Positivity Sized-Up Effectively

Assessing Stochastic Positivity in Causal Inference via Effective Sample Size

by

Bo Hofstede

to obtain the degree of Master of Science
in Computer Science – Data Science & Technology track
at the Delft University of Technology,
to be defended publicly on Friday May 1st, 2026 at 10:00.

Student number:	5099404	
Project duration:	September 15th, 2025 – May 1st, 2026	
Thesis committee:	Dr. ir. J. H. Krijthe,	TU Delft, advisor
	Dr. A. Lukina,	TU Delft, committee member

An electronic version of this thesis is available at https://repository.tudelft.nl/.

Positivity Sized-Up Effectively

Assessing Stochastic Positivity in Causal Inference via Effective Sample Size

Bo Hofstede
Pattern Recognition Lab
Delft University of Technology
The Netherlands

***Abstract*—**
Causal inference relies on several key identifying assumptions, including positivity: all treatment levels must have non-zero probability for every possible covariate combination. Violations lead to unreliable causal effect estimates, yet positivity is often overlooked, and existing diagnostics have limitations. This assumption is particularly relevant for observational data, because treatment assignment is not independent of confounders. To remove this dependence, Inverse probability of treatment weighting (IPTW) estimators can be used. However, IPTW relies on the positivity assumption, and near-violations lead to extreme weights and unstable estimates. We investigate *effective sample size* (ESS) as a practical diagnostic for evaluating the estimability of causal effects in the face of near-positivity violations. The key contribution is a theoretical definition of 'targeted ESS' that aligns with causal inference. Targeted ESS can quantify how many observations effectively contribute to weighted estimates and can serve as an intuitive tool for communicating positivity concerns. Through analysis and simulations, we demonstrate its strengths and limitations. Notably, targeted ESS cannot detect severe cases of positivity violations or propensity model misspecifications. Additionally, we show why conventional ESS is not generally suitable in this setting. This work offers practical guidance for assessing IPTW estimate reliability in observational causal inference.

***Index Terms*—Causal Inference, Positivity, Effective Sample Size, Overlap, Common support, Inverse Probability of Treatment Weight, Horvitz-Thompson, Hájek**

I. Introduction

There are many situations in which researchers are interested in the causal effects of a treatment. This includes, for example, testing the effect of a new social benefit program (Imbens & Xu, 2024; LaLonde, 1986), evaluating the efficacy of a sugical procedure (Austin & Stuart, 2015), or comparing users of a consumer product (Kuesten et al., 2016). Randomised experiments provide a solution under certain assumptions, but are not always feasible. Observational data is often easier to collect but requires greater consideration of assumptions. This includes considering the effect of any *confounder*, a variable that can both affect the treatment and the outcome. For example, young people might be less likely to apply to a social benefit program but would, on average, gain more from it. Then, ignoring that there are only a few young individuals in the data would make the program seem less effective. In observational studies, it is arguably easy to overlook aspects of the question and the data that can bias estimates and lead to incorrect inferences. Causal inference relies on assumptions, but these assumptions should be carefully tested.

One of the key assumptions for causal inference, and the focus of this paper, is *positivity*: all treatment levels must have non-zero probability for every combination of confounder values that occur in the population. Each individual who can exist given the confounders should be able to receive or not receive the social benefits. An exponential amount of data, relative to the number of confounders, is required to prove this assumption holds. For example, if a the effect of a social benefits program study includes the following confounders: income (3 groups), income certainty (3 groups), age (5 groups), household composition (4 categories), home type (4 categories), home location (urban/rural) and education level (3 groups) where we consider any combination possible, then we would need to have at least $8,640$ unique samples, one for each possible combination, to guarantee that positivity is satisfied. Additionally, if we measure age continuously, there are an infinite number of ages that one would need to observe across all treatment levels to guarantee that positivity is satisfied. This is only feasible if we make further assumptions. Compared to positivity, there is a greater focus on the assumption of no unmeasured confounders in the causal inference literature (Petersen et al., 2012; Zhu et al., 2021). The direct approach to avoid unmeasured confounders is to include more covariates in the analysis, but this makes it more difficult to satisfy positivity. Despite its critical importance, positivity violations are common in practice yet often go undetected or unreported in empirical research (Bettega et al., 2024; Stuart, 2010; Westreich & Cole, 2010). There is a need to emphasise the effects of positivity when considering causal claims.

A common way to handle confounding in observational data is by using *inverse probability of treatment weights* (IPTW). There, the probability of treatment conditional on covariates, also known as the propensity score, is used to weight the sample and address confounding based on the covariates. Consequently, not only the theoretical support but also the observed overlap in covariate distributions between treatment levels becomes relevant. This also showcases two different types of positivity violations (Zivich et al., 2022). First, deterministic (or structural) violations occur when certain combinations of covariates are necessarily impossible

for a particular treatment level. Consider social benefits that are only given if a person's income is under a certain threshold. Second, stochastic (or random) violations occur when a covariate combination is possible for all treatment values but empirically rare. For example, if elderly people rarely apply for the benefit program in practice despite being eligible, it is difficult to confidently measure the program's effect on them. Stochastic positivity requires having sufficient data to answer the question of interest. When it is violated, the conditional probability of treatment is either 0 or 1. Violations, or near-violations, can lead to extreme inverse probability weights and unstable estimates. It calls into question the reliability of causal conclusions.

Existing diagnostic methods for detecting positivity violations are limited. Some focus on univariate balance rather than multivariate overlap, while most questions involve multiple confounding variables (Stuart, 2010). Others are estimator-specific, rather than data-centric (Petersen et al., 2012). Finally, some require arbitrary hyperparameters, which makes it difficult to compare and trust (Bao & Schomaker, 2025; Danelian et al., 2023; Ring & Schomaker, 2025).

In this paper, we investigate *effective sample size* (ESS) as a diagnostic to evaluate the effect of positivity violations on IPTW estimators and give a precise definition for causal inference. ESS quantifies how many samples from a randomised experiment the IPTW observed data sample can represent, and hence provides an intuitive check for non-positivity. It provides a single, interpretable scalar summary, familiar to researchers, that can serve as an effective communication tool (Thomassen et al., 2024). While estimator and model assumptions must be considered, this approach handles multivariate overlap without introducing arbitrary choices. We discuss ESS's mathematical properties in relation to causal inference and link them to findings from the *importance sampling* and *survey sampling* literature (Kish, 1965; Kong, 1992). These fields provide a framework for formulating a definition of ESS suitable for causal inference. We provide a theoretical analysis of multiple interpretations of ESS in the context of positivity. We examine their implicit assumptions and highlight their differences. Finally, we demonstrate its strengths and weaknesses as a diagnostic tool using simulations. Our results offer practical guidance for applied researchers on using ESS to assess the reliability of IPTW estimates before drawing causal conclusions.

This paper contributes to the literature in three ways. First, we define ESS to quantify the estimability of causal effects and to fairly communicate the uncertainty of IPTW estimates. This includes cases with near-violations of stochastic positivity for some samples. Second, we provide simulation results to analytically show its strengths and limitations, which inform recommendations for empirical research. Finally, we show why conventional approaches are not generally suitable for causal inference and can underestimate the uncertainty of estimates.

The content of this paper is structured as follows. First, background knowledge of causal inference, relevant estimators and the positivity assumption is provided. Then, we provide an exact definition of ESS for causal inference. Third, we test the properties of ESS to demonstrate its strengths and limitations, and consider the effects of misspecification in propensity models. Finally, we discuss the findings, discuss limitations, and offer recommendations for practitioners before concluding.

II. Background

A. Causal Inference

In this paper, we follow the potential outcomes framework proposed by Neyman, as discussed by Rubin (2005), and formalised by Holland (1986). We consider a setting with observational data for a treatment (or 'intervention'), A, with possible values $a \in \mathcal{A}$, a list of observed confounders X, and an outcome Y. Causal relations between these variables, including confounding, can be characterised graphically using d-separation, as introduced by Pearl (2000). We focus our analysis on binary treatments $\mathcal{A} \equiv \{0, 1\}$ corresponding to no treatment (control) and treatment, respectively. We denote the potential outcome of treatment a on Y as Y^a. This corresponds to $Y \mid \mathrm{do}(a)$ for those more familiar with do-notation.

The value of interest, the estimand, is the *average treatment effect*, $\mathrm{ATE} = \mathbb{E}[Y^1] - \mathbb{E}[Y^0]$. Where $\mathbb{E}[Y^a]$ is the mean potential outcome if everyone received treatment level a. Intuitively, the goal is to estimate the average change in outcome if everyone were treated versus if no one were. It is also important to distinguish between the two types of estimands that we refer to. First, the quantity that we wish to estimate, the ATE, is our *causal estimand*. This value cannot be estimated directly as only one potential outcome can be observed per sample. The other potential outcome, for the treatment level that was not assigned, remains *counterfactual*. This is known as the *fundamental problem of causal inference* (Holland, 1986). In practice, we estimate the second type of estimand: the *statistical estimand*. If specific *identifying assumptions* hold, this corresponds to the causal estimand. There are other causal estimands that empirical researchers might wish to pursue, depending on research goals or if it is infeasible to satisfy the identifying assumptions with the given data. For example, the average treatment effect on the treated (ATT) or the conditional average treatment effect (CATE).

Randomised experiments, or randomised controlled trials (RCTs), are studies in which treatment A is assigned independent of any confounders X. We denote this as $A \perp\!\!\!\perp X$. This means that confounding is not a concern in expectation and that 'treatment' can be explicitly defined. Thus, with relatively few assumptions, we can estimate the average treatment effect. Our focus is on observational data. A common way to make causal inferences from observational data is to view it as a conditionally randomised experiment, in which treatment is assigned at random based on the confounders (Hernan & Robins, 2020, p. 27). Figure 1 highlights the difference between the observational and randomised settings.

