Ress, Vanessa; Wild, Eva-Maria

Comparing methods for estimating causal treatment effects of administrative health data: A plasmode simulation study

Health Economics

Provided in Cooperation with:
John Wiley & Sons

Suggested Citation: Ress, Vanessa; Wild, Eva-Maria (2024) : Comparing methods for estimating causal treatment effects of administrative health data: A plasmode simulation study, Health Economics, ISSN 1099-1050, Wiley Periodicals, Inc., Hoboken, NJ, Vol. 33, Iss. 12, pp. 2757-2777, https://doi.org/10.1002/hec.4891

Received: 4 July 2023 | Revised: 12 August 2024 | Accepted: 24 August 2024

DOI: 10.1002/hec.4891

RESEARCH ARTICLE

Health Economics WILEY

Comparing methods for estimating causal treatment effects of administrative health data: A plasmode simulation study

Vanessa Ress[1,2] | **Eva-Maria Wild**[1,2]

[1]Department of Health Care Management, University of Hamburg, Hamburg, Germany

[2]Hamburg Center for Health Economics (HCHE), Hamburg, Germany

Correspondence
Eva-Maria Wild, Hamburg Center for Health Economics (HCHE), University of Hamburg, Esplanade 36, Hamburg 20354, Germany.
Email: eva.wild@uni-hamburg.de

Funding information
Innovation Fund of the German Federal Joint Committee

Abstract

Estimating the causal effects of health policy interventions is crucial for policymaking but is challenging when using real-world administrative health care data due to a lack of methodological guidance. To help fill this gap, we conducted a plasmode simulation using such data from a recent policy initiative launched in a deprived urban area in Germany. Our aim was to evaluate and compare the following methods for estimating causal effects: propensity score matching, inverse probability of treatment weighting, and entropy balancing, all combined with difference-in-differences analysis, augmented inverse probability weighting, and targeted maximum likelihood estimation. Additionally, we estimated nuisance parameters using regression models and an ensemble learner called superlearner. We focused on treatment effects related to the number of physician visits, total health care cost, and hospitalization. While each approach has its strengths and weaknesses, our results demonstrate that the superlearner generally worked well for handling nuisance terms in large covariate sets when combined with doubly robust estimation methods to estimate the causal contrast of interest. In contrast, regression-based nuisance parameter estimation worked best in small covariate sets when combined with singly robust methods.

KEYWORDS
administrative health care data, best practice, causal inference, simulation

1 | INTRODUCTION

Evaluating health policy interventions is crucial for optimizing their effectiveness and efficiency, as well as for identifying and mitigating any unintended health and economic consequences that may arise (Husereau et al., 2022; Luyten et al., 2022). Such evaluations also play a broader role in ensuring the efficient allocation of scarce resources and informing future policymaking (Clarke et al., 2019). Estimating causal treatment effects is essential for achieving these

goals (Crown, 2019). However, this often requires the use of administrative health care data, which presents several methodological challenges, such as a lack of randomization, unobserved heterogeneity, and the presence of a wide range of variables with complex and unknown dependencies. To address these challenges, various methods have been developed, including propensity score techniques (Rosenbaum & Rubin, 1983), difference-in-differences (DiD) analysis (Card & Krueger, 1993) and doubly robust semi-parametric methods (Robins et al., 1994; van der Laan & Rubin, 2006). However, despite continuous efforts to develop and refine these methods, our understanding of their performance in estimating causal effects based on administrative health care data remains limited.

