Received: 28 October 2020 | Revised: 7 September 2021 | Accepted: 24 October 2021

DOI: 10.1002/bimj.202000323

Biometrical Journal

RESEARCH ARTICLE

Causal inference in case of near-violation of positivity: comparison of methods

Maxime Léger[1,2] | **Arthur Chatton**[1,3] | **Florent Le Borgne**[1,3] | **Romain Pirracchio**[4] | **Sigismond Lasocki**[2] | **Yohann Foucher**[1,5]

[1] INSERM UMR 1246 - SPHERE, Université de Nantes, Université de Tours, Nantes, France

[2] Département d'Anesthésie-Réanimation, Centre Hospitalier Universitaire d'Angers, Angers, France

[3] IDBC-A2COM, Nantes, France

[4] Department of Anesthesia and Perioperative Care, University of California, San Francisco, CA, USA

[5] Centre Hospitalier Universitaire de Nantes, Nantes, France

Correspondence
Yohann Foucher, Centre Hospitalier Universitaire de Nantes, Nantes, France.
Email: Yohann.Foucher@univ-nantes.fr

Funding information
French National Research Agency, Grant/Award Number: ANR-16-LCV1-0003-01

This article has earned an open data badge "**Reproducible Research**" for making publicly available the code necessary to reproduce the reported results. Results in this article were not fully reproducible due to insufficient quality of code and documentation.

Abstract

In causal studies, the near-violation of the positivity may occur by chance, because of sample-to-sample fluctuation despite the theoretical veracity of the positivity assumption in the population. It may mostly happen when the exposure prevalence is low or when the sample size is small. We aimed to compare the robustness of g-computation (GC), inverse probability weighting (IPW), truncated IPW, targeted maximum likelihood estimation (TMLE), and truncated TMLE in this situation, using simulations and one real application. We also tested different extrapolation situations for the sub-group with a positivity violation. The results illustrated that the near-violation of the positivity impacted all methods. We demonstrated the robustness of GC and TMLE-based methods. Truncation helped in limiting the bias in near-violation situations, but at the cost of bias in normal conditions. The application illustrated the variability of the results between the methods and the importance of choosing the most appropriate one. In conclusion, compared to propensity score-based methods, methods based on outcome regression should be preferred when suspecting near-violation of the positivity assumption.

KEYWORDS

causal inference, doubly robust estimators, g-computation, positivity, propensity score, real-world evidence, simulations

1 | INTRODUCTION

There is growing interest in causal methods (Hernán & Robins, 2020), notably the propensity score (PS)-based methods (Austin, 2011; Williamson et al., 2012). The PS is related to the exposure prediction. One can distinguish four different approaches: matching, stratification, conditional adjustment, and inverse probability weighting (IPW) (Robins et al., 2000; Rosenbaum & Rubin, 1983). IPW and matching on PS estimate marginal effects, while stratification and conditioning estimate conditional effects. In the settings of nonlinear link functions, marginal and conditional estimates may differ due to the noncollapsibility issues. IPW and matching emerge as preferable methods for estimating marginal effects (Austin,

2013). However, both IPW and matching suffer efficiency limitations: IPW due to extreme weights and matching due to nonmatching subjects resulting in loss of information. Despite the problems of the most extreme subjects, IPW emerges as a preferable options in terms of both bias and precision (Abdia et al., 2017; Hajage et al., 2016; Le Borgne et al., 2016; Lendle et al., 2013). An alternative is g-computation (GC), also known as parametric g-formula or (g-)standardisation (Robins, 1986; Snowden et al., 2011; Vansteelandt & Keiding, 2011). The latter estimator relies on an outcome model rather than an exposure model like for PS-based methods. Some estimators combine GC and PS to create doubly robust estimators (DREs) aiming to minimize the impact of model misspecification on consistency (Bang & Robins, 2005; Neugebauer & van der Laan, 2005). One of the most studied doubly robust methods is the targeted maximum likelihood estimation (TMLE) (van der Laan & Rubin, 2006).

1.1 | Positivity violation

Regardless of the type of estimator, in order to conclude causally, one has to make several assumptions: consistency, conditional exchangeability, and positivity. Positivity is met if, for any combination of the covariates, there is a nonnull probability of being exposed or unexposed. Positivity violations can occur in two situations: (i) theoretical violation: we know that there are patients with a null probability of being exposed or unexposed, for example, if certain patients present a contraindication to receiving a treatment of interest; (ii) near or practical violation, sampling variability may result in subjects having a null probability of being exposed or unexposed for certain combinations of covariate values. This may be particularly frequent for cases of low exposure prevalence or small sample sizes (Westreich & Cole, 2010).

The theoretical violation is a consequence of a conceptual problem in the study design and calls for restricting the studied population (Petersen et al., 2012; Westreich & Cole, 2010), that is, excluding patients with a theoretical null probability of being exposed or unexposed (for instance, patients with a counter-indication for one of the studied treatments). In contrast, in case of near-violation, the target population is well defined. In this situation, the goal is to select an estimator that does not suffer from the near-violation. For IPW, one can empirically set threshold values for truncating (Cole & Hernán, 2008) or trimming the PS (Crump et al., 2009). These approaches aim to limit the maximum contribution of extreme observations. Truncation has the advantage of preserving clinical equipoise in the target population, whereas excluding certain subjects would result in a trimmed population that would change the estimand.

1.2 | Extrapolation issue

For the methods based on the outcome regression, the problem is an extrapolation of the outcome prediction for the patients affected by the near-violation, rather than using actual observations in the data (Neugebauer & van der Laan, 2006; van der Laan & Robins, 2003).

Let (Y, A, Z) denote the binary outcome ($Y = 1$ for events and 0 otherwise), the binary exposure ($A = 1$ for exposed individuals and 0 otherwise), and the p baseline covariates $(Z_1, \dots, Z_p)$. Let us define $f(Z_1|A)$, the density function of the quantitative covariate Z_1 conditional to A, Z_1 being a true confounder that causes both the exposure status A and the outcome Y. As illustrated in Figure 1, consider a near-violation of the positivity for $Z_1 > \alpha$ and an effect of a theoretical increase in the conditional probability of the outcome under $A = 1$ for larger values of Z_1.

Because of the lack of information when $Z_1 > \alpha$ due to the near-violation of the positivity assumption, the estimation of the exposure effect relies on extrapolating the observed effect, that is, when $Z_1 \leq \alpha$. Even when the outcome model is adequately specified in the region supported by data, the model may be inadequate for the region suffering from positivity near-violation.

The causal inferences will depend on the formulation of nontestable hypotheses.

One can note that this illustration (with Z_1 as a quantitative confounder can be extended for a binary confounder. Consider that Z_1 represents the gender. If there is no information regarding the outcomes among exposed women, one cannot properly infer the average exposure effect in the target population.