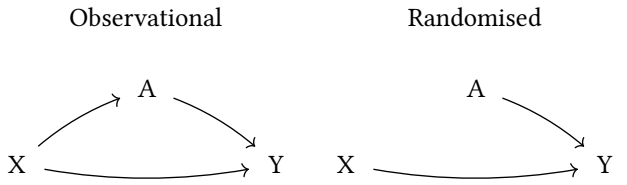

Fig. 1. Graphical representation of how we assume variables affect each other in observational versus randomised settings. The joint distribution of the observed data, or a conditionally randomised experiment, is as follows: $f(X, A, Y) = f(Y|A, X)\ f(A|X)\ f(X)$. X are covariates that confound treatment with probability density $f(X)$, A are possible treatments with probability mass $f(A|X)$, Y is a measured outcome with probability density $f(Y|A, X)$.

To consider an observational study as a conditionally randomised experiment, we require three identifying assumptions (Hernan & Robins, 2020). First, we need *consistency* in the treatment: the observed outcome for an individual who received treatment a is the same as their potential outcome under that treatment. Formally, $A_i = a \Rightarrow Y_i = Y_i^a$ for all individuals and all treatment levels. Intuitively, the 'treatment' must be well-defined and data must be recorded correctly. Second, we need *exchangeability*: that the potential outcomes are independent of treatment assignment conditional on measured covariates X. Mathematically, $Y^a \perp\!\!\!\perp A \mid X, \quad \forall a \in \mathcal{A}$. This assumption implies that there is no unmeasured confounding (or no omitted-variable bias). Third, there is the *positivity* assumption, which is the focus of this paper[1] :

Assumption - Positivity

For all covariate values x with positive density $f(x)$, and all treatment levels $a \in \mathcal{A}$,

$$f(X = x) > 0 \Rightarrow f(A = a \mid X = x) > 0 \tag{1}$$

where $f(A = a \mid X = x)$ is a probability mass function. This ensures that all treatment levels are possible for any covariate combination that occurs with positive probability density in the population.

This is also referred to as 'overlap' or 'common support' (D'Amour et al., 2021).

B. Deterministic versus Stochastic Positivity

A key distinction is that the definition of positivity can be split into two types: *deterministic* and *stochastic* positivity[2]. This was first discussed by Westreich & Cole (2010) and subsequently refined by Zivich et al. (2022) in a nonparametric setting. Deterministic positivity relates to the identifiability of the causal estimand and is defined by Eq. 1[3]. If violated, then the estimand is ill-defined.

Stochastic positivity requires having sufficient data and is a concern for the estimability of the causal estimand. It is defined as follows:

$$\hat{f}_n(A = a \mid X = x) > 0, \tag{2}$$

for all $a \in \mathcal{A}$, and $x = \{x_1, x_2, ..., x_n\}$, where $\hat{f}_n$ is an estimate based on n samples. This can be checked directly by computing the estimated propensity scores for each observed sample. That said, the estimated conditional probabilities do not need to be zero exactly for estimates to become unstable.

This distinction is relevant to later analysis. The diagnostics developed in this paper specifically focus on stochastic positivity. Hence, they focus on the estimateability of the problem given the data rather than on the identifiability of the causal estimand, which is a distinct concern requiring separate consideration.

C. Inverse Probability of Treatment Weighting

Inverse Probability of Treatment Weights (IPTW) is a weighting scheme that reweights samples to create a *pseudo-population* where treatment is independent of its confounders (Hernan & Robins, 2020, pp. 22-23, 163-164). Afterwards, a variety of estimators can be applied to the data, including model-based approaches. Crucially, these methods rely on the conditional probability mass function of treatment $f(A = a|X)$ being correctly specified. In practice, this is generally estimated from observational data.

This method is interesting for causal inference because it can yield a consistent and unbiased estimate of various causal estimands. This is possible if (1) the identifying assumptions hold, (2) $f(A = a|X)$ is known. The direct mean over the weighted samples estimator corresponds to the mean potential outcome under treatment a and can be used to estimate the ATE; it is called the Horvitz-Thompson (HT) estimator (Horvitz & Thompson, 1952). It is defined as follows:

$$\hat{\mathbb{E}}[Y^a] = \hat{\mathbb{E}}\left[\frac{\mathbb{1}(A = a)Y}{f(A = a \mid X)}\right] \tag{3}$$

In practice, this is a simple weighted sample average, $\frac{1}{n}\sum_{i=1}^{n} w_i\, \mathbb{1}(A_i = a)\, Y_i$. As the name suggests, the weights are inversely proportional to the conditional probability of treatment, i.e. $w_i \propto \hat{f}(A_i = a \mid X_i)^{-1}$.

Alternatively, the Hájek (or self-normalising) estimator can be used. It rescales all the weights proportional to the total weight. It is defined as:

$$\hat{\mathbb{E}}[Y^a] = \frac{\hat{\mathbb{E}}\left[\frac{\mathbb{1}(A=a)Y}{f(A=a \mid X)}\right]}{\hat{\mathbb{E}}\left[\frac{\mathbb{1}(A=a)}{f(A=a \mid X)}\right]} \tag{4}$$

This is also estimated with a simple weighted sample average where the weights are normalised, i.e. $\breve{w}_i = \frac{w_i}{\frac{1}{n}\sum_{j=1}^{n} w_j\, \mathbb{1}(A_j=a)}$. If positivity holds and the propensity score estimates are correct, this estimator is asymptotically unbiased (Hernan & Robins, 2020, p. 162) and consistent (Austin

[1] Formally, positivity can also be written as $\inf_{a \in \mathcal{A}} \Pr(A = a \mid X) > 0$, -a.e. This form is used by Petersen et al. (2012).

[2] Also referred to as 'structural' and 'random' positivity respectively.

[3] Zivich et al. (2022) define deterministic positivity as $\Pr(A = a \mid X = x) \geq \epsilon > 0$, $\forall a \in \mathcal{A}$, and x where $\Pr(x) > 0$ given an arbitrarily small constant ϵ.

& Stuart, 2015). Datta & Polson (2025), citing Särndal et al. (2003, p. 183), state that it is preferred over HT when (a) there is low outcome variance, (b) samples have equal inclusion probabilities, and the sample size is variable, or (c) when outcomes are negatively associated with sampling probability.

D. IPTW framed as Importance Sampling

Inverse Probability of Treatment Weighting (IPTW) can be reframed as an importance sampling method (Elvira et al., 2022; Matsouaka & Zhou, 2024; Shook-Sa & Hudgens, 2020). In importance sampling, the goal is to estimate an expectation $\mathbb{E}_p[g(Z)]$ of some target statistic g under a target distribution p by sampling a random variable Z from a proposal distribution f. The key idea is that:

$$\mathbb{E}_p[g(Z)] = \mathbb{E}_f\left[g(Z) \cdot \frac{p(Z)}{f(Z)}\right] \tag{5}$$

where $w(p, f) = \frac{p(Z)}{f(Z)}$ is known as the importance weights. Similarly, in IPTW, the goal is to estimate $\mathbb{E}_p[Y^a]$, where p is a distribution such that $A \perp\!\!\!\perp X$, but the observed samples come from the distribution f, where $A \not\perp\!\!\!\perp X$. The inverse probability weights serve as importance weights, reweighting the observed distribution toward the target distribution in which treatment is independent of covariates. For the IPTW, the random variable is the combination of the confounders, treatment, and outcome: $Z = (X, A, Y)$. Importance weights can be expressed as follows:

$$\begin{aligned} w(p, f) &= \frac{p(Z)}{f(Z)} = \frac{p(X, A, Y)}{f(X, A, Y)} \\ &= \frac{p(Y \mid a, X)p(X)}{f(Y|A, X)f(A|X)f(X)} \\ &= \frac{\mathbb{1}(A = a)}{f(A|X)} \end{aligned} \tag{6}$$

where we assume that $p(Y \mid a, X) = f(Y|A, X)\mathbb{1}(A = a)$ by consistency, and that $p(X) = f(X)$. Many statistical estimands can be transformed to fit this form. The mean potential outcome can be framed as an importance sampling problem with $g(Z) = \mathbb{1}(A = a)Y$. Likewise, the ATE can be expressed as importance sampling with $g(Y) = \mathbb{1}(A = 1)Y - \mathbb{1}(A = 0)Y$.

This is an important link because there is extensive literature on importance sampling. This includes research regarding poor overlap between the proposal and target distributions, which relates directly to empirical non-positivity. Diagnostics developed for importance sampling include ESS, which we adapt to the causal inference context.

E. Balance

It is important to distinguish between *positivity* and *balance*. These concepts are related but differ in key ways. Balance refers to the similarity in the confounder distributions across treatment groups within a sample (StataCorp LLC, 2021, p. 208; Stuart, 2010). Ho et al. (2007) point out that balance is a property of the observed sample, not of a hypothetical underlying distribution. This differs from positivity, which focuses on both observed and theoretical support of these distributions. In randomised experiments, balance holds in expectation by design. In observational data, this is not the case. Instead, methods such as IPTW can reweight samples to induce balance (Rosenbaum & Rubin, 1983).

Non-positivity can cause imbalance. For example, consider a situation in which the treated group corresponds to only a small subset of the observed population. In this case, there will be untreated individuals with $f(X) > 0$, for whom no comparable treated individual was observed. This could be a stochastic positivity violation, since $\hat{f}_n(A = 1 \mid X)$ might be zero, or a deterministic violation if $f(A = 1|X) = 0$. In either case, the observed confounder distributions in the observed samples differ between the groups.