To address this deficit, Monte Carlo simulation studies have been conducted to evaluate and compare the performance of different methods (Radice et al., 2012; Kreif et al., 2013; Kreif, Gruber, et al., 2016; O'Neill et al., 2016; Hwang et al., 2017). For instance, Radice et al. (2012) compare propensity score methods and genetic matching; Kreif et al. (2013) compare doubly robust methods to regression and propensity score methods; Kreif, Gruber, et al. (2016) compare a targeted maximum likelihood estimation (TMLE) with bias-corrected matching utilizing the superlearner for estimating the propensity score and regression function; Hwang et al. (2017) compare a new approach for estimating mean lifetime healthcare costs to a popular approach; and O'Neill et al. (2016) compare DiD estimation to the synthetic control method, a regression approach, and matching. Many simulation studies, however, are characterized by relatively simple confounding structures with few variables, leading to varying results depending on the data structure modeled and the methods under consideration (Franklin et al., 2014). Because the optimal choice for an estimation strategy depends on the research question, data features, population characteristics and method assumptions, simulation results are only applicable to the specific simulation setting (Varga et al., 2023). Moreover, these studies cannot accommodate the complexity and characteristics of real-world administrative health care data. As a result, researchers working with such data face a lack of guidance when it comes to selecting the most suitable method for their analysis.

Plasmode simulations have been proposed as an alternative to traditional simulations (Franklin et al., 2014; Vaughan et al., 2009). In a plasmode simulation, the covariates from a real dataset are used without alteration, while the values for the outcome variables are simulated based on the estimated associations between covariates and outcomes from the original data, ensuring that the true effect size is known. The advantage of this approach is that it preserves the high-dimensional and complex covariate structure of the source data, providing a simulation environment that closely resembles real-world conditions (Ripollone et al., 2020). Some previous plasmode simulations comparing the performance of various methods are either based on other settings, such as Sekhon & Grieve, 2012, whose simulation uses data from randomized controlled trials (RCTs), or focus on specific parts of the estimation strategy, such as Jones et al., 2015, who model health care costs only. Other plasmode simulations have compared the performance of various methods using administrative health care data but have focused mainly on continuous normally distributed (e.g., Meng & Huang, 2021) or binary outcomes (e.g., Franklin et al., 2014). Consequently, there remains a lack of guidance when it comes to choosing the most appropriate method for estimating causal effects for outcomes with other distributions commonly encountered in health economic evaluations, such as count data (e.g., the number of physician visits) and non-normally distributed continuous outcomes (e.g., expenditures) (Baxter et al., 2018).

To address this important research gap and provide practical suggestions for evaluating health policy interventions, we conducted a simulation study using a plasmode dataset derived from real-world administrative health care data. The data originate from a policy initiative launched in 2017 to improve access to care, optimize the use of resources and reduce costs in a deprived urban area in Germany. Using this dataset to simulate real-world conditions, our study evaluated and compared a range of methods for estimating causal effects across outcomes with different distributions. Our aim in doing so was to identify which causal estimation methods performed best in settings similar to that in our study. We focused on count data (number of physician visits), non-normally distributed continuous data (total health care costs), and binary data (indicator for hospitalization).

For our analysis, we selected five causal estimation methods that can be categorized into two groups. The first group comprises three commonly used methods in the field of health economics, namely propensity score matching, inverse probability of treatment weighting (IPW), and entropy balancing, all used in combination with a DiD framework. These have been employed in the case of propensity score matching by researchers such as Schreyögg et al. (2011), Fu et al. (2017), Strumpf et al. (2017), Somé et al. (2019), and Flawinne et al. (2023); in the case of IPW by researchers such as Nasseh et al. (2017), Strumpf et al. (2017), Sarma et al. (2018), Urwin et al. (2021), and Bijwaard (2022); and in the case of entropy balancing by Marcus (2013), Somé et al. (2019), Aranda et al. (2021), Bäuml et al. (2023), and Urwin et al. (2023). The second group of approaches consists of the doubly robust methods known as augmented inverse probability weighting (AIPW) and TMLE. These have been recommended in simulation studies, such as those conducted by Schuler and Rose (2017), Naimi et al. (2021), and Zivich and Breskin (2021). We deployed various simulation scenarios, which involved varying the set of covariates used for analysis and the approach employed for estimating the

nuisance parameters. In each scenario, we estimated average treatment effects on the treated (ATT) for count data and continuous outcomes and odds ratios (OR) for binary outcomes. These estimands are widely used in the field and are relevant to policymakers, underscoring the need for practical guidance in their use for evaluations. Lastly, we conclude with some suggestions based on our findings that may be useful for designing estimation strategies when conducting health economic evaluations based on administrative health care data.