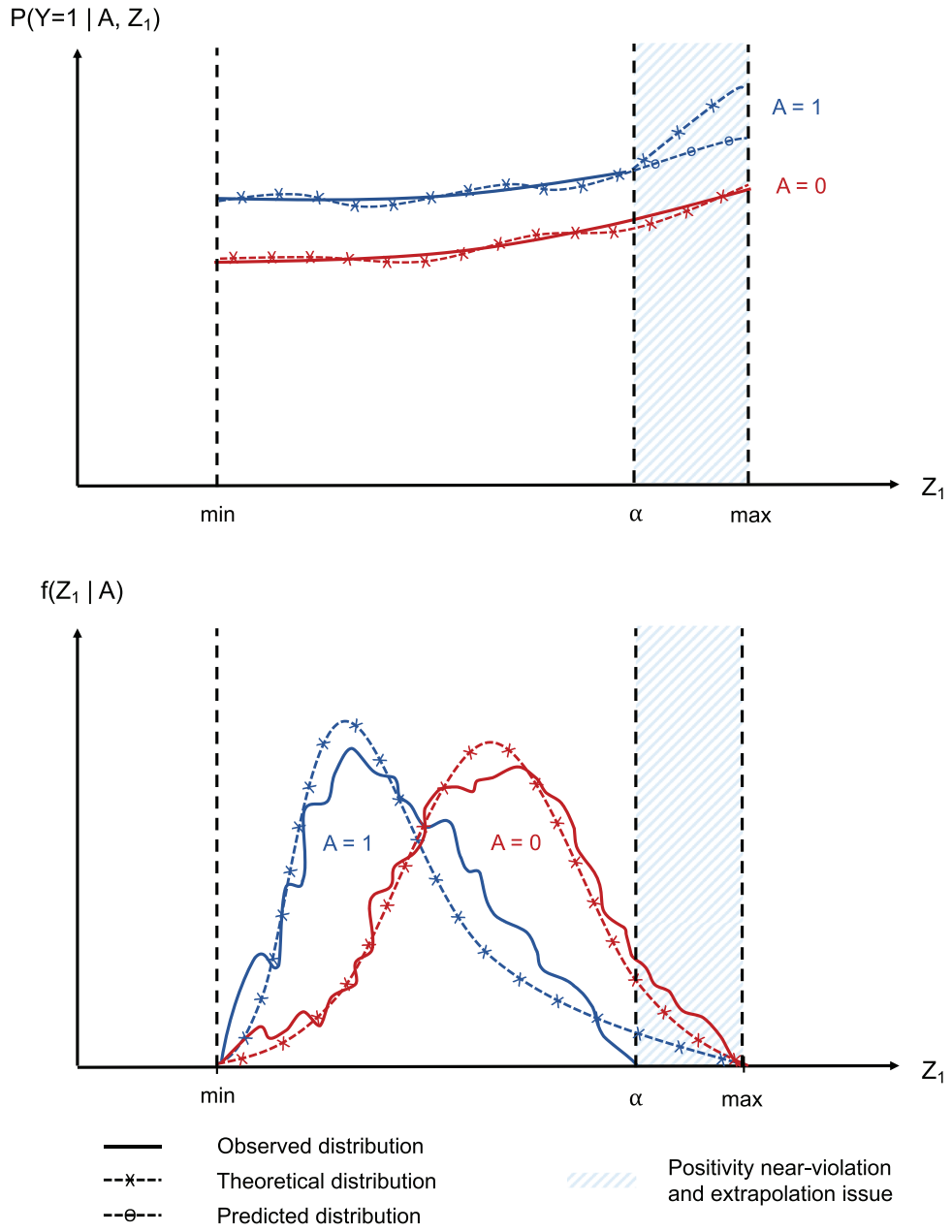

FIGURE 1 Representative illustration of the extrapolation issue occurring with a positivity near-violation. The left y-axis represents the conditional distribution function of the covariate Z_1 according to the exposure status. The right y-axis represents the conditional probability of the outcome

1.3 | Framework

The literature does not provide a clear answer as to the most reliable method in cases of positivity near-violation. Indeed, even though several studies have compared the previous methods in the context of positivity violation (Lendle et al., 2013; Moore et al., 2012; Petersen et al., 2012), suggesting better stability and reduced bias for GC and DRE, they did not investigate the extrapolation issue.

In the situation of positivity near-violation, Petersen et al. (2012) introduced the problem of extrapolation. Nevertheless, they did not study its impact.

In this context, we performed a simulation-based study to evaluate the robustness of IPW, truncated IPW, GC, TMLE, and truncated TMLE in the situation of the extrapolation issue and positivity near-violation. We also evaluated one application from a real dataset. This study is structured as follows: Section 2 outlines the methods used, Section 3 presents

the design and the results of the simulation study, in Section 4, we apply the developed application to a real dataset, and finally we discuss the results and provide recommendations.

2 | METHODS

2.1 | Setting and notations

Consider a resulting sample of size n in which one can observe the realizations of these random variables (y, a, z). Define $\pi_a = P(Y = 1|do(A = a))$ as the expected proportions of event if the entire population is exposed ($do(A = 1)$) or unexposed ($do(A = 0)$) (Pearl et al., 2016). The average exposure effect on the entire population is defined as $\Delta = \pi_1 - \pi_0$. The corresponding marginal causal odds ratio is expressed as $\text{OR} = (\pi_1/(1 - \pi_1))/(\pi_0/(1 - \pi_0))$.

2.2 | Inverse Probability Weighting

Formally, the PS for a subject i ($i = 1, \dots, n$) is $p_i = P(A = 1|z_i)$, that is, the probability that a subject is exposed according to her/his observed characteristics z_i (Rosenbaum & Rubin, 1983). The PS is often estimated from logistic regression, but other models or algorithms can be used such as random forest, boosting, or super learner (Austin, 2012; Pirracchio & Carone, 2018). IPW results in weighting the contribution of each subject i by $\omega_i = A_i P(A_i = 1)/p_i + (1 - A_i)P(A_i = 0)/(1 - p_i)$, where $P(A_i = 1)$ and $P(A_i = 0)$ denote the marginal probability of exposure and its complementary. The use of such stabilized weights are preferred to optimize the variance estimation (Robins et al., 2000; Xu et al., 2010). Based on ω_i, the maximization of the weighted likelihood of the logistic regression with Y as the outcome and A as the unique explanatory variables allows us to obtain $\hat{\pi}_0^{IPW}$, $\hat{\pi}_1^{IPW}$, and $\widehat{\text{OR}}^{IPW}$.

2.3 | Truncated IPW

The weights ω_i can largely inflate for a subject i concerned by positivity near-violation. The usual approach is to truncate the lowest and the highest p_i estimations by the 10th and 90th percentiles, respectively (Cole & Hernán, 2008). We also analyzed alternative thresholds, including the 5th and 95th percentiles, as well as the 2.5th and 97.5th percentiles of the estimated PS. We obtained truncated stabilized weights, and the estimations $\hat{\pi}_0^{T-IPW}$, $\hat{\pi}_1^{T-IPW}$, and $\widehat{\text{OR}}^{T-IPW}$.

2.4 | G-computation

GC is based on the outcome regression, frequently called the Q-model (Snowden et al., 2011). The logistic regression is often used when Y is binary. Other models or algorithms can constitute alternatives (Austin, 2012). Consider the following Q-model: $\text{logit}\{P(Y = 1|A, Z)\} = \gamma A + \beta Z$. Once fitted, one can compute for each subject i the two expected probabilities of events if she/he is exposed or unexposed, that is, $\hat{P}(Y_i = 1|do(A_i = 1), z_i)$ and $\hat{P}(Y_i = 1|do(A_i = 0), z_i)$, respectively (Snowden et al., 2011). One can then obtain $\hat{\pi}_a^{GC} = n^{-1} \sum_i \hat{P}(Y_i = 1|do(A_i = a), z_i)$ for $a = 0, 1$; $\hat{\Delta}^{GC} = \hat{\pi}_0^{GC} - \hat{\pi}_1^{GC}$ and $\widehat{\text{OR}}^{GC} = (\hat{\pi}_1^{GC}/(1 - \hat{\pi}_1^{GC}))/(\hat{\pi}_0^{GC}/(1 - \hat{\pi}_0^{GC}))$. This method is implemented in the `RISCA` package, in R (Foucher et al., 2019).