Imbalance does not necessarily mean that positivity is violated. The confounder distributions can differ between the treated and untreated groups as long as their supports, the values the confounders can take on, overlap. Intuitively, each untreated individual in the data could have had a non-zero chance of receiving treatment, even though it was much more likely that they would remain untreated, and vice versa. As such, balance by itself cannot directly indicate whether positivity is violated. That said, it is a desirable property for the estimability of the causal estimand and can be used to evaluate the IPTW performance (Austin & Stuart, 2015).

F. Diagnostics

In general, balance diagnostics are more commonly discussed than the positivity assumption. This group of diagnostics is relevant because IPTW should balance the confounder distributions (Hernan & Robins, 2020, p. 197; Rosenbaum & Rubin, 1983). These diagnostics include comparisons of means and higher-order moments either numerically or graphically. Austin & Stuart (2015) review biomedical literature and provide an overview of relevant IPTW diagnostics. The most common recommendation is to report the (absolute) standardised mean difference (SMD), for example, in a side-by-side boxplot[4] to compare different models. For binary treatments we have $\text{SMD} = \sqrt{2}\big(\bar{X}_1 - \bar{X}_0\big)\big(s_1^2 + s_0^2\big)^{-\frac{1}{2}}$. Alternatively, the Kolmogorov-Smirnov test can be used for nonparametric balance testing. Stuart (2010) emphasises that diagnostics often focus on the balance of "lower-dimensional" summaries, such as a single confounder at a time. It is not feasible to directly examine the joint balance of all confounders. This is a risk because positivity violations can occur for a specific combination of confounder values.

For causal inference, several checks have been suggested to test specifically for positivity, though these are few. The most commonly used method is to report the mean-stabilised weight (MSW), which should be close to 1, as recommended

[4]Multiple sources also use love plots that show the absolute standardised mean difference (x-axis) for the adjusted and unadjusted samples per confounder (y-axis), see for example Austin (2009), Austin & Stuart (2015), Chesnaye et al. (2022), Zhou et al. (2022), and Liu et al. (2025)

by Hernan & Robins (2020). Additionally, we can examine the standard deviation, minimum and maximum of the weights, as extreme weights can be indicative of positivity violations or propensity model misspecification (Austin & Stuart, 2015, citing Cole & Hernan, 2008). Alternatively, it is common to examine the distribution of propensity scores of the observed samples (Liu et al., 2025; McCaffrey et al., 2013).

The second, more rigorous option is to use a parametric bootstrap method proposed by Petersen et al. (2012). By repeatedly resampling from the observed data under a fitted propensity score model, one can obtain an optimistic estimate of the bias due to positivity violations. The problems are that it is computationally expensive and requires both care and expert knowledge during the implementation. Ring & Schomaker (2025) also point out that it is estimator-specific and therefore captures other forms of sparsity.

Third, King & Zeng (2006) propose a fully data-centric method to measure the degree of extrapolation. They suggest comparing the distance between a point and a convex hull over the confounders, the smallest polygon in the confounder space that contains all samples of a specific treatment level. Then, the Gower distance is used to determine if a sample is outside the convex hull. Strong extrapolation suggests positivity violations and increased model-dependence, if applicable.

A fourth method, by Danelian et al. (2023), uses regression trees to identify in-sample positivity violations (PoRT). The regression tree approach enables them to identify regions of the covariate space where treatment assignment is near-deterministic. This results in interpretable regions of (near) positivity violations, although it is restricted to categorical treatment variables.

Ring & Schomaker (2025) recently proposed *effective data points* (EDP), which uses kernel density estimation to measure the number of observations that effectively contribute to the estimation even for continuous or Modified Treatment Policies (MTPs). It includes a solely data-centric and estimator-specific approach. A downside is that this method requires the researcher to specify a kernel, which is an arbitrary hyperparameter.

Finally, Bao & Schomaker (2025) proposed the *non-overlap ratio*, which quantifies the proportion of the covariate distribution that has inadequate overlap between treatment groups. It is based on the concept of highest density regions (HDRs). Similar to Ring & Schomaker (2025), a downside is that it requires the ad hoc setting of a threshold hyperparameter.

These diagnostic approaches are complementary, with each offering different insights into positivity violations. Some require arbitrary hyperparameters, can only be used in certain settings, require expert knowledge, are computationally expensive, etc. In the next section we investigate whether effective sample size can address some of these limitations.

III. Analysis

A. Effective Sample Size

The term *effective sample size* has been used in multiple contexts. We focus on the definition from the survey and importance sampling literature; see, for example, Kish (1965) and Kong (1992). This should not be confused with the use in Markov Chain Monte Carlo (MCMC), where it indicates the number of independent samples in a chain based on autocorrelation. Nor the effective size mentioned in kernel density estimation literature, mentioned for example in Ring & Schomaker (2025). The definition we use is as follows:

$$\text{ESS} = \frac{n\text{Var}_p\big(\hat{\psi}\big)}{\text{Var}_f\big(\tilde{\psi}\big)} \tag{7}$$

where n is the actual sample size, $\hat{\psi}$ represents a simple estimator on the target distribution p, and $\tilde{\psi}$ is the (complex) weighted estimator on the observed data distribution f. The key idea is that it compares the efficiency between two different distributions.

In our case, we focus on semi-parametric[5] estimators. We consider simple (weighted) average estimators, such that $\hat{\psi} = \frac{1}{n}\sum_{i=1}^{n}\psi_i$, and that $\tilde{\psi} = \frac{1}{n}\sum_{i=1}^{n} w_i\psi_i$ is the IPTW estimator, where $\psi_i = Y_i(2A_i - 1)$ for the ATE. It is also possible to consider model-based estimators; this is a practical extension but adds complexity.

Originally, ESS originated in the survey sampling literature. It was introduced by Kish (1965), although his focus was on the *design effect* (DEFF)[6] (Shook-Sa & Hudgens, 2020). The idea of the DEFF is to compare a simple design for a randomised parallel study with a complex design.

$$\text{DEFF} = \frac{\text{Var}\big(\tilde{\psi}_{\text{complex}}\big)}{\text{Var}\big(\hat{\psi}_{\text{simple}}\big)} \tag{8}$$

While several earlier works examined ratios of variances, Kish (1965) popularised the concept (Kish, 1995; Park & Lee, 2004). Notably, for this paper, Kish elaborated that $\frac{n}{\text{DEFF}}$ is an "estimate of the effective n". However, the goal of survey design differs from ours. It seeks to perform power calculations and determine how many samples to collect before collecting the data. In survey design, variances are not estimated from the (not yet collected) dataset but from past surveys or pilot studies. Regardless, it highlights the effect that a non-randomised sampling procedure has on the uncertainty of statistical inferences.

The importance sampling literature has repurposed the definition of ESS. The main goal of importance sampling is to "reduce variance for very skewed sampling prob-

[5]Technically, these estimators can be considered nonparametric if all confounders are categorical and a nonparametric propensity model is used.

[6]The design effect has also been referred to as the "relative efficiency" and "variance inflation factor" (VIF) (Hsieh et al., 2003).

lems" (Owen, 2013, p. 35). This is not the same as causal analysis, which aims for an unbiased and precise estimate of a causal effect (Hernán, 2022). The definition we use stems from importance sampling literature, for example, Elvira et al. (2022) and Martino et al. (2017). However, it contains variances from unknown distributions, which must therefore be estimated. This is where differences in the goals of importance sampling and causal analysis become relevant.

Why do we consider ESS as a diagnostic tool when detecting positivity violations? In IPTW, we weight the data to estimate effects as if they were randomised, which requires the three identifying assumptions. Assuming that we have tested and are confident in consistency and exchangeability, then the remaining concern is the positivity assumption. Stochastic positivity concerns the estimability of the statistical estimand, and weighting reduces the it. The ESS quantifies the efficiency of using weights to mimic a randomised experiment. Low ESS suggests that precision is lost when mimicking a randomised experiment and that stochastic positivity may be violated.

B. Conventional ESS

Usually, based on importance sampling theory, ESS is approximated using the following formula (Elvira et al., 2022; Owen, 2013):

$$\mathrm{ESS}_{\mathrm{conv}} = \frac{\left(\sum_{i=1}^{n} w_i\right)^2}{\sum_{i=1}^{n} w_i^2} \tag{9}$$

Intuitively, this could be relevant for detecting positivity violations. For IPTW estimators, as the name suggests, weights are inversely proportional to the probability of treatment conditional on the confounders. If positivity is violated, the propensity nears zero, and the weights will become very large. As a result, the ESS would decrease, since the denominator increases more rapidly than the numerator.

As Zhang et al. (2024) state, this definition of ESS makes four assumptions:

1) Estimand is a population mean	$\mathbb{E}[\psi]$
2) Homoskedastic outcome data	$\mathrm{Var}(\psi_i) = \sigma_\psi^2,\ \forall i$
3) Independent outcome samples	$\mathrm{Cov}(\psi_i, \psi_j) = 0,\ \forall i \neq j$
4) Weights are constants	$\mathrm{Var}(w_i\psi_i) = w_i^2\ \mathrm{Var}(\psi_i)$

The first assumption is generally satisfied. For causal inference, we mainly want to compute expected values. These can be estimated with a sample average of the target function ψ. Second, whether homoskedasticity and independence are reasonable assumptions depends on the empirical research that is conducted. It should mirror the assumptions made by the estimator that is used. This is the case for the HT (Eq. 3) and Hájek (Eq. 4) estimators we are considering, but not generally. Finally, the assumption that the weights are constant is concerning for causal inference because we assume that the estimated weights and the outcome both depend on the confounders.

