 25% of subjects (and the remaining 75% were censored) to have a power of 80% to detect a hazard ratio of at least 1.50 using a significance level of 0.05 (Source: PASS 2020, version 20.0.3, NCSS LLC, Kaysville, UT).

Given the within-subject homogeneity in outcomes induced by weighting, we would multiply the above sample size (865) by the estimated VIF. Using the VIFs obtained using the R function (with a c-statistic of 0.88 producing an empirical c-statistic of 0.83), we would need a sample size of 3244 (IPTW-ATE), 3823 (IPTW-ATT), 1237 (IPTW-OW), 1272 (IPTW-MW), or 1254 (IPTW-EW). Thus, depending on the weights used, the observational study that we are designing would require between 1237 subjects and 3823 subjects. Note that the choice of which set of weights to use should not be decided primarily by which requires the lowest sample size or which results in the greatest statistical power. Instead, the decision should be informed, at least in part, by which target estimand is most appropriate for addressing the investigators' researcher question. The original study by Zheng and colleagues included 4046 subjects, and thus a study of that size would have been adequately powered, regardless of which set of weights the authors elected to use.

6 | DISCUSSION

Sample size and power calculations for studies using propensity score weighting require knowledge of the weights to allow the computation of the VIF. However, the weights are only known once the study sample has been assembled. The reporting of post-hoc power calculations have been criticized by statistical authors.[4] The results provided in the current study can facilitate sample size and power calculations before conducting weighted analyses. These results can facilitate good study design by allowing for sample size and power calculations to be conducted prior to the study analyses being conducted.