2.5 | Targeted Maximum Likelihood Estimation

The first step is to fit the Q-model and estimate the two expected probabilities of events $\hat{\pi}_1^{GC}$ and $\hat{\pi}_0^{GC}$. The additional "targeting" step involves the estimation of p_i, which is then used to update the initial estimates obtained by the Q-model. This step aims to compute first: the clever covariates $H(1, Z) = A/(\text{expit}(\hat{p}_i))$ and $H(0, Z) = (1 - A)/(1 - \text{expit}(\hat{p}_i))$, where expit() represents the inverse logit function ($\frac{\exp()}{1+\exp()}$), and second: a vector fluctuation parameter $\hat{\epsilon} = (\hat{\epsilon}_0, \hat{\epsilon}_1)$ estimated

through a maximum likelihood procedure. The fluctuation parameter is computed using an outcome model where the logit of the initial prediction of the Q-model is an offset in an intercept-free logistic regression with the clever covariates as explanatory variables (Luque-Fernandez et al., 2018). Therefore, we can generate updated estimates of the set of potential outcomes (Y_1^* and Y_0^*) by incorporating information from the mechanisms to reduce potential biases. We generate $\text{logit}(\hat{Y}_1^*) = \text{logit}(\hat{Y}_1) + \hat{\epsilon} \times H_1$ and $\text{logit}(\hat{Y}_0^*) = \text{logit}(\hat{Y}_0) + \hat{\epsilon} \times H_0$ (Schuler & Rose, 2017). In the presence of residual confounders, the PS provides additional information to improve the initial estimates. It results in the estimations $\hat{\pi}_0^{TMLE}$ and $\hat{\pi}_1^{TMLE}$, that is, the updated values of $\hat{\pi}_0^{GC}$ and $\hat{\pi}_0^{GC}$, respectively. This method is implemented in the `tmle` package, in R (Gruber & van der Laan, 2012).

2.6 | TMLE with truncated PS

As for IPW, the TMLE can use truncated PS in its second stage. The usual method is the truncation of the lowest and highest values of p_i by 0.1 and 0.9, respectively. We also analyzed other alternative truncation levels: 0.05/0.95 and 0.025/0.975. One can then obtain $\hat{\pi}_1^{T-TMLE}$, $\hat{\pi}_0^{T-TMLE}$ and $\widehat{\text{OR}}^{T-TMLE}$. We used the `gbounds` arguments in the `tmle` function of the `tmle` package in R (Gruber & van der Laan, 2012).

2.7 | Variance estimators

For each method, the variance was obtained from the usual and well-validated method. For IPW, we used a robust sandwich-type variance estimator (Robins et al., 2000), with the `sandwich` package in R (Zeileis, 2006). For GC, we generated 1000 bootstrapped samples. This method is implemented in the `RISCA` package in R (Foucher et al., 2019). For TMLE, we used the efficient curve based variance estimator, implemented in the `tmle` package in R (Gruber & van der Laan, 2012).

To improve the comparability of the results, we additionally used the bootstrap for IPW and TMLE-based methods.

3 | SIMULATION STUDY

3.1 | Data generation

Figure S1 (Supplementary Material) represents the directed acyclic graph of the simulations. We first independently generated covariates $Z = (Z_1, \ldots, Z_9)$: six binary covariates using Bernoulli distributions with different probabilities (0.1 for Z_1, 0.4 for Z_2, 0.7 for Z_4, 0.5 for Z_5, 0.3 for Z_7, and 0.8 for Z_8), and three continuous covariates using a Gaussian distribution with mean at 0 and standard deviation at 1. We generated the exposure A according to a Bernoulli distribution with probability obtained from a logistic model with the following linear predictor: $\alpha_0 + \alpha Z_1 + \alpha Z_2 + \alpha Z_4 + \alpha Z_6 + \alpha Z_7 + \alpha Z_8$, α being the regression coefficients associated with the covariates as detailed in Table S1, and α_0 was set to 1.05 or -0.45 to simulate a prevalence of exposed patients at 80% or 50%, respectively. This design allows us to expect situations of positivity near-violation (Figures S2 and S3), especially for Z_1 which was generated with a 10% prevalence. A prevalence of 50% improved the PS distribution overlap between exposed and nonexposed subjects, and reducing the risk of positivity near-violation. Furthermore, because the near-violation is more susceptible for small samples, we studied several sample sizes: $n = 100, 200, 500$, and 1000.

We randomly generated the outcome from a Bernoulli distribution with probability obtained from a logistic model with the following linear predictor: $-0.8 + \beta_A A + \beta_Z Z_1 + \beta_Z Z_2 + \beta_Z Z_3 + \beta_Z Z_4 + \beta_Z Z_5 + \beta_Z Z_6 + \beta_{A,Z_1} A * Z_1$, where (β_A, β_Z) were the regression coefficients of A and Z, respectively. To create an extrapolation issue as illustrated in Figure 1, we considered an interaction between A and Z_1 in the outcome-generating model to obtain a poorly calibrated model in the area where Z_1 violated the positivity assumption. The values of β_A and β_Z are presented in Table S1. The regression coefficient β_{A,Z_1} of the interaction ranged from $0.0 \cdot \beta_A$ to $2.0 \cdot \beta_A$, according to the intensity of the extrapolation issues : 0.0 for no issue, 0.3 for low issues, 0.9 for moderate issues, and 2.0 for high issues.

For each of the 32 scenarios (four sample sizes, two exposures, and four extrapolation scenarios), we generated 1000 datasets. Among the generated datasets for a 50% exposure prevalence, the near-violation of the positivity assumption

(no unexposed subjects with $Z_1 = 1$) concerned 0.0% of the datasets for $n = 1000$ or 500 subjects, 1.3% for $n = 200$ subjects, and 14.1% for $n = 100$ subjects. For an 80% exposure prevalence, this near-violation concerned 0.2% of the datasets for $n = 1000$, 7.2% for $n = 500$ subjects, 31.8% for $n = 200$ subjects, and 58.2% for $n = 100$ subjects.

3.2 | Estimations

We used correctly specified exposure and outcome models to study the impact of positivity near-violation and the extrapolation issue. The interaction between Z_1 and A was introduced in both the models for data generation and the models estimated in each simulated dataset. Even if the outcome model was theoretically well specified, its estimation could result in poor calibrated predictions where there was no data support in the near-violation area.

The interaction between Z_1 and A was introduced in both the models for data generation and the models estimated in each simulated dataset. Even if the outcome model is theoretically well specified, its estimation may result in poor calibrated predictions where there is no data support in the near-violation area.

We estimated the true values of π_1 and π_0 by averaging the values obtained from a univariate logistic model (the exposure as the only covariate), fitted from datasets generated as above, except that the exposure A was simulated independently of the covariates Z (Gayat et al., 2012).

To ensure comparability between methods, we decided to set the same strategy of variables' selection. Our set of covariates corresponded to all the outcome causes, theoretically defined by the simulation design (Figure S1), that is, Z_1, Z_2, Z_3, Z_4, Z_5, and Z_6 (Chatton et al., 2020). We did not study data-adaptive methods to optimize our set of covariates (for instance, the collaborative targeted maximum likelihood estimation (van der Laan & Gruber, 2010)), or even a data-adaptive choice of the truncated PS threshold (Bembom & van der Laan, 2008).

The main estimand was the log(OR). We reported several associated criteria: the mean absolute bias (MAB) ($MAB = E(\log(\widehat{\mathrm{OR}})) - \log(\mathrm{OR}))$, the variance estimation ratio (VER) by the ratio of estimated model standard deviation to empirical standard deviation ($VER = (\sqrt{E[\widehat{\mathrm{V}}\mathrm{ar}(\log(\widehat{\mathrm{OR}}))]}/\sqrt{\mathrm{Var}(\log(\widehat{\mathrm{OR}}))})$, the mean square error (MSE) ($MSE = E[(\log(\widehat{\mathrm{OR}}) - \log(\mathrm{OR}))^2]$) , the coverage rate of the 95% confidence interval (95%CI), and the statistical power. We also reported the mean bias of the probability of an event under the two counterfactual treatments as well as their difference (Δ). We computed the Monte Carlo standard errors for each metric (Morris et al., 2019). We performed all of the analyses using the R software package (R Core Team, 2014).

3.3 | Results

The results are presented in Figures 2–4 for an 80% exposure prevalence. For the methods with truncation, we report in this subsection the results obtained by using the 10th and 90th percentiles, which were associated with the lower MSE values. We also performed the analyses for the 5th and 95th percentiles, and the 2.5th and 97.5th percentiles. These additional results are detailed as Supplementary Information in Tables S2 and S3 for an exposure prevalence at 80%, and in Tables S6 and S7 for an exposure prevalence at 50% (with Figures S6–8). The bootstrap-based results for an 80% exposure prevalence are presented in Tables S4, S5, with Figures S4 and S5. The results under the null hypothesis are presented in Tables S8–9 with Figures S9–11 for a prevalence of 80%, and in Tables S10–11 with Figures S12–14 for a prevalence of 50%. The standard Monte Carlo errors were negligible and are not presented in the results.