The assumption that weights are constant masks a more fundamental problem. The conventional ESS (Eq. 9) is equivalent to a comparison of two estimands on the observed data distribution. In other words, it indirectly assumes that the target distribution is the same as the observed distribution, i.e. that $p = f$, when deriving the ESS, or at least that $\mathrm{Var}_p(\psi_i) = \mathrm{Var}_f(\psi_i) = \sigma_\psi^2$. To highlight this, we can examine the ESS from Eq. 7 for the Hájek estimator:

$$\begin{aligned}
\widehat{\mathrm{ESS}}_{\mathrm{Hájek}} &= n\frac{\mathrm{Var}_p\left(\frac{1}{n}\sum_{i=1}^{n}\psi_i\right)}{\mathrm{Var}_f\left(\frac{\sum_{i=1}^{n} w_i\psi_i}{\sum_{i=1}^{n} w_i}\right)} \\
&= n\frac{\frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}_p(\psi_i)}{\frac{1}{\left(\sum_{i=1}^{n} w_i\right)^2}\sum_{i=1}^{n}\mathrm{Var}_f(w_i\psi_i)} \\
&= \frac{\frac{1}{n}\left(\sum_{i=1}^{n} w_i\right)^2\left(n\sigma_\psi^2\right)}{\sum_{i=1}^{n} w_i^2\sigma_\psi^2} \\
&= \frac{\left(\sum_{i=1}^{n} w_i\right)^2}{\sum_{i=1}^{n} w_i^2}
\end{aligned} \tag{10}$$

This corresponds to the conventional ESS formula (see Eq. 9). The first simplification assumes that the samples are independent, and the second that the weights are constant and that the variance of ψ_i is homoskedastic (and equal) across both distributions. A similar calculation of the HT estimator leads to a slightly different result:

$$\begin{aligned}
\widehat{\mathrm{ESS}}_{\mathrm{HT}} &= n\frac{\mathrm{Var}_p\left(\frac{1}{n}\sum_{i=1}^{n}\psi_i\right)}{\mathrm{Var}_f\left(\frac{1}{n}\sum_{i=1}^{n} w_i\psi_i\right)} \\
&= n\frac{\frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}_p(\psi_i)}{\frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}_f(w_i\psi_i)} \\
&= n\frac{\left(n\sigma_\psi^2\right)}{\sum_{i=1}^{n} w_i^2\sigma_\psi^2} \\
&= \frac{n^2}{\sum_{i=1}^{n} w_i^2}
\end{aligned} \tag{11}$$

The issue is that f represents the observed data distribution and not the target distribution p. What, then, are we comparing with conventional ESS? In causal inference, unbiased estimation of the causal effect from the observed data distribution f without weighting the samples is generally not possible because $A \not\perp\!\!\!\perp X$. A simple random sample will therefore not yield a valid causal effect estimate from our observed data. Instead, we want to know how our estimator performs relative to a setting where the stochastic positivity assumption holds, such as a randomised experiment. It is contradictory to assume that these two settings yield the same variance, since we use IPTW because the distributions differ in the first place.

Given the limitations of the conventional ESS, we propose focusing on comparing estimators under the appropriate

distributions when discussing positivity. Instead, the idea is that we compare the relative efficiency of estimating the variance of a simple estimator defined on the target distribution with the IPTW estimator on the observed data. To be specific, the target is a distribution where $A \perp\!\!\!\perp X$:

$$\begin{aligned} &\text{Target} && p(Y, A, X) = p(Y \mid a, X)\, p(X) \\ &\text{Observed} && f(Y, A, X) = f(Y \mid A, X)\, f(A \mid X)\, f(X) \end{aligned}$$

The survey design literature worked on the premise that the ESS would be computed before the main data collection to decide whether the research is feasible and how many samples to collect. That said, we are interested in the situation where we already have access to the data, even though the target distribution p remains unknown. Consequently, we have to estimate the variance of our simple estimator under the targeted distribution using the observed data.

C. Estimating Targeted ESS

For causal inference, we recommend that researchers focus on comparing the right distributions. To distinguish the various forms of ESS, we consider a *targeted ESS* to be any effective sample size that (1) adheres to Eq. 7, (2) that observes data from a distribution where treatment is confounded, and (3) that targetes a distribution where treatment is fixed, independent of X. This is a group of diagnostics rather than a single closed form formula because targeted ESS depends on the estimator. To estimate a targeted ESS, we need to account for how the estimator works, including any relevant assumptions.

Targeted ESS can be estimated directly for the IPTW estimators under independence and homoskedasticity. The idea is to modify the ESS estimand to be in terms of statistics on the observed data, such as an expected value or variance, which can be estimated directly. For both the HT and Hájek estimators, the main question is how to estimate the numerator of Eq. 7, since the denominator is simply the variance of IPTW estimates. Given homoskedasticity and independence, we can create an estimand on the observed data as follows:

$$\begin{aligned} \operatorname{Var}_p\left(\frac{1}{n}\sum_{i=1}^{n}\psi_i\right) &= \frac{1}{n^2}\left(\sum_{i=1}^{n}\operatorname{Var}(\psi_i)\right) \\ &= \frac{1}{n^2}(n\ \operatorname{Var}(\Psi)) \\ &= \frac{1}{n}\big(\mathbb{E}_p[\Psi^2] - \mathbb{E}_p[\Psi]^2\big) \\ &= \frac{1}{n}\big(\mathbb{E}_f[\Psi^2 W] - \mu^2\big) \end{aligned} \tag{12}$$

where $W = w(X)$ and $\mu = \mathbb{E}_f[\Psi W]$. The first step assumes the independence between samples. The second step assumes homoskedasticity, $\operatorname{Var}(\psi_i) = \operatorname{Var}(\Psi)$. Recall that Ψ is a functional transformation of Y. In the last two steps, we rewrite the expression and use the weights to convert the expectation to be over the observed distribution f instead of p. We should also be careful here and note that this requires that $\operatorname{supp}(p) \subseteq \operatorname{supp}(f)$, which is the positivity. This also means that this diagnostic will only capture near-violations under this design, as will be evident in Section IV. Similarly, we can apply the same steps to the HT estimator.:

$$\mathrm{ESS}_{\mathrm{HT}} = n\frac{\mathbb{E}_f[\Psi^2 W] - \mu^2}{\operatorname{Var}_f(\Psi W)} \tag{13}$$

We note that the $\frac{1}{n}$ term from Eq. 12 is lost because it is also present in the denominator, where $\operatorname{Var}_f\big(\frac{1}{n}\sum_{i=1}^{n}\Psi_i W_i\big) = \frac{1}{n}\operatorname{Var}_f(\Psi W)$. A similar direct estimand can be derived for the Hájek estimator.

$$\mathrm{ESS}_{\text{Hájek}} = n\frac{\mathbb{E}_f\big[\Psi^2 \breve{W}\big] - \mu^2}{\operatorname{Var}_f\big(\Psi \breve{W}\big)} \tag{14}$$

where $\breve{W}$ are the normalised weights. There are different ways to estimate ESS, and the assumptions of independence and homoskedasticity are not strictly necessary. However, this requires more specific domain knowledge regarding the causal estimand. The way ESS can and should be estimated depends on the estimator.

D. Link to the Conventional ESS

Where did the conventional ESS formula come from? To answer this question, we examine a different approximation of the ESS from importance sampling literature. Kong (1992) developed a method to estimate the ESS for the Hájek estimator. Her method was later extended to the conventional ESS formula. She used delta approximations, i.e., Taylor series approximations of the $\operatorname{Var}_f\big(\tilde{\Psi}\big)$ at the expected value of Ψ, and found that ESS could be simplified to the following expression.

$$\widehat{\mathrm{ESS}}_{\mathrm{Kong}} = \frac{n}{1 + \operatorname{Var}_f(W)} \tag{15}$$

If $\operatorname{Var}_f(W)$ can be identified up to a constant, then is one can get an unbiased estimate of the propensity scores. The main reason the delta-approximated version is commonly preferred is that it relies only on the weights, not the scale of ψ. This means that it can be used to compare the efficiency of different ψ, and makes it easier to compute. This approach is not possible for the HT estimator.

Kong notes that the $\operatorname{Var}_f(W)$ can be estimated using the sample variance of standardised weights. This leads to the definition $\mathrm{ESS}_{\mathrm{cv}} = n\big(1 + \mathrm{cv}^2(W)\big)^{-1}$, where $\mathrm{cv}(W) = \bar{w}^{-1}\sqrt{(n-1)^{-1}\sum_{i=1}^{n}(w_i - \bar{w})^2}$ is the coefficient of variation created by Liu & Chen (1995). This suggestion is made to ensure that ESS does not depend on the scale of the weights and the target distribution (Martino et al., 2017).

It is known that this definition, using the coefficient of variation, is equivalent to the conventional ESS (Elvira et al., 2022). See Martino et al. (2017) for a demonstration. Comparably, this also means that the delta-approximation version of ESS for the Hájek estimator is equal to the conventional definition of ESS for the HT estimator.

$$\begin{aligned}\mathrm{ESS}_{\text{Hájek}} &= \frac{n^2}{1+\mathrm{Var}_f(W)} \\ &= \frac{n^2}{1+\left(\mathbb{E}_f[W^2]-\mathbb{E}_f[W]^2\right)} \\ &= \frac{n^2}{\mathbb{E}_f[W^2]}\end{aligned} \tag{16}$$

which can be estimated with $\frac{n^2}{\sum_{i=1}^n w_i^2}$.