The rest of this paper is organized as follows: Section 2 provides a brief introduction to the estimands of interest and the estimators compared in this study. Section 3 describes the data used for the plasmode simulation, the simulation approach employed and the specific scenarios that were simulated. Section 4 presents the results of the simulation, highlighting the performance of the different methods. Lastly, Section 5 discusses the results and provides suggestions that may be useful for researchers seeking to estimate treatment effects in causal analysis of administrative health care data.

2 | METHODS

2.1 | Target parameters

We used the potential outcome framework by Rubin (1974) to define our causal estimands of interest. Let (X_i, Y_i, Z_i) be the data of the ith subject in an independent and identically distributed dataset containing n subjects. $Y_i \in \mathbb{R}$, where $\mathbb{R}$ denotes the set of real numbers, is the observed outcome, $Z_i \in \{0, 1\}$ is a binary indicator for the received treatment, and $X_i \in \mathbb{R}^d$ denotes the d covariates of the ith subject. Furthermore, the potential outcomes for subject i, which represent the outcomes that subject i would have under the control treatment and the treatment of interest, respectively, given the same circumstances, are defined as $(Y_i(0), Y_i(1))$.

The average treatment effect (ATE) is then defined as

$$\tau_{ATE} = \mathrm{E}[Y_i(1) - Y_i(0)],$$

the ATT as

$$\tau_{ATT} = \mathrm{E}[Y_i(1) - Y_i(0) |\, Z_i = 1]$$

and the marginal causal OR for a binary outcome as

$$\tau_{OR} = \frac{\mathrm{E}[Y_i(1) = 1]}{1 - \mathrm{E}[Y_i(1) = 1]} \Big/ \frac{\mathrm{E}[Y_i(0) = 1]}{1 - \mathrm{E}[Y_i(0) = 1]}.$$

Note that the ATE and the OR measure the effect of the intervention for the whole population under consideration, meaning they represent the effect if every subject were treated. In contrast, the ATT measures the effect on those who were actually treated. Because the potential outcomes are subject to uncertainty and only one can be observed for each subject (Pearl, 2009), certain methods and assumptions are necessary to overcome this fundamental problem of causal inference (Holland, 1986).

2.2 | Propensity score: Matching and IPW

The propensity score, which we denote as $\mathrm{e}(X_i) = \mathrm{P}(Z_i = 1|X_i)$, is the probability of receiving the treatment of interest given the observed covariates X. In the presence of unconfoundedness, as well as positivity and consistency assumptions, the propensity score can be used to control for confounding and calculate unbiased estimators of treatment effects. The assumptions are defined as follows:

- Conditional exchangeability/Unconfoundedness: $\{Y_i(1), Y_i(0)\} \perp Z_i|X_i$
- Positivity: $0 < \mathrm{P}(Z_i = 1|X_i) < 1$
- Consistency: $Y_i = (1 - Z_i)Y_i(0) + Z_i Y_i(1)$.

In addition to the identifying assumptions above, the propensity score model ê needs to be correctly specified. The propensity score must be estimated based on the available data and can be referred to as a nuisance parameter because it is not the primary parameter of interest. This estimation is commonly performed using logistic regression models, although machine learning approaches can also be employed.

One method that uses the propensity score is known as matching. The fundamental concept of matching is to identify subjects in the control group who closely resemble the treated subjects. In nearest neighbor matching (Caliendo & Kopeinig, 2008), a distance is calculated between the propensity score of each treated and control subject, and, subsequently, each treated subject is matched with the closest control subject. As a result, only a subset of subjects who are similar with respect to their propensity scores are used for effect estimation. The expected outcomes are identified by

$$\mathrm{E}[Y_i(z)] = \frac{1}{m} \sum_{i:Z_i=z,\ i\in M} Y_i,$$

where m is the number of matched subjects in each treatment group and M indexes the set of matched subjects.