3.3.1 | Mean bias

The truncated IPW estimator was biased in almost every situation. For the other methods, the bias increased as the near-violation of positivity was accentuated, that is, when the sample size decreased (Figure 2). This increase was more significant for IPW estimators. For instance, for the scenario without an extrapolation issue, the MAB was 0.065 for a sample size of 100 subjects, versus −0.002 for 1000 subjects.

The extrapolation issue increased the MAB for methods based on the outcome modeling (GC, TMLE, and truncated TMLE), but only when the level was high. For instance, for 200 subjects without extrapolation issue, the MAB for GC and truncated TMLE were −0.006 and 0.000, respectively, versus −0.038 and −0.032 with high extrapolation issue. Even

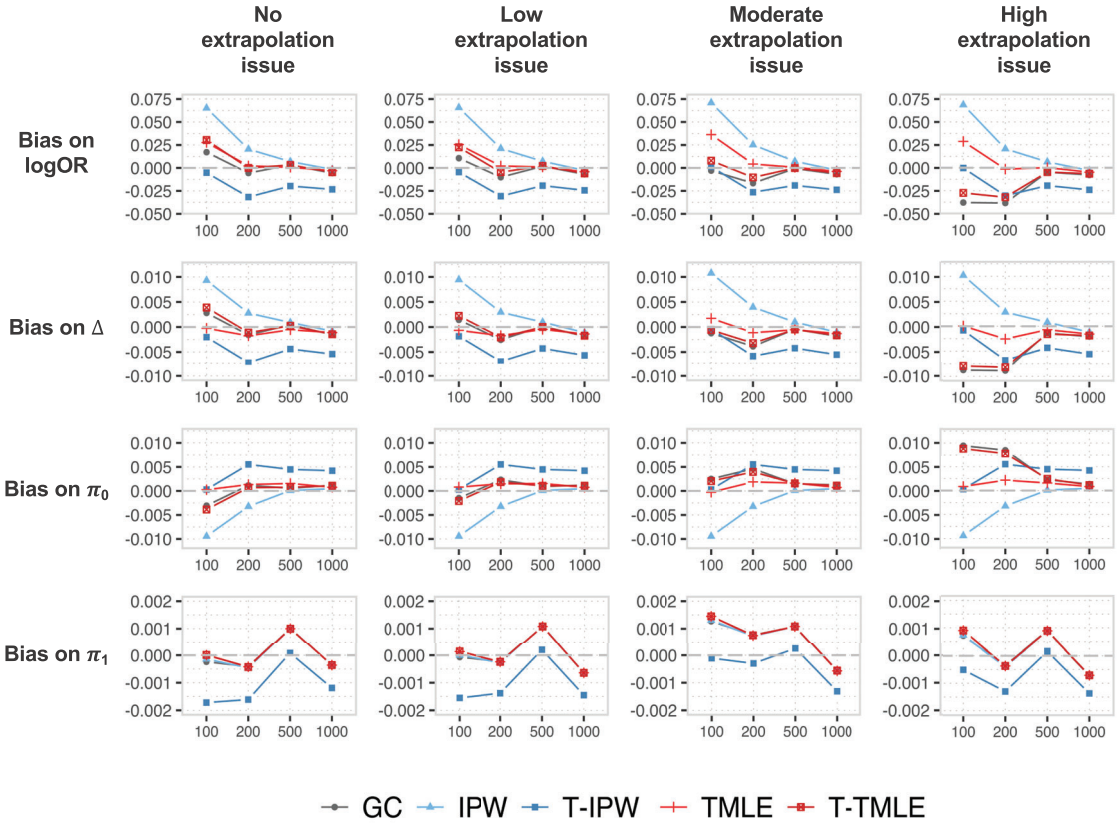

FIGURE 2 The mean absolute bias (y-axis) according to different sample size (from 100 to 1000, x-axis) and extrapolation issue. Abbreviations: GC, g-computation; IPW, inverse probability weighting; T-IPW, truncated inverse probability weighting (thresholds: 10th and 90th percentiles); TMLE, targeting maximum likelihood estimator; T-TMLE, truncated targeting maximum likelihood estimator (thresholds: bounds at 0.1 and 0.9); π_1, the expected proportions of event if the entire population is exposed; π_0, the expected proportions of event if the entire population is unexposed; Δ, the corresponding difference (π_1 - π_0); OR, the corresponding odds-ratio

when its level was high, the extrapolation issue had minor consequences when the sample size was equal to or higher than 500 subjects. The TMLE seemed to be the most robust method across all scenarios, especially for small sample sizes ($n = 200$).

For a prevalence of exposure of 50% where the positivity near-violation was lower, the MABs were lower for all of the methods, with comparable results in terms of bias. Even in the most extreme situations (100 subjects with high extrapolation issue), the methods remained robust.

3.3.2 | Variance

As illustrated in Figure 3, the decreases in the variance associated with the sample size was comparable across all methods. The extrapolation issue did not affect the variance estimation. However, GC was associated with larger variance when the sample size was smaller. The estimated standard deviation for GC was 1.167 for 100 subjects, 0.399 for 200 subjects, 0.244 for 500 subjects, and 0.177 for 1000 subjects. GC was the only method based on bootstrapping, which can explain this result. Therefore and for comparability sake, we subsequently used bootstrapping for the other methods (Tables S4, S5 and Figure S4). In this situation, variance was similar among all methods for 100 subjects.

Note that regardless of the method used for variance estimation, the standard deviations were similar when the prevalence was 50%. For 100 subjects without extrapolation issue, we estimated a standard deviation at 0.447 for GC, 0.458 for the IPW, 0.433 for the truncated IPW, 0.406 for the TMLE, and 0.405 for the truncated TMLE.

The VER was lower for the TMLE-based methods. This over-optimistic estimation of the variance was partially corrected for the largest sample sizes. More precisely, the VER for TMLE were 0.715 for 100 subjects, 0.798 for 200 subjects, 0.839 for