This shows where the difference between the conventional ESS and the targeted ESS comes from. First, we apply the delta method approximation. This makes the ESS estimates independent of the estimand Ψ. Second, the assumption that the mean weight is one: $\mathbb{E}_f[W]=1$. It was already invoked in the delta approximation, but used again to derive the conventional ESS. The problem is that this holds only for the IPT weights of the mean potential outcome (Hernan & Robins, 2020, p. 163), not for the ATE in general. The exception is that the Hájek estimator can multiply the weights by a factor[7], such as $\frac{1}{2}$ or $\Pr(A=a)$, to make this true without necessarily changing the estimand. Third, a second normalisation step was used to estimate the weights. These steps combined resulted in a diagnostic that essentially ignores the difference in the observed and targeted distributions and is not suitable for causal inference.

IV. Simulations

A. Goals

In this section, we aim to demonstrate how effective sample size behaves in practice through simulations. First, we consider scenarios in which targeted ESS, estimated using direct sample average estimators of Eq. 13 and Eq. 14, works or does not, to highlight its strengths and weaknesses. Second, we examine how targeted ESS reacts to separate concerns for estimability unrelated to positivity, which may be relevant to researchers. Third, we consider the effect of a misspecified propensity model. Identifying the correct propensity model is a core assumption for IPTW to work, but it is not guaranteed in practice. It is generally unknown and has to be estimated.

The general setup of these simulations will compare relative effective sample sizes ($\frac{\mathrm{ESS}}{n}$) for the average treatment effect, $\mathbb{E}[Y^1-Y^0]$. This means that $\psi_i = A_iY_i-(1-A_i)Y_i = Y_i(2A_i-1)$. Each simulation involves setting up a *data-generating process* (DGP) that creates samples from a specified distribution over confounders, treatment and outcomes. Given two treatment levels with five normally distributed confounders, we generate 500 samples and replicate the simulation 100 times. For all simulations, we assume that the other causal assumptions are upheld. In particular, this means that there is no unmeasured confounding. Unless otherwise specified, the ESS estimates have access to the true propensity scores from the DGP; we assume that a researcher accurately identified the true propensities. Plots show points representing the mean across the 100 replications, with corresponding standard error bars. Each simulation is designed to highlight a particular aspect of effective sample size.

B. Targeted ESS and Positivity

The main question we want to answer is whether targeted ESS is a good diagnostic for detecting positivity violations. First, we examine the effects of deterministic positivity and show that targeted ESS is not a good diagnostic in such a situation. Second, we consider stochastic positivity. Specifically, we investigate how the distribution of propensity scores affects the targeted ESS. More extreme propensity scores, i.e., those close to either 0 or 1, should make it more difficult to estimate the ATE. We simulate cases in which it would become increasingly difficult to trust causal estimates to verify how targeted ESS responds.

a) *Deterministic Positivity:*
Targeted ESS cannot capture deterministic positivity violations. We can create a simple thought experiment to show this. Consider a data-generating process in which the treated and control groups are completely distinct and have distinct supports. Then, a correctly estimated propensity model estimates a conditional probability of treatment of 1 for the treated samples. Similarly, it will predict a conditional probability of receiving no treatment of one for the control samples. Mathematically, $\hat{f}_n(A|X) = f(A|X) = 1$ for all observed A. This means that all weights will be set to 1 for the HT estimator. As a result, the targeted ESS from Eq. 13 reduces to n. For the Hájek estimator, the weights instead become $\frac{1}{\widehat{\Pr}_n(A=a)}$. This results in a deceptive ESS value since even in the setting with no overlap, the ESS would not be zero. Optimally, a deterministic positivity diagnostic would conclude that the causal question cannot be identified with the data, but the targeted ESS does not. This example shows that targeted ESS is not reliable in these situations because, if deterministic positivity is violated, we cannot identify the causal effect.

b) *Stochastic Positivity:*
To show that effective sample size captures the estimability under different degrees of stochastic non-positivity, we consider balance. We set up the confounder distribution as a mixture of two Gaussians, one for the treated and one for the control group. The marginal probability of treatment is the same for each treatment level: $\Pr(A=1)=\Pr(A=0)=0.5$, and the outcome Y is a linear combination of the confounders, treatment, an intercept, and noise. In the simulation, we vary the distance between these two distributions, i.e. $\|\mu_1-\mu_0\|$. Since we control the DGP, there is no hidden confounding, and any imbalance is caused by the treatment allocation. A two-dimensional projection of how the confounder distributions per treatment level are affected by the shift in means is shown in Figure 2. This shift in distribution simulates increasing difficulties of stochastic non-positivity, because more samples will gradually have their estimated conditional probability of treatment approach

[7]This is referred to as a *projection function.*

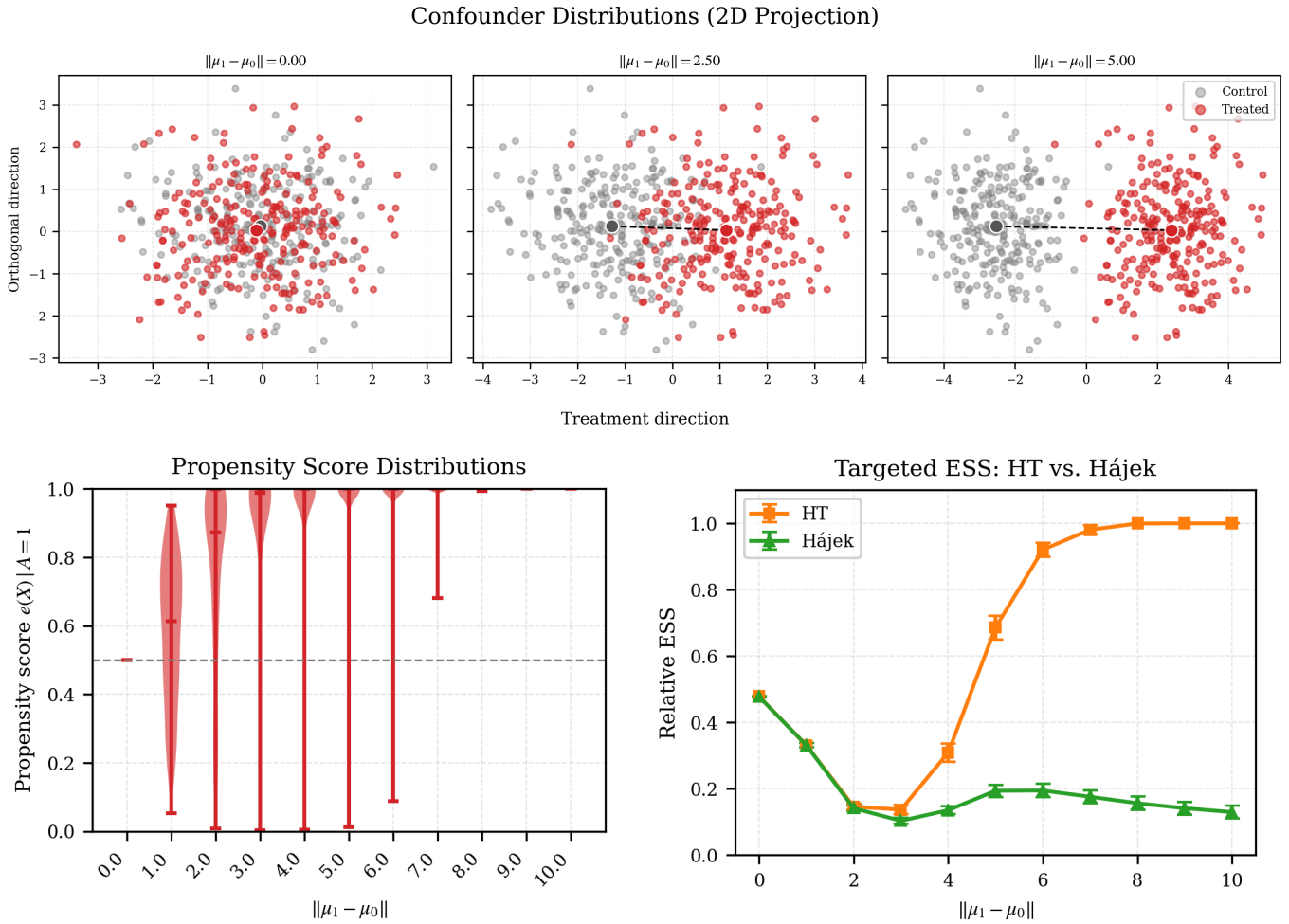

Fig. 2. Top: Vizualizes the confounder distributions per treatment level projected onto two dimensions, the direction of the mean shift and the orthogonal component via Gram-Schmidt, across degrees of imbalance created by a distribution shift. Bottom-Left: Shows the propensity score distribution for the observed treated samples. The control samples will have similar distributions for the conditional probability of receiving no treatment. Bottom-Right: Shows the targeted ESS for different degrees of imbalance.

0. This is illustrated by the propensity distribution plot in Figure 2. In this simulation, we focus on stochastic non-positivity, as the design ensures deterministic positivity.

The simulation results are shown in the bottom-right plot of Figure 2. We can focus on two different regions in the plot. These regions correspond to situations where stochastic positivity is (1) a significant concern for most samples, or (2) when it is potentially a concern for some of the samples.