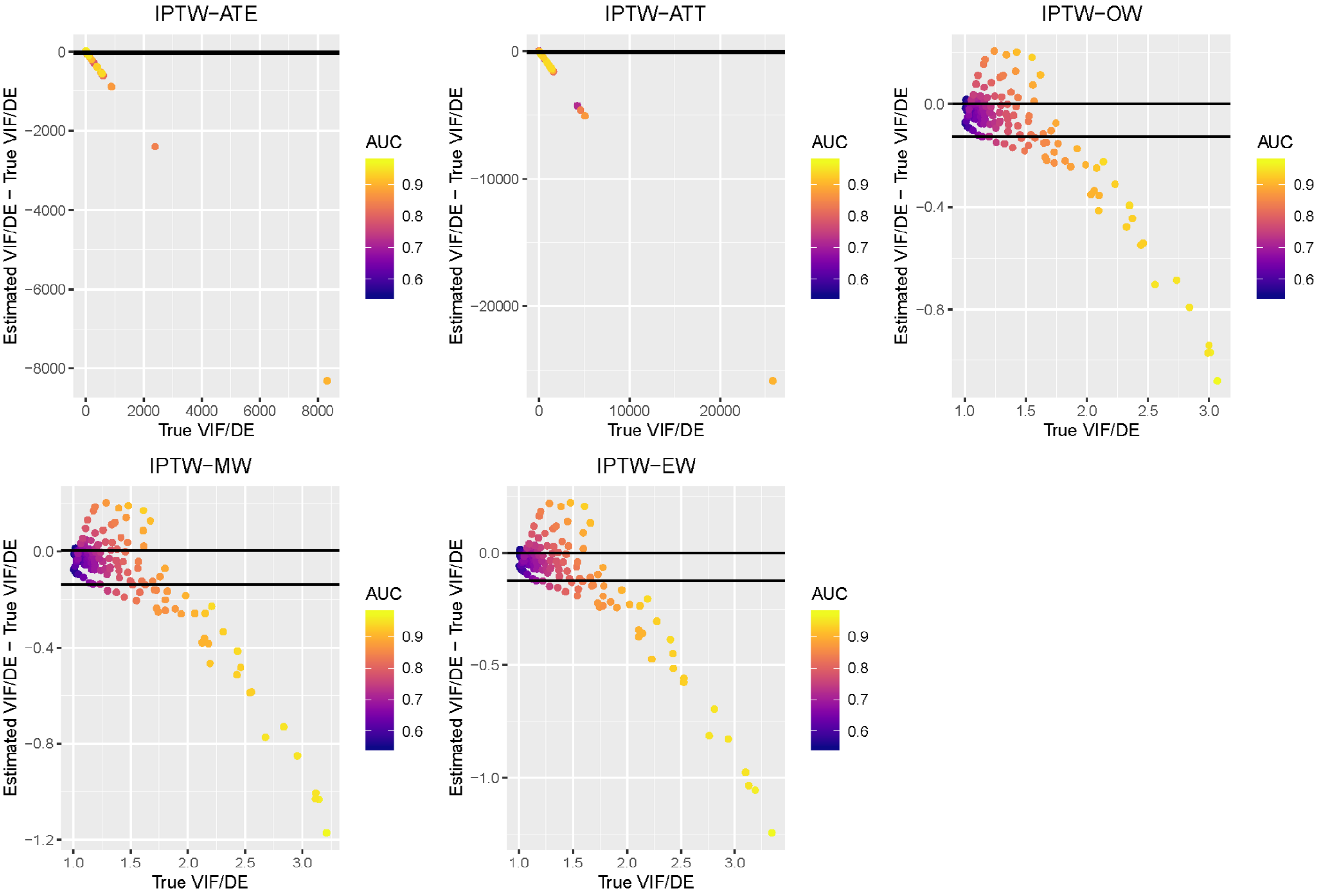

FIGURE 6 Comparing estimated and true VIF/DE: log-normal distribution [Colour figure can be viewed at wileyonlinelibrary.com]

Propensity-score weighting accounts for confounding using weighting. In the weighted sample, the distribution of measured baseline covariates is similar in treated subjects as in control subjects.[6] Consequently, the statistical analyses conducted in the weighted sample can reflect the analyses that would be conducted in an RCT with a similar outcome variable. Thus, for continuous outcomes, the mean outcome can be computed in each arm and the difference in means reported. For binary outcomes, the probability of success or failure can be reported in each arm and the difference and ratio of these probabilities can be reported as the absolute risk difference and the relative risk. For time-to-event outcomes, survival curves can be reported along with the absolute difference in survival at clinically relevant times. These can be complemented with the reporting of a hazard ratio from a univariate Cox regression model.[24] While the estimation of the point estimate of the effect of treatment in a weighted analysis can reflect what would be done in an RCT, variance estimation must account for the within-subject homogeneity in outcomes induced by weighting. Note that the validity of conclusions drawn from propensity score analyses rest on two assumptions: (i) the assumption of no unmeasured confounders; (ii) the positivity assumption.[5] The latter assumption is that all subjects have a non-zero probability of receiving either treatment.

We suggest that investigators using weighting in observational studies proceed as follows: first, obtain estimates of the c-statistic of the propensity score model and of the prevalence of treatment. These can be informed by previously published research, pilot studies, or by the investigators' clinical judgment. Granger and colleagues conducted a review of the use of propensity score diagnostics in papers published in the medical literature and noted which studies reported estimates of the c-statistic of the propensity score model.[25] Similarly, Sturmer and colleagues reviewed the use of propensity score methods in the medical literature.[26] They note that the c-statistic of the propensity score model was presented in 73 studies and provide a table containing the study-specific c-statistics. These reviews can serve as resources for obtaining plausible estimates of the c-statistic across different areas of medical research and for different exposures. Alternatively, investigators can select a range of plausible values of the c-statistic. Second, compute the VIF using either the regression results in Table 1 or the R function provided in the

Appendix A. Third, determine the necessary sample size for a two-armed RCT with the same prevalence of treatment and with the same type of outcome variable. Fourth, inflate the estimated sample size by the VIF obtained in the second step.