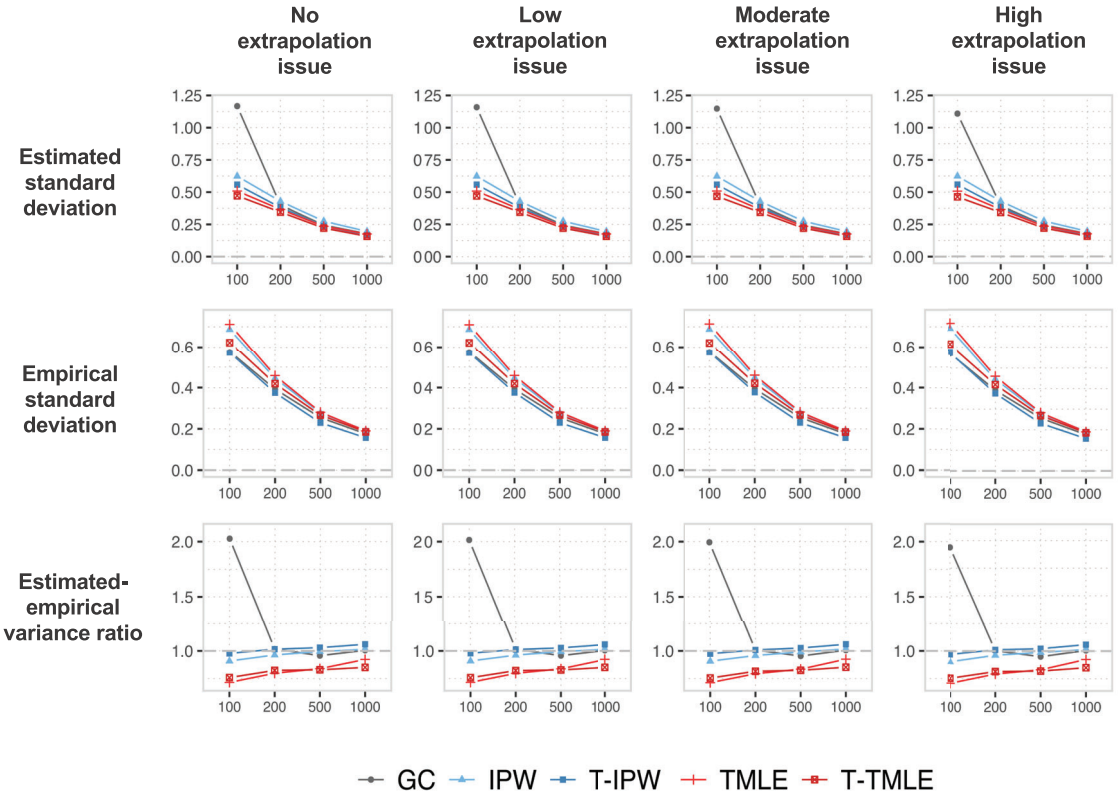

FIGURE 3 Graphical representation of the evolution of accuracy (empirical standard deviation, estimated standard deviation, and the variance estimation ratio) according to different sample size (from 100 to 1000, x-axis) and extrapolation issue. The target parameter was log(OR). Abbreviations: GC, g-computation; IPW, inverse probability weighting; T-IPW, truncated inverse probability weighting (thresholds: 10th and 90th percentiles); TMLE, targeting maximum likelihood estimator; T-TMLE, truncated targeting maximum likelihood estimator (thresholds: bounds at 0.1 and 0.9)

500 subjects, and 0.923 for 1000 subjects. The use of bootstrapping corrected this over-optimistic estimation (Figure S4). Note that truncated TMLE was associated with lower variances (Tables S2 and S3).

3.3.3 | MSE, coverage and power

As illustrated in Figure 4, we observed an increase in the MSE values with the level of the positivity near-violation, in agreement with the previously reported increase in the MAB values. Nevertheless, the MSE was not significantly affected by the problem of extrapolation. The MSE was lower for GC and truncated methods in the most extreme situation. For instance, for 100 subjects, MSE values were 0.331, 0.507, 0.388, 0.475, and 0.326 for GC, TMLE, truncated TMLE, IPW, and truncated IPW, respectively. The lowest MSE was always obtained with the truncated IPW. The second method was GC. Truncated IPW and GC were the two methods with the best bias-variance trade-off. Note that when the prevalence was 50% (Tables S6 and S7), the MSE for the different methods was similar. However, truncated IPW remained the method with the lowest MSE.

As presented in Figure 4, IPW-based methods and GC resulted in nominal coverage values regardless of the sample sizes. TMLE and truncated TMLE underestimated the variance, resulting in coverage issues. For TMLE-based methods, the underestimated variance results in anti-conservative confidence intervals. More precisely, for scenarios without extrapolation issues, the coverage value of TMLE was 84.6% for 100 subjects, 88.4% for 200 subjects, 88.6% for 500 subjects, and 91.4% for 1000 subjects. The use of bootstrapping allowed to correct this underestimation. However, as reported in Tables S4 and S5, we obtained values greater than 95%, regardless of the extrapolation issue: 97.1% for 100 subjects, 96.4% for 200 subjects, 95.7% for 500 subjects, and 94.7% for 1000 subjects. The previous results under the alternative hypothesis remained consistent under the null hypothesis. The type I error rate was close to the nominal 5% value at for all methods, except for the TMLE-based methods (variance estimation with efficient curves), with values close to 10% throughout the scenarios.

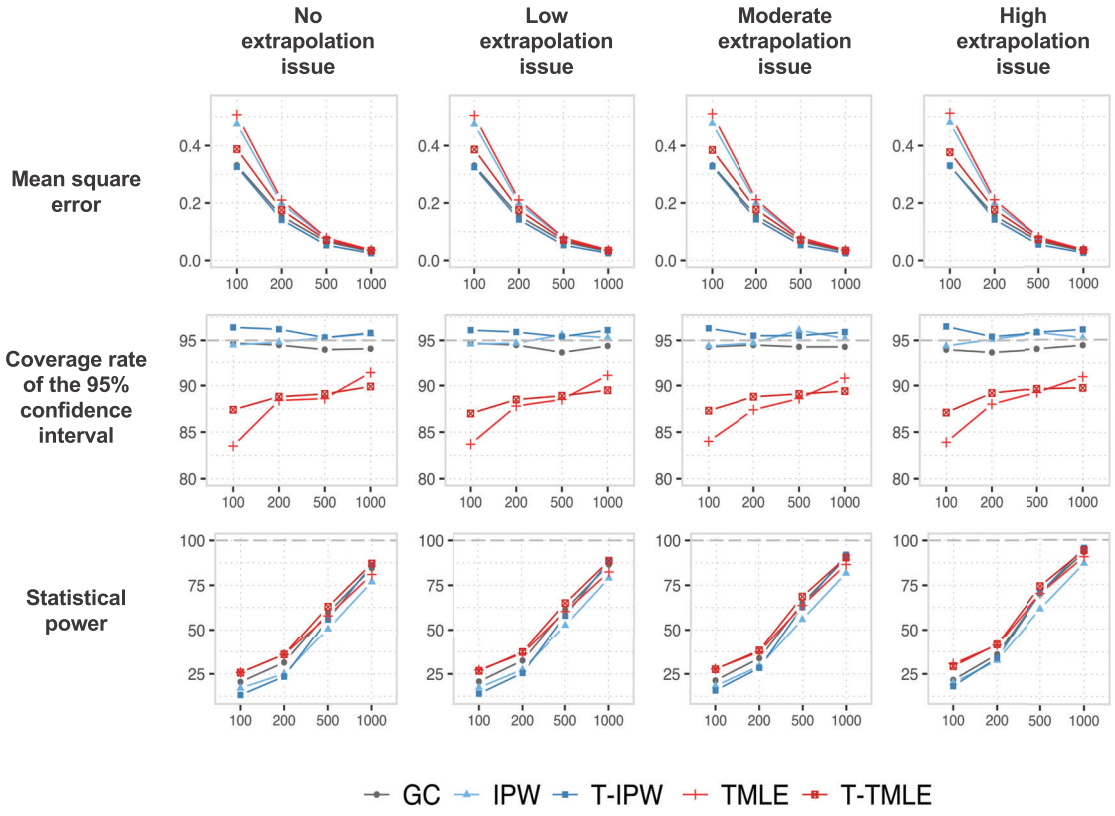

FIGURE 4 The mean square error, the coverage of the 95% confidence interval and the statistical power according to different sample size (from 100 to 1000, x-axis), and extrapolation issue. The target parameter was log(OR). Abbreviations: GC, g-computation; IPW, inverse probability weighting; T-IPW, truncated inverse probability weighting (thresholds: 10th and 90th percentiles) ; TMLE, targeting maximum likelihood estimator; T-TMLE, truncated targeting maximum likelihood estimator (thresholds: bounds at 0.1 and 0.9)

The progressive increase in the extrapolation issue allowed a slight increase in the statistical power for all methods, regardless of the magnitude of the positivity near-violation. In contrast, the statistical power was strongly impacted by the size of the population for all methods, with values around 20% per 100 subjects compared to values around 90% per 1000 subjects. The IPW presented the lowest values, while the truncated TMLE was the highest statistical power method. These results were in agreement with the over-optimistic estimation of the variance for TMLE-based methods. The power of GC was close to the truncated TMLE. The use of truncated methods improved the statistical power. For example, for 200 subjects without extrapolation issues, the powers were 36.0% for truncated TMLE, 35.9% for TMLE, 31.4% for GC, 25.1% for IPW, and 23.5% for truncated IPW. For 1000 subjects, the powers were 87.1% for truncated TMLE, 81.0% for TMLE, 85.6% for GC, 76.9% for IPW, and 86.3% for truncated IPW.