First, we consider the region where stochastic positivity is a concern for most samples. In the plot, this corresponds to the region where the mean shift is roughly greater than 2. There, the expected propensity score for the treated is nearly 1, a clear signal that stochastic positivity is a problem. In this region, targeted ESS still provides results for both the HT and Hájeks estimator that should not be trusted. The Horvitz-Thompson's ESS actually increases, and the Hájek's ESS stays deceptively low. When the mean shift exceeds 7, stochastic positivity is a concern for all samples. The expected propensity score is not only near 1, but there are no observed samples with a propensity score near 0.5. In this case, the HT estimator's relative ESS becomes 1 and the Hájek estimator's ESS becomes deceptively low, for the same reasons explained when discussing deterministic violations. This shows a minor weakness of the targeted ESS: it is unreliable when stochastic non-positivity is very strong. It is only a minor weakness because it is apparent when the violations are this severe.

Second, we have the region where stochastic positivity may be a concern for some samples. This corresponds to the region of the graph where the mean shift is between 0 and roughly 2. In this region, as the imbalance increases, some propensity scores become more extreme, and the targeted ESS decreases. This is a desirable property of the ESS and happens for both the HT and Hájek estimators. They show very similar ESS values because the conditional probability of treatment is very stable for the samples. Intuitively, a poorer balance indicates less information regarding the causal estimand, which is captured by the ESS. This is the more interesting and relevant region because we can still have some trust in the causal estimates. Targeted ESS provides a way to assess the estimability of the causal effect, assuming it can be identified.

Overall, this shows that targeted ESS is a decent measure of the estimability of the causal effect. There are some caveats. It does not detect deterministic positivity violations or widespread stochastic positivity violations. Essentially, it optimistically estimates the efficiency given that the target distribution can be identified with our finite sample. Whether it can be identified with the finite sample is an important question that researchers should discuss. Both the HT and Hájek estimators are unreliable when this is not the case, and are equally unhelpful. That said, if the ESS is low, then it is a good indicator that we cannot be certain whether the causal effect is estimated correctly.

C. Estimability Beyond Positivity

Beyond its purpose of detecting positivity violations, the targeted ESS can be viewed as a general measure of the estimability of the causal estimand, provided the identifying assumptions hold. To demonstrate this, we consider two scenarios that make estimation of causal effects more difficult without affecting the positivity assumption. The examples illustrate how common differences between observed datasets affect targeted ESS.

a) *Prevalence:*
The first consideration we make with regard to estimabilty, which does not violate positivity, is the prevalence of treatment. For this, we use the same DGP as described for Section IV.B.b, but we keep the difference in means at 1. Instead, we vary the prevalence of treatment, i.e. the marginal probability of receiving treatment $\Pr(A = 1) = 1 - \Pr(A = 0)$. The results are shown in Figure 3. Intuitively, fewer samples make it harder to estimate the causal effect. This is captured by the targeted ESS, which decreases when there are fewer samples in either group. As desired, the targeted ESS is symmetric for prevalences above or below 0.5, because it considers control and treated samples equally important for estimating the ATE. This is a relevant result because most observational studies will not have equal amounts of treated and control samples. Targeted ESS reflects on the impact that this has on the estimates.

b) *Propensity Score Variance:*
The second form of increased estimability difficulty is the variance of the propensity scores. To test the effects of variance, we alter the spread of the propensity scores. This is done by directly modifying α in the following DGP:

$$\begin{aligned} X &\sim N(0, I) \\ A \mid X &\sim Ber(\text{expit}(\alpha_0 + \alpha^T X)) \\ Y \mid A, X &\sim N(\beta_0 + \tau A + \beta^T X, I) \end{aligned} \tag{17}$$

The results are shown in Figure 4. The simulation shows that targeted ESS for both the HT and Hájek estimators decreases with increasing variance in the conditional probabilities. This is favourable behaviour, as the target's estimability depends on the propensity estimates. If these are imprecise, then it should be more difficult to estimate our causal effect.

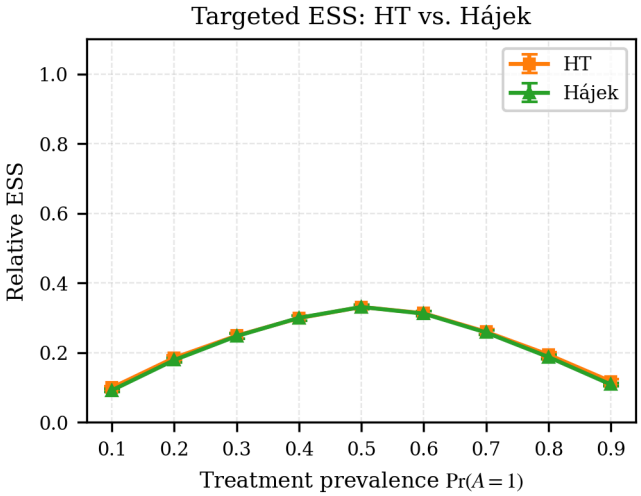

Fig. 3. Shows the effects of different prevalence on the estimated targeted ESS for the HT and Hájek estimators.

D. Propensity Model Misspecification

A significant problem is that the propensity (model) must be estimated and is typically unknown in observational data. It is not certain that the propensity scores can be estimated without bias. This raises to the question: Is ESS robust to propensity-score misspecification? For these simulations, the estimated propensity scores no longer correspond to the true propensity scores defined by the DGP. Instead, we estimate them from the generated data using a simple logistic regression model. This allows us to examine the effects of structural bias and misspecification for the propensity model.

a) *Bias:*
To simulate bias in the weights, we simply alter the estimated propensity scores. This is done by shifting the estimated propensity scores on the logit scale after fitting the logistic regression estimator. We then consider the absolute

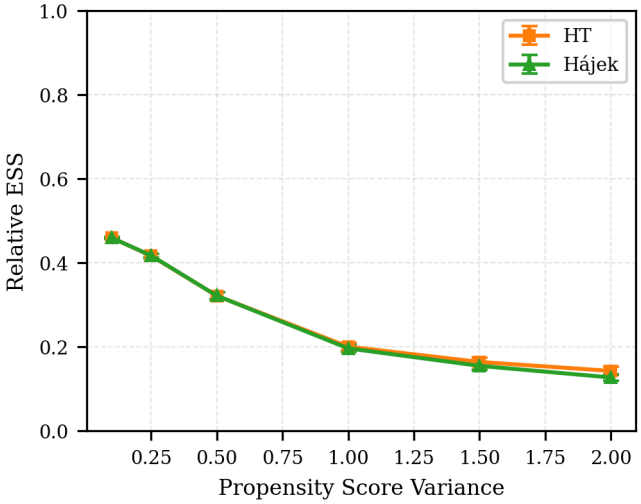

Fig. 4. Shows the effects of increased variance of the DGP on the estimated targeted ESS. Larger values mean more misspecification.

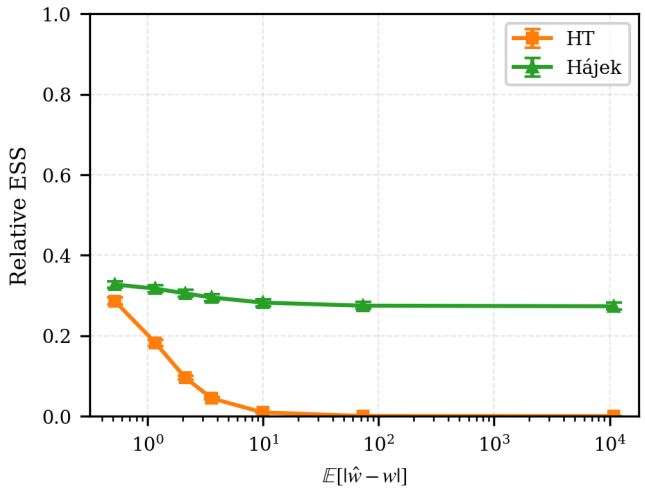

Fig. 5. Shows the effects of a bias in propensity score estimation on the estimated targeted ESS. Larger values mean more misspecification.

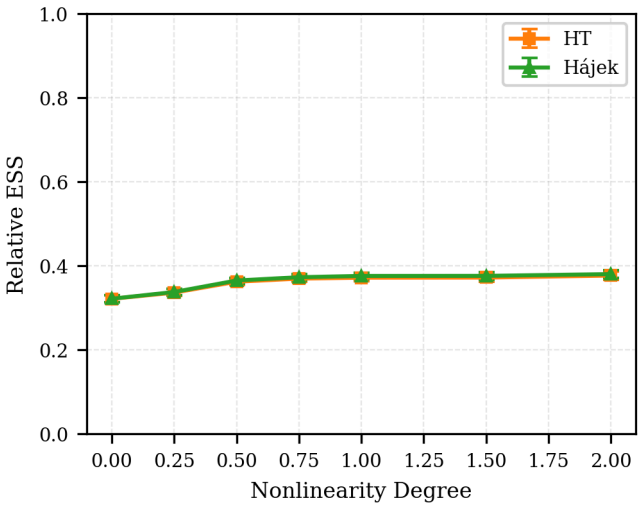

Fig. 6. Shows the effects of stronger model misspecification of the DGP on the estimated targeted ESS. To be specific, the influence of the interaction and quadratic term on the actual propensity score. Larger values mean more misspecification.

bias in the estimated weights to make the results more interpretable with respect to propensity score bias. For these simulations, we use the same DGP as in Section IV.C.b (Eq. 17). The results are shown in Figure 5. First, the Hájek estimator's ESS stays relatively the same, independent of the bias. It completely misses the shift in the estimated propensity score. This simulation oversimplifies how bias works in practice, but it further highlights that self-normalisation in the Hájek makes it difficult to trust the estimated ESS values. Second, the HT estimator's targeted ESS drops to zero. This happens because the weights become relatively more extreme. Less extreme weights may increase the ESS. This is favourable from the perspective of estimability, but it does not address positivity. The takeaway is that targeted ESS cannot distinguish between biased propensity scores and non-positivity indicated by the weights.

b) *Model specification:*

To test the effects of propensity model misspecification, we use a new DGP where the functional form of the propensity score does not match that of the propensity estimator. The true propensity score will include an interaction and a second-order exponential term with the confounders, which are controlled as independent variables. The estimated propensity score will remain a simple logistic regression that does not capture the full complexity. Figure 6 shows that targeted ESS is unable to detect strong model misspecification. Neither the HT nor the Hájek estimator's targeted ESS respond to the increased non-linearity in the true propensity scores.