We examined the VIF for five different sets of weights. The VIFs for the use of IPTW-OW, IPTW-MW, and IPTW-EW tended to be much smaller than those for IPTW-ATE and IPTW-ATT. This finding complements those of Zhou and colleagues who found that the latter three sets of weights results in estimates that had lower standard errors and that displayed lower variability compared to estimates obtained using the first two sets of weights.[10]

The choice of which set of weights to use should not be dictated solely by which induces the lowest VIF, and thus results in the highest statistical power (assuming a fixed sample size). Instead, the primary motivation should be which target estimands best addresses the investigators' research question. The differences between the VIFs for the first two sets of weights and the last three sets of weights were amplified as the c-statistic increased. As the c-statistic increased, differences in the distribution of the propensity score between treated and control subjects were amplified. This suggest that in settings with a high c-statistic, there is likely a sub-population of subjects who would most often receive the control exposure and a different sub-population of subjects who would most often receive the active exposure. In these settings, the focus of interest may be on those subjects for whom there is clinical equipoise. Thus, in settings with a very high c-statistic, the use of IPTW-OW, IPTW-MW, or IPTW-EW may be preferable to the use of IPTW-ATE or IPTW-ATT weights. It is these settings that these three sets of weights offer a striking advantage, resulting in much smaller VIFs that the use of the other two sets of weights.

There are certain limitations to the current study. First, there are other propensity score methods, such as matching on the propensity score. We have focused on methods to estimate sample size requirements and statistical power when using weighting-based methods. Future research should focus on developing comparable methods for matching-based methods. Second, there is another set of weights that we have not examined: stabilized weights, which are defined as $w_{\text{stab}}(\mathbf{X}) = \Pr(Z = 1)(Z/e(\mathbf{X})) + \Pr(Z = 0)(1 - Z)/(1 - e(\mathbf{X}))$.[8,27] However, if one computes the VIF associated with the use of stabilized weights, it is equal to the VIF associated with the use of IPTW-ATE weights. Thus, all our findings for IPTW-ATE weights would apply to their stabilized counterpart. Third, our simulations were restricted to settings with a single covariate. However, our findings should be generalizable to more complex settings. If there were multiple covariates (either continuous, categorical, or a mixture of the two), we could consider the linear predictor that combined these covariates. As the number of covariates increased, the central limit theorem would suggest that, in many settings, the linear predictor would be approximately normally distributed. Finally, we have assumed a prospective sample size calculation and that the assembled cohort will reflect this calculation. However, the proposed method may adapted to work with dynamic sample size methods or the necessary sample size may be re-estimated once interim data are available and accurate estimates of the c-statistic and prevalence of treatment are available.

In summary, knowledge of the VIF allows for conducting sample and power size calculations for observational studies that use propensity score weighting. We provide methods to estimate the VIF based on two characteristics of the observational study: the prevalence of treatment and the anticipated c-statistic of the propensity score model. Implementation of these methods allows for improvements in the design and reporting of observational studies that use propensity score weighting.

ACKNOWLEDGEMENTS

This study was supported by ICES, which is funded by an annual grant from the Ontario Ministry of Health and Long-Term Care (MOHLTC). The opinions, results, and conclusions reported in this paper are those of the authors and are independent from the funding sources. No endorsement by ICES or the Ontario MOHLTC is intended or should be inferred. This research was supported by operating grant from the Canadian Institutes of Health Research (CIHR) (PJT 166161). Dr. Austin is supported in part by a Mid-Career Investigator award from the Heart and Stroke Foundation of Ontario.

DATA AVAILABILITY STATEMENT

Data sharing not applicable to this article as no empirical datasets were analyzed during the current study.