4 | APPLICATION: EFFECT OF BARBITURATES IN PATIENTS WITH INTRACRANIAL HYPERTENSION

We compared the five methods on a real dataset, in situations that could suggest a near-violation of the positivity assumption. We studied barbiturate prescription for the treatment for refractory intracranial hypertension during the first 24 h post-admission, and its relationship to, in-hospital mortality.

4.1 | Methods

We included 1584 patients from the AtlanREA cohort (www.atlanrea.org, CNIL DR-2013-047). These patients were admitted to an intensive care unit (ICU) in France's western region between March 2013 and February 2018, and were monitored for intracranial pressure.

For covariates selection, to be consistent with the simulations, we selected the covariates causing the outcome (Chatton et al., 2020). For this purpose, as proposed by VanderWeele and Shpitser (2011), we asked experts which covariates caused the outcome (i.e., a history of head trauma, use of osmotherapy, type of brain injury, age, SAPS II score, signs of intracranial hypertension on admission, lactate, and creatinine levels on admission). We did not test interactions. We applied B-spline transformations to continuous covariates when the log-linearity assumption did not hold. For IPW-based approaches, we additionally checked the balance between the two weighted groups with standardized differences. We performed complete case analyses.

4.2 | Description of the cohort

Among the 1584 patients, 1119 had no missing data on the outcome or covariates. One hundred and twenty-seven (127) patients were in the treated group versus 992 control patients (no barbiturate during the first 24 h post-admission).

We performed a comparison of analyzed patients versus patients excluded due to missing data and the results are shown in Table S12. Excluded patients were mainly less severe (higher Glasgow scores and lower SAPS II scores), with a higher proportion of women, and a different distribution of hospital care centers. Table S13 provides a comparison between the control and barbiturate-treated groups.

Sixty-six patients in the group administered barbiturates died in ICU compared to 256 in the control group. One can note that only six patients in the treated group (4.7%) were over 70 years old versus 126 (12.7%) in the control group (Figure S15). The age ranged was from 19 to 90 years old in the control group versus 19 to 76 years old in the treated group. One can explain a near-violation of the positivity because of two main reasons. First, elderly patients have a lower probability of receiving last-line treatment for intracranial hypertension because of therapy limitations (Calland et al., 2012). Second, the treatment prevalence was small, resulting in only 127 patients with barbiturates and the possible sample-to-sample fluctuation.

4.3 | Marginal effects estimates

In situations where the age-related near-violation of the positivity concerned 10% of the sample, we first performed an analyses of the overall sample. Next, we restricted the inclusion of patients to those younger than 70 years old. Figure S16 confirms that the patient age, for which we described the positivity violation, was associated with in-hospital mortality. The results are presented in Table 1 and plotted in Figure S17.

By observing the entire sample results, one can notice significant differences between the different methods. The most extreme effects were obtained with the truncated methods, while the previous simulation-based results highlighted their higher bias. More precisely, the truncated IPW (10th and 90th percentiles) had the highest OR (2.909, 95%CI from 1.990 to 4.254), while the truncated TMLE (bounds at 0.1 and 0.9) had the lowest OR (1.043, 95%CI from 0.814 to 1.338). The IPW and the TMLE were the two methods with the highest variance (0.362 and 0.299, respectively). The techniques with the lowest variances were the truncated approaches (0.127 for truncated TMLE and 0.194 for truncated IPW). Only the methods based on the TMLE have a 95% CI for the OR incorporating the value 1.

By comparing the results obtained from the entire sample with those reduced to patients under 70 years old, one can note relative stability in the estimates achieved by the five methods. Nevertheless, the estimations did not vary in the same direction: a slight increase between the estimations performed on the entire sample versus those in the subgroup for the GC, IPW, and truncated IPW, and a modest decrease in values for the TMLE-based methods. The methods with the closest results between the entire cohort and the sub-sample were based on the outcome model (TMLE, truncated TMLE, and GC). We reported a more considerable difference for IPW and truncated IPW. For instance, the OR obtained with truncated TMLE varied from 1.043 (95%CI from 0.814 to 1.338) to 1.047 (95%CI from 0.791 to 1.396), whereas the values obtained with IPW ranged from 2.158 (95%CI from 1.060 to 4.390) to 2.237 (95%CI from 1.082 to 4.624). Population restriction leads to an increase in variance, especially for TMLE-based methods (Figure S17).

The conclusions that can be drawn from the 95%CI did not change between the overall population and the restricted population. However, one can note that only the TMLE-based methods resulted in nonsignificant statistical effects, that is, rendering the study statistically "inconclusive," in contrast to the results obtained by the other methods.

TABLE 1 Results obtained by using g-computation (GC), inverse probability weighting (IPW), truncated IPW, targeted maximum likelihood estimator (TMLE), and truncated TMLE for estimating barbiturates effects

AtlanREA cohort: the barbiturates effect							
	π_1	π_0	**Δ**	**log(OR)**	**SD**	**OR**	**95%CI OR**
Whole sample							
GC	0.396	0.272	0.124	0.560	0.194	1.750	1.207 - 2.524
IPW	0.447	0.273	0.174	0.769	0.362	2.158	1.060 - 4.390
Truncated IPW [10–90%]	0.515	0.267	0.248	1.068	0.194	2.909	1.990 - 4.254
Truncated IPW [5–95%]	0.470	0.271	0.199	0.872	0.236	2.391	1.506 - 3.795
Truncated IPW [2.5–97.5%]	0.467	0.272	0.195	0.853	0.245	2.347	1.453 - 3.792
TMLE	0.320	0.289	0.031	0.146	0.299	1.158	0.645 - 2.079
Truncated TMLE [0.1–0.9]	0.298	0.288	0.010	0.043	0.127	1.043	0.814 - 1.338
Truncated TMLE [0.05–0.95]	0.311	0.289	0.022	0.107	0.156	1.112	0.820 - 1.509
Truncated TMLE [0.025–0.975]	0.311	0.289	0.022	0.108	0.219	1.114	0.725 - 1.711
Restricted sample							
GC	0.370	0.243	0.127	0.606	0.196	1.833	1.259 - 2.670
IPW	0.418	0.243	0.175	0.805	0.371	2.237	1.082 - 4.624
Truncated IPW [10–90%]	0.499	0.238	0.261	1.160	0.203	3.188	2.142 - 4.746
Truncated IPW [5–95%]	0.447	0.242	0.205	0.932	0.242	2.539	1.579 - 4.082
Truncated IPW [2.5–97.5%]	0.447	0.243	0.204	0.902	0.250	2.464	1.509 - 4.025
TMLE	0.280	0.263	0.017	0.090	0.344	1.094	0.558 - 2.145
Truncated TMLE [0.1–0.9]	0.272	0.263	0.009	0.046	0.143	1.047	0.791 - 1.386
Truncated TMLE [0.05–0.95]	0.274	0.263	0.011	0.056	0.183	1.058	0.739 - 1.515
Truncated TMLE [0.025–0.975]	0.275	0.263	0.012	0.066	0.269	1.068	0.630 - 1.809

Abbreviations: π_1, the expected proportions of event if the entire population is exposed; π_0, the expected proportions of event if the entire population is unexposed; Δ, the risk difference (π_1 - π_0); log(OR), the logarithm of the odds ratio; SD, the standard deviation for the logarithm of the odds ratio; OR, the corresponding odds-ratio ; 95%CI, 95% confidence interval of the odds-ratio.