To summarise, targeted ESS cannot diagnose propensity model misspecification. It answers a conditional question: given the estimated weights, how much information does the weighted sample contain about the interventional distribution that, under the identifying assumptions, lets us estimate the causal estimand? First, structural bias did not significantly impact the Hájek estimator's targeted ESS. While the HT estimator's ESS decreased, this can be explained by the weights becoming more extreme across all samples. Second, the functional misspecification of the propensity score barely affected the ESS. That said, it is not the ESS's main goal to determine whether the propensity score is correctly specified. Rather, its goal is to determine whether the targeted estimand can be estimated appropriately. Under substantial propensity model misspecification, neither the ESS nor the IPTW estimator will identify the targeted estimand, i.e. the interventional distribution. This can be a significant problem as the causal question will not be answered by the estimates. This emphasises that practitioners should validate their propensity model through separate tests before estimating the ESS.

V. Discussion

A. Interpretation and Limitations

Targeted ESS describes the estimability of the causal effect. The simulations in Section IV show that it reliably decreases as estimability worsens and when stochastic positivity is a concern for some samples. This includes, for example, when the unweighted data is imbalanced within reason or when the propensity score variance is high. This is a strength of targeted ESS, since in practice our effect estimates rely on the precision of the propensity score estimates. Similarly, it reacts to reduced estimability due to unequal marginal probabilities of treatment. It provides an interpretable value that puts the estimability of the causal question in perspective, given the observed data and the IPTW estimator. These are desirable aspects of the targeted ESS, which make it a useful diagnostic to consult. That said, there are limitations to what targeted ESS can do.

The most significant concern is that the targeted ESS requires the same change in probability measure as the IPTW estimator, which in turn requires positivity. In practice, this means that ESS does not highlight deterministic positivity

violations or strong stochastic violations. When positivity fails, the weights become unreliable, as does any diagnostic derived from them. However, ESS captures how some near-violations of stochastic positivity reduce the estimability of the causal effects, given that the identifying assumptions hold. Targeted ESS's main use is as a tool to understand and communicate the uncertainty of one's inferences.

Another concern is that propensity scores are generally unknown in observational data. This questions whether targeted ESS should be used given possible estimation errors in the propensity scores. The simulations showed that this concern is warranted. Targeted ESS did not meaningfully capture bias, nor does it show a practitioner whether the propensity model is incorrect. It remains important for researchers to consistently estimate the propensity scores and consider multiple model specifications if necessary. Propensity model misspecification is a general concern for IPTW methods, not specific to ESS. Future work can consider how effective sample size can be extended to different causal estimators that are less sensitive to propensity model misspecification, such as doubly robust methods, e.g. Augmented IPW (AIPW).

There are also several practical concerns regarding targeted ESS that we have not discussed so far. First, targeted ESS is a diagnostic based on finite data. The estimator and ESS are fitted on the estimated data distribution, $\hat{f}_n$. As the sample size grows, this converges to f, and ESS can be interpreted as the 'true' efficiency.[8]. Furthermore, the targeted ESS requires the estimation of variances, which can be unstable. This is an inherent problem for a diagnostic based on variance. A practitioner should first consider whether the confounders' distribution in their observed data is suitable for their causal question. For the ATE, this means that it needs to be close to f. If that is the case, then one can proceed with estimating the targeted ESS. However, we recommend robust methods, such as bootstrap sampling, to account for the confidence intervals of the targeted ESS.

Second, targeted ESS does not consider the bias of the estimated weights, nor of an outcome model if used. Several authors (Elvira et al., 2022; Martino et al., 2017) suggested using the mean squared error (MSE) instead of the variance, which is equal to the variance plus the bias squared. The main issue is that the truth is unknown in causal inference. One possible solution is to use the ETA.bias from Petersen et al. (2012) to obtain an optimistic estimate of the MSE. Using MSE also means that the ESS-based diagnostics loses its interpretation as a relative efficiency. That said, it is a potential avenue for further research.

Finally, it is important to consider the potential use of targeted ESS in practice. While targeted ESS is estimator-specific, which makes it more difficult to specify for researchers, it can still be systematically applied in empirical research through software. As mentioned, design effects or variance inflation factors are different expressions from ESS but describe the same type of problem while being estimator-specific. That said, they have been well documented for various sampling designs so that they can be used consistently in empirical research, as was done by Hsieh et al. (2003). Similarly, targeted ESS can also be directly incorporated into software packages such as svy in Stata or R's survey package .

In summary, targeted ESS can capture the loss in estimability from the IPT weights. It is most informative when positivity is not the main concern for most samples and the propensity model is correctly specified. Its main limitations are that it cannot detect deterministic or severe stochastic violations and that it does not capture bias. These limitations motivate the following section.

B. Recommendations

Based on the analysis and simulations, we can make several recommendations for practitioners who aim to use IPTW estimators. Before discussing these, we need to reiterate that we cannot recommend using conventional ESS with IPTW in observational data. First, it requires homoskedasticity and independence, which limits the situations where it could be applied. More importantly, it assumes that the target and the observed distributions are identical, which contradicts the idea behind IPTW. We warn practitioners that the estimated conventional ESS may not be interpretable due to this distributional difference. There are several steps that we can recommend.

First, we recommend that practitioners use a rigorous selection process for the propensity score model. Incorrect functional forms can significantly affect estimates. This is not reflected in the targeted ESS and can make it deceptive with regard to the causal question.

Second, we recommend a separate assessment of severe positivity violations. For deterministic positivity, existing metrics should be used, such as the convex hull approach by King & Zeng (2006) or the PoRT algorithm by Danelian et al. (2023). Stochastic violations can be identified when the propensity scores are exactly 1 or 0 for a given sample. Furthermore, we recommend examining the distribution of propensity scores. If the mean propensity score is near zero or one, this is a clear signal that IPTW and targeted ESS estimates will not be reliable.

Third, we recommend that targeted ESS be used as a communication tool, as noted by Thomassen et al. (2024). A targeted ESS, based on the estimator used in the analysis, should be reported with findings. This targeted ESS should reflect the assumptions underlying the estimator, such as homoskedasticity and independence. It is not meant to be a pass/fail threshold, but a low relative ESS should alert a researcher to potential estimability concerns regarding their causal estimand.

Finally, we extend the previous recommendation and suggest that targeted ESS be used to compute standard errors, $\frac{\hat{\sigma}}{\sqrt{\text{ESS}}}$, where $\hat{\sigma} = \sqrt{\text{Var}_f(\tilde{\psi})}$. Standard errors reflect the uncertainty of the estimates. With IPTW, we make inferences about a distribution different from the observed

[8]The IPT weighting projects the data onto the interventional distribution corresponding to the observed data $\hat{p}_n$ where $A \perp\!\!\!\perp X$. By the continuous mapping theorem, given the mapping $\psi(\hat{f}_n) = \hat{p}_n$, then $\hat{f}_n \xrightarrow{d} f$ implies that $\psi(\hat{f}_n) \xrightarrow{d} p$.

one, which is not equally efficient; ESS captures this relative efficiency. Using the ESS instead of the observed sample size corresponds to the standard error one would obtain from ESS independent observations under the target distribution. As such, using the targeted ESS for the standard error is more honest about the confidence associated with their estimates.

VI. Conclusion

Causal inference from observational data relies on the positivity assumption. In practice, this assumption is difficult to verify and is often left unexamined. Even if positivity is technically satisfied, near-violations can significantly reduce the estimability of the causal effects. Existing diagnostics have limitations, such as a focus on univariate balance, the need for ad hoc hyperparameters, or the requirement for expert knowledge. In this paper, we propose the targeted effective sample size, which quantifies the efficiency loss of IPTW estimators relative to a randomised experiment and can thus be used to infer whether some near-violations of stochastic positivity are a concern.

We adapted the definition of ESS by adding specific criteria that align it with causal inference. This was necessary as we found that the conventional ESS is not generally suitable for causal inference. It implicitly assumes that the targeted and the observed distributions are identical, which directly contradicts the idea behind IPTW estimators.

Through simulations, we demonstrate the strength of targeted ESS in describing the estimability of the causal effect under identifying assumptions. This includes scenarios where some near-violations of stochastic positivity exist. Additionally, we highlight its limitations. ESS uses the same weights as the IPTW estimator and hence also relies on the positivity assumption. As such, it cannot identify deterministic non-positivity, severe stochastic non-positivity, or misspecification of the propensity model. It also focuses only on the variance and is blind to potential bias in the estimated propensity or, if applicable, the outcome model. These limitations mean that targeted ESS should be used in combination with other diagnostics that can detect these blind spots.

Overall, targeted ESS quantifies the efficiency cost of weighting samples to estimate a causal effect and allows applied researchers to understand and communicate the uncertainty in their study effectively.

Remark on AI-Usage

Claude (Anthropic) and ChatGPT (OpenAI) were used for brainstorming, idea testing, and learning of (mathematical) concepts during the research process. Claude Code was used for code assistance for the simulations, the visualisation of results and for querying read sources for specific information. Grammarly was used for grammar and spelling checks during writing. All outputs were verified by the author. The author retains full responsibility for the content, arguments, and conclusions presented in this work.