ORCID

Peter C. Austin https://orcid.org/0000-0003-3337-233X

REFERENCES

1. Dorn HF. Philosophy of inference from retrospective studies. *Am J Public Health*. 1953;43:692-699.
2. Rubin DB. *Matched Sampling for Causal Effects*. New York, NY: Cambridge University Press; 2006.
3. Hernan MA, Robins JM. Using big data to emulate a target trial when a randomized trial is not available. *Am J Epidemiol*. 2016;183(8):758-764.
4. Hoenig JM, Heisey DM. The abuse of power: the pervasive fallacy of power calculations for data analysis. *Am Stat*. 2001;55(1):19-24.
5. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika*. 1983;70:41-55.
6. Austin PC. An introduction to propensity-score methods for reducing the effects of confounding in observational studies. *Multivar Behav Res*. 2011;46:399-424.
7. Rosenbaum PR. Model-based direct adjustment. *J Am Stat Assoc*. 1987;82:387-394.
8. Hernan MA, Brumback B, Robins JM. Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. *Epidemiology*. 2000;11(5):561-570.
9. van der Wal WM, Geskus RB. Ipw: an R package for inverse probability weighting. *J Stat Softw*. 2011;43(13).
10. Zhou Y, Matsouaka RA, Thomas L. Propensity score weighting under limited overlap and model misspecification. *Stat Methods Med Res*. 2020;29(12):3721-3756.
11. Hsieh FY, Lavori PW, Cohen HJ, Feussner JR. An overview of variance inflation factors for sample-size calculation. *Eval Health Prof*. 2003;26(3):239-257.
12. Donner A, Klar N. *Design and Analysis of Cluster Randomization Trials in Health Research*. London, UK: Arnold; 2000.
13. Lohr SL. *Sampling: Design and Analysis*. Vol 2. Brooks/Cole: Boston, MA; 2010.
14. Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. *J Educ Psychol*. 1974;66:688-701.
15. Morgan SL, Winship C. *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. New York, NY: Cambridge University Press; 2007.
16. Li L, Greene T. A weighting analogue to pair matching in propensity score analysis. *Int J Biostat*. 2013;9(2):215-234.
17. Thomas LE, Li F, Pencina MJ. Overlap weighting: a propensity score method that mimics attributes of a randomized clinical trial. *JAMA*. 2020;323(23):2417-2418.
18. Golinelli D, Ridgeway G, Rhoades H, Tucker J, Wenzel S. Bias and variance trade-offs when combining propensity score weighting and regression: with an application to HIV status and homeless men. *Health Serv Outcomes Res Methodol*. 2012;12(2–3):104-118.
19. Austin PC, Steyerberg EW. Interpreting the concordance statistic of a logistic regression model: relation to the variance and odds ratio of a continuous explanatory variable. *BMC Med Res Methodol*. 2012;12:82.
20. Austin PC, Schull MJ. Quantile regression: a statistical tool for out-of-hospital research. *Acad Emerg Med*. 2003;10(7):789-797.
21. Austin PC, Tu JV, Daly PA, Alter DA. The use of quantile regression in health care research: a case study examining gender differences in the timeliness of thrombolytic therapy. *Stat Med*. 2005;24(5):791-816.
22. Krouwer JS. Why bland-Altman plots should use X, not (Y+X)/2 when X is a reference method. *Stat Med*. 2008;27(5):778-780.
23. Zheng Z, Xu B, Zhang H, et al. Coronary artery bypass graft surgery and percutaneous coronary interventions in patients with unprotected left Main coronary artery disease. *JACC Cardiovasc Interv*. 2016;9(11):1102-1111.
24. Austin PC, Laupacis A. A tutorial on methods to estimating clinically and policy-meaningful measures of treatment effects in prospective observational studies: a review. *Int J Biostat*. 2011;7(1):6.
25. Granger E, Watkins T, Sergeant JC, Lunt M. A review of the use of propensity score diagnostics in papers published in high-ranking medical journals. *BMC Med Res Methodol*. 2020;20(1):132.
26. Sturmer T, Joshi M, Glynn RJ, Avorn J, Rothman KJ, Schneeweiss S. A review of the application of propensity score methods yielded increasing use, advantages in specific settings, but not substantially different estimates compared with conventional multivariable methods. *J Clin Epidemiol*. 2006;59(5):437-447.
27. Cole SR, Hernan MA. Adjusted survival curves with inverse probability weights. *Comput Methods Programs Biomed*. 2004;75:45-49.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Austin PC. Informing power and sample size calculations when using inverse probability of treatment weighting using the propensity score. *Statistics in Medicine*. 2021;40(27):6150-6163. https://doi.org/10.1002/sim.9176