5 | DISCUSSION

The results of the simulations illustrated that the near-violation of the positivity assumption could impact the bias and precision of the five methods. In terms of MAB, one can conclude that methods based on the outcome modeling showed the best results. The addition of an extrapolation issue altered the MAB for these methods, but in a magnitude similar to the one observed for the IPW-based approaches. Whilst the truncated methods introduced bias, they reduced the variance estimation, as previously described by Moore et al. (2012). Methods with the best balance between variance and bias were truncated IPW and GC. TMLE-based methods were associated with an over-optimistic estimate of the variance, resulting in lower coverage than the nominal value. We did not observe this issue when the prevalence of exposure was 50%, that is, reducing the positivity near-violation. Although the TMLE is a doubly robust estimator consistent when at least one nuisance model is well-specified, the variance estimation can be challenging. Petersen et al. (2014) and Lendle et al. (2017) reported the potential inflation of the type I error and poor coverage in the presence of positivity near-violations. Our results confirm their findings and the potential of a bootstrap-based approach as an alternative. We performed additional simulations with the average exposure effect as the estimand (instead of the logOR), our results were consistent (data not shown).

Whilst the simulations illustrated important differences between each methods performance, the "real-dataset" application emphasized the importance of the method chosen. Indeed, the clinical conclusion varied according to the specific method. In agreement with the simulations, the variances of truncated methods were smaller, but this benefit has to be counterbalanced with the risk of bias (Cole & Hernán, 2008; Ju et al., 2019). The main concern lies in their optimal cut-off choice, giving us the best bias-variance trade-off. We have studied consensual thresholds, defined either by a bound value of PS (for TMLE) or by the value of a percentile of the weights (for IPW). An alternative would lie in establishing an algorithm seeking the best bias-variance trade-off, which would be guided by the data. This solution has recently been studied to choose data-driven PS truncation thresholds adapted to IPW (Bembom & van der Laan, 2008) or to TMLE (Ju et al.,

2019), with promising results for positivity violation situations. Another solution may also lie in the use of modern methods such as limited overlap, matching and entropy weightings to reduce the influence of the most extreme observations and focus on the data area with the most overlap, therefore capturing the processing effect for which we have the most information (Zhou et al., 2020). These techniques enable us to estimate an average treatment effect on the population overlap (Li et al., 2018).

Causal inference in observational studies relies directly on the assumption that all participants are eligible to be exposed (or unexposed). Our results confirmed the importance of this assumption since all the methods compared were affected in terms of bias and/or variance. This assumption's violation is more identifiable by using PS-based approaches as it consists of regressing the exposure probability. In contrast, GC involves outcome modeling, and this violation can remain unidentified (Kang & Schafer, 2007). For IPW, subjects who have a low likelihood of exposure but who are exposed, results in extreme weights with unstable estimations and high variances (Kang & Schafer, 2007). The inflated variance and the associated extreme weighting obtained in this way can alert investigators. Unfortunately, the situations at risk of extrapolation are not directly identifiable, and only the violations of positivity can be revealed.

The near-violation of the positivity represents an obstacle to causal inferences only when it concerns true confounders, that is, those associated with both the exposure and the outcome (Westreich & Cole, 2010). In contrast, imbalance of variables was only associated with exposure, also called instrumental variables, and will have no impact on the bias.

Several authors have previously documented different techniques for detecting restrictions on the positivity assumption in the context of PS analysis (Austin & Stuart, 2015; Cole & Hernán, 2008). The first approach is to study the distribution of the exposure regimen for each covariate, but this can become tedious when dealing with many covariates. One can also use standardized differences (Austin & Stuart, 2015). Another possibility is to compare a groups weights distributions, or even to focus on the distribution of PS. Histogram of the PS distribution by exposure group is an example of interesting representation. In practice, one can assess the positivity assumption by searching for a lack of sufficient overlap of the PS distributions between the exposure groups. However, while useful to diagnose potential positivity violations, these techniques do not provide any quantitative estimate of the estimator bias due to positivity near-violation. Petersen et al. (2012) proposed a parametric bootstrap approach to provide an optimistic bias estimate specifically targeted for positivity violations and near-violations.

Our study has several limitations. First, we only considered TMLE-based methods, while other DRE approaches exist, such as the augmented inverse probability of treatment weighting (A-IPTW) (Glynn & Quinn, 2010). We focused on TMLE because of its better stability compared to A-IPTW (Luque-Fernandez et al., 2018; Neugebauer & van der Laan, 2005; Porter et al., 2011). Second, we did not study the different methods for the construction of the model, as this would have multiplied the number of possible approaches to compare. For instance, an alternative to reduce the variance of TMLE is the collaborative TMLE (C-TMLE), which uses a sequential selection of covariates estimating PS (Lendle et al., 2013; Pirracchio et al., 2018; Porter et al., 2011). Machine learning techniques were also proposed for GC (Austin, 2012), or for PS-based methods (Pirracchio et al., 2015). The improvement of the methods we studied by machine learning techniques is an interesting perspective of our work, especially because it can help to reduce the problem of extrapolation. Third, our simulation-based study was not associated with theoretical justification, and it does not demonstrate which method is the best in all situations. Even though our results are in agreement with the current literature, additional studies are required, such as incorporating the extrapolation issue for patients with a higher susceptibility of positivity near-violation.

To conclude, our study illustrates that all the causal methods were sensitive to the near-positivity violation. Nevertheless, we reported the methods' robustness based on the outcome model (GC and TMLE), even with an extrapolation issue. The truncated method, whilst attractive in terms of variance reduction, should be used with caution due to the associated risk of increased bias. G-computation appears to present the best compromise when considering its ability to reduce the bias and its statistical power.

ACKNOWLEDGMENTS

The authors would like to thank the members of AtlanREA Group for their involvement in the study, the physicians who helped recruit patients, and all patients who participated in this study. We also thank the clinical research associates who participated in the data collection. The analyses and interpretation of these data are the responsibility of the authors. This work was partially supported by a public grant overseen by the French National Research Agency (ANR) to create the Common Laboratory RISCA (www.labcom-risca.com, reference: ANR-16-LCV1-0003-01). The funder had no role in study design, analyses, and interpretation of data, writing the report, or the decision to submit the report for publication.

CONFLICT OF INTEREST

The authors have declared no conflict of interest.

DATA AVAILABILITY STATEMENT

The data that supports the findings of this study are available in the supplementary material of this article.

OPEN RESEARCH BADGES

This article has earned an Open Data badge for making publicly available the digitally-shareable data necessary to reproduce the reported results. The data is available in the Supporting Information section.

This article has earned an open data badge "**Reproducible Research**" for making publicly available the code necessary to reproduce the reported results. Results in this article were not fully reproducible due to insufficient quality of code and documentation.

ORCID

Maxime Léger https://orcid.org/0000-0002-5447-6595

REFERENCES

Abdia, Y., Kulasekera, K. B., Datta, S., Boakye, M., & Kong, M. (2017). Propensity scores based methods for estimating average treatment effect and average treatment effect among treated: A comparative study. *Biometrical Journal*, *59*, 967–985.

Austin, P. C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behav Res*, *46*, 399–424.

Austin, P. C. (2012). Using ensemble-based methods for directly estimating causal effects: An investigation of tree-based G-computation. *Multivariate Behavioral Research*, *47*, 115–135.

Austin, P. C. (2013). The performance of different propensity score methods for estimating marginal hazard ratios. *Statistics in Medicine*, *32*, 2837–2849.

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*, 3661–3679.

Bang, H., & Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics*, *61*, 962–973.

Bembom, O., & van der Laan, M. J. (2008). Data-adaptive selection of the truncation level for inverse-probability-of-treatment-weighted estimators. U.C. Berkeley Division of Biostatistics Working Paper Series.

Calland, J. F., Ingraham, A. M., Martin, N., Marshall, G. T., Schulman, C. I., Stapleton, T., Barraco, R. D., & Eastern Association for the Surgery of Trauma (2012). Evaluation and management of geriatric trauma: An eastern association for the surgery of trauma practice management guideline. *Journal of Trauma and Acute Care Surgery*, *73*, S345–350.

Chatton, A., Le Borgne, F., Leyrat, C., Gillaizeau, F., Rousseau, C., Barbin, L., Laplaud, D., Léger, M., Giraudeau, B., & Foucher, Y. (2020). G-computation, propensity score-based methods, and targeted maximum likelihood estimator for causal inference with different covariates sets: A comparative simulation study. *Scientific Reports*, *10*, 9219.