References

Austin, P. C. (2009). Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Statistics in Medicine*, *28*(25), 3083–3107. https://doi.org/10.1002/sim.3697

Austin, P. C., & Stuart, E. A. (2015). Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. *Statistics in Medicine*, *34*(28), 3661–3679. https://doi.org/10.1002/sim.6607

Bao, H., & Schomaker, M. (2025, September 22). *Addressing Positivity Violations in Continuous Interventions through Data-Adaptive Strategies*. arXiv. https://doi.org/10.48550/arXiv.2502.14566

Bettega, F., Mendelson, M., Leyrat, C., & Bailly, S. (2024). Use and reporting of inverse-probability-of-treatment weighting for multicategory treatments in medical research: a systematic review. *Journal of Clinical Epidemiology*, *170*, 111338. https://doi.org/10.1016/j.jclinepi.2024.111338

Chesnaye, N. C., Stel, V. S., Tripepi, G., Dekker, F. W., Fu, E. L., Zoccali, C., & Jager, K. J. (2022). An introduction to inverse probability of treatment weighting in observational research. *Clinical Kidney Journal*, *15*(1), 14–20. https://doi.org/10.1093/ckj/sfab158

Cole, S. R., & Hernan, M. A. (2008). Constructing Inverse Probability Weights for Marginal Structural Models. *American Journal of Epidemiology*, *168*(6), 656–664. https://doi.org/10.1093/aje/kwn164

Danelian, G., Foucher, Y., Léger, M., Borgne, F. L., & Chatton, A. (2023). Identification of in-sample positivity violations using regression trees: The PoRT algorithm. *Journal of Causal Inference*, *11*(1). https://doi.org/10.1515/jci-2022-0032

Datta, J., & Polson, N. (2025, April 14). *Inverse Probability Weighting: from Survey Sampling to Evidence Estimation*. arXiv. https://doi.org/10.48550/arXiv.2204.14121

D'Amour, A., Ding, P., Feller, A., Lei, L., & Sekhon, J. (2021). Overlap in observational studies with high-dimensional covariates. *Journal of Econometrics*, *221*(2), 644–654. https://doi.org/10.1016/j.jeconom.2019.10.014

Elvira, V., Martino, L., & Robert, C. P. (2022). Rethinking the Effective Sample Size. *International Statistical Review*, *90*(3), 525–550. https://doi.org/10.1111/insr.12500

Hernan, M. A., & Robins, J. M. (2020). *Causal Inference: What If*. Chapman & Hall/CRC.

Hernán, M. A. (2022). Causal analyses of existing databases: no power calculations required. *Journal of Clinical Epidemiology*, *144*, 203–205. https://doi.org/10.1016/j.jclinepi.2021.08.028

Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2007). Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference. *Political Analysis*, *15*(3), 199–236. https://doi.org/10.1093/pan/mpl013

Holland, P. W. (1986). Statistics and Causal Inference. *Journal of the American Statistical Association*, *81*(396), 945–960. https://doi.org/10.1080/01621459.1986.10478354

Horvitz, D. G., & Thompson, D. J. (1952). A Generalization of Sampling Without Replacement From a Finite Universe. *Journal of the American Statistical Association*, *47*(260), 663–685. https://doi.org/10.2307/2280784

Hsieh, F. Y., Lavori, P. W., Cohen, H. J., & Feussner, J. R. (2003). An overview of variance inflation factors for sample-size calculation. *Evaluation & the Health Professions*, *26*(3), 239–257. https://doi.org/10.1177/0163278703255230

Imbens, G. W., & Xu, Y. (2024, May 27). *Comparing Experimental and Nonexperimental Methods: What Lessons Have We Learned Four Decades After LaLonde (1986)?*. Social Science Research Network. https://doi.org/10.2139/ssrn.4849285

King, G., & Zeng, L. (2006). The Dangers of Extreme Counterfactuals. *Political Analysis*, *14*(2), 131–159. https://doi.org/10.1093/pan/mpj004

Kish, L. (1965). *Survey Sampling*. Wiley.

Kish, L. (1995). *Methods for Design Effects*. 55. https://www.proquest.com/docview/1266820489?pq-origsite=gscholar&fromopenview=true

Kong, A. (1992). A note on importance sampling using standardized weights. *University of Chicago, Dept. Of Statistics, Tech. Rep*, *348*, 14.

Kuesten, C., Dang, J., Nakagawa, M., Bi, J., & Meiselman, H. L. (2016). Propensity score analysis (PSA) for sensory causal inference – Global consumer psychographics and applications for phytonutrient supple-

ments. *Food Quality and Preference*, *51*, 77–88. https://doi.org/10.1016/j.foodqual.2016.02.020

LaLonde, R. J. (1986). Evaluating the Econometric Evaluations of Training Programs with Experimental Data. *The American Economic Review*, *76*(4), 604–620. https://www.jstor.org/stable/1806062

Liu, J. S., & Chen, R. (1995). Blind Deconvolution via Sequential Imputations. *Journal of the American Statistical Association*, *90*(430), 567–576. https://doi.org/10.1080/01621459.1995.10476549

Liu, J., Liu, Y., Zhou, Y., & Matsouaka, R. A. (2025). Assessing racial disparities in healthcare expenditure using generalized propensity score weighting. *BMC Medical Research Methodology*, *25*(1), 64. https://doi.org/10.1186/s12874-025-02508-2

Martino, L., Elvira, V., & Louzada, F. (2017). Effective sample size for importance sampling based on discrepancy measures. *Signal Processing*, *131*, 386–401. https://doi.org/10.1016/j.sigpro.2016.08.025

Matsouaka, R. A., & Zhou, Y. (2024). Causal inference in the absence of positivity: The role of overlap weights. *Biometrical Journal*, *66*(4), 2300156. https://doi.org/10.1002/bimj.202300156

McCaffrey, D. F., Griffin, B. A., Almirall, D., Slaughter, M. E., Ramchand, R., & Burgette, L. F. (2013). A tutorial on propensity score estimation for multiple treatments using generalized boosted models. *Statistics in Medicine*, *32*(19), 3388–3414. https://doi.org/10.1002/sim.5753

Owen, A. B. (2013). 9 Importance sampling. In *Monte Carlo theory, methods and examples: Monte Carlo theory, methods and examples* (p. 46). https://artowen.su.domains/mc/. https://artowen.su.domains/mc/Ch-var-is.pdf

Park, I., & Lee, H. (2004). Design effects for the weighted mean and total estimators under complex survey sampling. *Survey Methodology*, *30*(2), 183–193.

Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*. Cambridge University Press.

Petersen, M. L., Porter, K. E., Gruber, S., Wang, Y., & Laan, M. J. van der. (2012). Diagnosing and responding to violations in the positivity assumption. *Statistical Methods in Medical Research*, *21*(1), 31–54. https://doi.org/10.1177/0962280210386207

Ring, K., & Schomaker, M. (2025, February 17). *A Diagnostic to Find and Help Combat Positivity Issues – with a Focus on Continuous Treatments*. arXiv. https://doi.org/10.48550/arXiv.2502.11820

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41–55. https://doi.org/10.1093/biomet/70.1.41

Rubin, D. B. (2005). Causal Inference Using Potential Outcomes: Design, Modeling, Decisions. *Journal of the American Statistical Association*, *100*(469), 322–331. https://doi.org/10.1198/016214504000001880

Shook-Sa, B. E., & Hudgens, M. G. (2020, March 12). *Power and Sample Size for Marginal Structural Models*. arXiv. https://doi.org/10.48550/arXiv.2003.05979

StataCorp LLC. (2021). *Stata Treatment-Effects Reference Manual: Potential Outcomes/Counterfactual Outcomes* (Release17 ed.). Stata Press.

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science : a Review Journal of the Institute of Mathematical Statistics*, *25*(1), 1–21. https://doi.org/10.1214/09-STS313

Särndal, C.-E., Swensson, B., & Wretman, J. (2003). *Model Assisted Survey Sampling*. Springer Science & Business Media.

Thomassen, D., Cessie, S. le, Houwelingen, H. C. van, & Steyerberg, E. W. (2024). Effective sample size: A measure of individual uncertainty in predictions. *Statistics in Medicine*, *43*(7), 1384–1396. https://doi.org/10.1002/sim.10018

Westreich, D., & Cole, S. R. (2010). Invited Commentary: Positivity in Practice. *American Journal of Epidemiology*, *171*(6), 674–677. https://doi.org/10.1093/aje/kwp436

Zhang, L., Bujkiewicz, S., & Jackson, D. (2024). Three new methodologies for calculating the effective sample size when performing population adjustment. *BMC Medical Research Methodology*, *24*(1), 287. https://doi.org/10.1186/s12874-024-02412-1

Zhou, T., Tong, G., Li, F., E. Thomas, L., & Li, F. (2022). PSweight: An R Package for Propensity Score Weighting Analysis. *The R Journal*, *14*(1), 282–300. https://doi.org/10.32614/RJ-2022-011

Zhu, Y., Hubbard, R. A., Chubak, J., Roy, J., & Mitra, N. (2021). Core Concepts in Pharmacoepidemiology: Violations of the Positivity Assumption in the Causal Analysis of Observational Data: Consequences and Statistical Approaches. *Pharmacoepidemiology and Drug Safety*, *30*(11), 1471–1485. https://doi.org/10.1002/pds.5338

Zivich, P. N., Cole, S. R., & Westreich, D. (2022, July 11). *Positivity: Identifiability and Estimability*. arXiv. https://doi.org/10.48550/arXiv.2207.05010