APPENDIX A

R code for computing VIF

```
# Function for estimating VIF as a function of c-statistic and
# prevalnce of treatment.
# THIS SOFTWARE IS PROVIDED FOR ILLUSTRATIVE PURPOSES AND COMES WITH ABSOLUTELY
# NO WARRANTY.

library(rms)

vif <- function(auc,prev.treat){
# auc: c-statistic of PS model.
# prev.treat: Prevalance of treatment.

set.seed(1)

N <- 1000000
# Population size.

########################################################################################
# Generate covariate.
########################################################################################

x <- rnorm(N)
# Assume standardized so that it is standard normal.

########################################################################################
# Generate beta for PS model.
########################################################################################

# auc = phi(sigma*beta/sqrt(2))

beta.ps <- sqrt(2) * qnorm(auc)

########################################################################################
# Determine intercept for PS model for desired prevalance of treatment.
########################################################################################

bias <- 1

lower.int <- -20
upper.int <- 20

while (bias > 0.001){

  beta0.treat <- (lower.int + upper.int)/2

  # Generate treatment status for each subject.
  logit.treat <- beta0.treat + beta.ps*x
  p.treat <- exp(logit.treat)/(1 + exp(logit.treat))
  treat <- rbinom(N,1,p.treat)
  emp.prev.treat <- mean(treat)

  bias <- abs(prev.treat - emp.prev.treat)

  if (emp.prev.treat > prev.treat) {
    upper.int <- beta0.treat
    } else {
        lower.int <- beta0.treat
    }
    remove(logit.treat,p.treat,emp.prev.treat)
}

remove(bias,lower.int,upper.int)

########################################################################################
# Compute weights
########################################################################################

psm <- lrm(treat ~ x)
auc.emp <- psm$stats["C"]
ps <- exp(psm$linear.predictor)/(1 + exp(psm$linear.predictor))

iptw <- (treat/ps) + (1-treat)/(1-ps)
att <- treat + (1-treat)*(ps/(1-ps))
ow <- treat*(1-ps) + (1-treat)*ps
mw <- treat*pmin(ps,1-ps)/ps + (1-treat)*pmin(ps,1-ps)/(1-ps)
entropy <- treat*(-ps*log(ps) - (1-ps)*log(1-ps))/ps +
    (1-treat)*(-ps*log(ps) - (1-ps)*log(1-ps))/(1-ps)

remove(x,beta.ps,psm,ps)

########################################################################################
# Compute VIFs
########################################################################################

N1 <- sum(treat)
N0 <- N - N1

VIF.iptw <- (N1*N0/N) * (sum(treat*iptw*iptw)/((sum(treat*iptw))^2) +
    sum((1-treat)*iptw*iptw)/((sum((1-treat)*iptw))^2) )
VIF.att <- (N1*N0/N) * (sum(treat*att*att)/((sum(treat*att))^2) +
    sum((1-treat)*att*att)/((sum((1-treat)*att))^2) )
VIF.ow <- (N1*N0/N) * (sum(treat*ow*ow)/((sum(treat*ow))^2) +
    sum((1-treat)*ow*ow)/((sum((1-treat)*ow))^2) )
VIF.mw <- (N1*N0/N) * (sum(treat*mw*mw)/((sum(treat*mw))^2) +
    sum((1-treat)*ow*ow)/((sum((1-treat)*ow))^2) )
VIF.entropy <- (N1*N0/N) * (sum(treat*entropy*entropy)/((sum(treat*entropy))^2) +
    sum((1-treat)*ow*ow)/((sum((1-treat)*ow))^2) )

remove(N1,N0,iptw,att,ow,mw,entropy)

return(c(auc,auc.emp,prev.treat,VIF.iptw,VIF.att,VIF.ow,VIF.mw,VIF.entropy))

}
```