Cole, S. R., & Hernán, M. A. (2008). Constructing inverse probability weights for marginal structural models. *American Journal of Epidemiology*, *168*, 656–664.

Crump, R. K., Hotz, V. J., Imbens, G. W., & Mitnik, O. A. (2009). Dealing with limited overlap in estimation of average treatment effects. *Biometrika*, *96*, 187–199.

Foucher, Y., Borgne, F. L., Dantan, E., Gillaizeau, F., Chatton, A., & Combescure, C. (2019). RISCA: Causal inference and prediction in cohort-based analyses. https://CRAN.R-project.org/package=RISCA.

Gayat, E., Resche-Rigon, M., Mary, J.-Y., & Porcher, R. (2012). Propensity score applied to survival data analysis through proportional hazards models: A Monte Carlo study. *Pharmaceutical Statistics*, *11*, 222–229.

Glynn, A. N., & Quinn, K. M. (2010). An introduction to the augmented inverse propensity weighted estimator. *Political Analysis*, *18*, 36–56.

Gruber, S., & van der Laan, M. J. (2012). TMLE: An R package for targeted maximum likelihood estimation. *Journal of Statistical Software*, *51*, 1–35. http://www.jstatsoft.org/v51/i13/. https://doi.org/10.18637/jss.v051.i13.

Hajage, D., Tubach, F., Steg, P. G., Bhatt, D. L., & De Rycke, Y. (2016). On the use of propensity scores in case of rare exposure. *BMC Medical Research Methodology*, *16*, 32.

Hernán, M. A., & Robins, J. M. (2020). *Causal Inference: What if?*. Chapman & Hall/CRC. https://www.hsph.harvard.edu/miguel-hernan/causal-inference-book/.

Ju, C., Schwab, J., & van der Laan, M. J. (2019). On adaptive propensity score truncation in causal inference. *Statistical Methods in Medical Research*, *28*, 1741–1760.

Kang, J. D. Y., & Schafer, J. L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science*, *22*, 523–539.

Le Borgne, F., Giraudeau, B., Querard, A. H., Giral, M., & Foucher, Y. (2016). Comparisons of the performance of different statistical tests for time-to-event analysis with confounding factors: Practical illustrations in kidney transplantation. *Statistics in Medicine*, *35*, 1103–1116.

Lendle, S. D., Fireman, B., & van der Laan, M. J. (2013). Targeted maximum likelihood estimation in safety analysis. *Journal of Clinical Epidemiology*, *66*, S91–98.

Lendle, S. D., Schwab, J., Petersen, M. L., & van der Laan, M. J. (2017). LTMLE: An R package implementing targeted minimum loss-based estimation for longitudinal data. *Journal of Statistical Software*, *81*, 1–21. https://www.jstatsoft.org/index.php/jss/article/view/v081i01.

Li, F., Morgan, K. L., & Zaslavsky, A. M. (2018). Balancing covariates via propensity score weighting. *Journal of the American Statistical Association*, *113*, 390–400.

Luque-Fernandez, M. A., Schomaker, M., Rachet, B., & Schnitzer, M. E. (2018). Targeted maximum likelihood estimation for a binary treatment: A tutorial. *Statistics in Medicine*, *37*, 2530–2546.

Moore, K. L., Neugebauer, R., van der Laan, M. J., & Tager, I. B. (2012). Causal inference in epidemiological studies with strong confounding. *Statistics in Medicine*, *31*, 1380–1404.

Morris, T. P., White, I. R., & Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, *38*, 2074–2102.

Neugebauer, R., & van der Laan, M. J. (2005). Why prefer double robust estimators in causal inference? *Journal of Statistical Planning and Inference*, *129*, 405–426.

Neugebauer, R., & van der Laan, M. J. (2006). G-computation estimation for causal inference with complex longitudinal data. *Computational Statistics & Data Analysis*, *51*, 1676–1697.

Pearl, J., Glymour, M., & Jewell, N. P. (2016). *Causal Inference in Statistics: A Primer*. John Wiley & Sons.

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

Petersen, M., Schwab, J., Gruber, S., Blaser, N., Schomaker, M., & van der Laan, M. (2014). Targeted maximum likelihood estimation for dynamic and static longitudinal marginal structural working models. *Journal of Causal Inference*, *2*, 147–185.

Pirracchio, R., & Carone, M. (2018). The balance super learner: A robust adaptation of the super learner to improve estimation of the average treatment effect in the treated based on propensity score matching. *Statistical Methods in Medical Research*, *27*, 2504–2518.

Pirracchio, R., Petersen, M. L., & van der Laan, M. (2015). Improving propensity score estimators' robustness to model misspecification using super learner. *American Journal of Epidemiology*, *181*, 108–119.

Pirracchio, R., Yue, J. K., Manley, G. T., van der Laan, M. J., & Hubbard, A. E. (2018). Collaborative targeted maximum likelihood estimation for variable importance measure: Illustration for functional outcome prediction in mild traumatic brain injuries. *Statistical Methods in Medical Research*, *27*, 286–297.

Porter, K. E., Gruber, S., van der Laan, M. J., & Sekhon, J. S. (2011). The relative performance of targeted maximum likelihood estimator. *International Journal of Biostatistics*, *7*, 1557–4679.

R Core Team. (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing.

Robins, J. M. (1986). A new approach to causal inference in mortality studies with a sustained exposure period-application to control of the healthy worker survivor effect. *Mathematical Modelling*, *7*, 1393–1512.

Robins, J. M., Hernán, M. A., & Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology*, *11*, 550–560.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*, 41–55.

Schuler, M. S., & Rose, S. (2017). Targeted maximum likelihood estimation for causal inference in observational studies. *American Journal of Epidemiology*, *185*, 65–73.

Snowden, J. M., Rose, S., & Mortimer, K. M. (2011). Implementation of G-computation on a simulated data set: Demonstration of a causal inference technique. *American Journal of Epidemiology*, *173*, 731–738.

van der Laan, M. J., & Gruber, S. (2010). Collaborative double robust targeted maximum likelihood estimation. *The International Journal of Biostatistics*, *6*, 17–17.

van der Laan, M. J., & Robins, J. M. (2003). *Unified Methods for Censored Longitudinal Data and Causality*. Springer.

van der Laan, M. J., & Rubin, D. B. (2006). Targeted maximum likelihood learning. *The International Journal of Biostatistics*, *2*, https://doi.org/10.2202/1557-4679.1043.

VanderWeele, T. J., & Shpitser, I. (2011). A new criterion for confounder selection. *Biometrics*, *67*, 1406–1413.

Vansteelandt, S., & Keiding, N. (2011). Invited commentary: G-computation-lost in translation? *American Journal of Epidemiology*, *173*, 739–742.

Westreich, D., & Cole, S. R. (2010). Invited commentary: Positivity in practice. *American Journal of Epidemiology*, *171*, 674–677; discussion 678–681.

Williamson, E. J., Morley, R., Lucas, A., & Carpenter, J. (2012). Propensity scores: From naïve enthusiasm to intuitive understanding. *Statistical Methods in Medical Research*, *21*, 273–293.

Xu, S., Ross, C., Raebel, M. A., Shetterly, S., Blanchette, C., & Smith, D. (2010). Use of stabilized inverse propensity scores as weights to directly estimate relative risk and its confidence intervals. *Value Health*, *13*, 273–277.

Zeileis, A. (2006). Object-oriented computation of sandwich estimators. *Journal of Statistical Software*, *16*, 1–16.

Zhou, Y., Matsouaka, R. A., & Thomas, L. (2020). Propensity score weighting under limited overlap and model misspecification. *Statistical Methods in Medical Research*, *29*, 3721–3756.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Léger, M., Chatton, A., Le Borgne, F., Pirracchio, R., Lasocki, S., Foucher, Y. (2022). Causal inference in case of near-violation of positivity: comparison of methods. *Biometrical Journal. 64,* 1389–1403. https://doi.org/10.1002/bimj.202000323