Another common method based on the propensity score is IPW. In IPW, weights are assigned to subjects to construct similar treatment and control populations. These weights are determined by the inverse of their propensity scores $w_i = \frac{Z_i}{\mathrm{e}(X_i)} + \frac{1-Z_i}{1-\mathrm{e}(X_i)}$ (Horvitz & Thompson, 1952; Robins et al., 2000). For estimating the ATT, the weight is multiplied by $e(X_i)$, resulting in treatment subjects having a weight of one and acting as the reference group (Morgan & Todd, 2008). The expected outcomes are identified by

$$\mathrm{E}[Y_i(z)] = \frac{1}{n} \sum_{i:Z_i=z} w_i Y_i.$$

2.3 | Entropy balancing

A generalization of the propensity score weighting method is entropy balancing (Hainmueller, 2012), which aims to achieve covariate balance directly instead of relying on a single score such as the propensity score. Under unconfoundedness, positivity and consistency assumptions, this is done by adjusting the weights assigned to the control group to satisfy a set of balancing conditions, while the subjects in the treated group are assigned a weight of one. The weights for estimation of the ATT are computed by solving

$$\hat{w} = \underset{w}{\operatorname{argmin}} \sum_{i:Z_i=0} w_i \log(w_i) \text{ subject to}$$

$$\sum_{i:Z_i=0} w_i \mathrm{c}_j(X_i) = \frac{1}{n_1} \sum_{i:Z_i=1} \mathrm{c}_j(X_i) \text{ for a set of functions } \mathrm{c}_j, j \in \{1, \cdots, J\} \text{ and}$$

$$\sum_{i:Z_i=0} w_i = 1,\ w_i \geq 0, i = 1, ..., n \text{ where } n_1 \text{ is the number of subjects in the treated group.}$$

The expected outcomes are identified by

$$\mathrm{E}[Y_i(0)|Z_i = 1] = \sum_{i:Z_i=0} w_i Y_i \text{ and } \mathrm{E}[Y_i(1)|Z_i = 1] = \sum_{i:Z_i=1} \frac{1}{n_1} Y_i.$$

2.4 | Difference-in-differences

The aforementioned methods are often used as preprocessing steps to improve covariate balance between the treatment and control groups before estimating the treatment effect in the preprocessed data, such as in matched or weighted

samples (Hainmueller, 2012). For longitudinal data, one commonly employed method is the DiD approach, which compares the average change over time in the outcome for the treatment group to the average change over time for the control group. This allows for selection on unobservables, assuming that the selection bias is additive and time-invariant, and that the parallel trend assumption holds (i.e., the two groups would have exhibited parallel trends in the absence of treatment conditional on observed covariates):

- Conditional parallel trends: $\mathrm{E}[Y_{i,1}(0) - Y_{i,0}(0)|Z_i = 1, X] = \mathrm{E}[Y_{i,1}(0) - Y_{i,0}(0)|Z_i = 0, X]$, where $t = 0$ before treatment and $t = 1$ after treatment, let $Y_{i,t}(1)$ and $Y_{i,t}(0)$ be the respective outcomes for treatment and control subjects i at time t.

As a consequence, the strong unconfoundedness assumption posed in the earlier sections is not required, while the assumptions of positivity and consistency remain (Abadie, 2005).

The DiD estimates the ATE on the treated via the regression

$$Y_{i,t} = \alpha + \beta Z_i t + \alpha_i + \alpha_t + \varepsilon_{i,t}$$

where α_i are individual fixed effects, α_t are time fixed effects, and $\varepsilon_{i,t}$ the additive time-varying error term. $\hat{\tau}^{\mathrm{DiD}}_{\mathrm{ATT}} = \hat{\beta}$ captures the difference in change between the intervention and control groups and therefore the effect of the intervention. As a consequence, statistical assumptions of the linear regression model, such as linearity, additivity and model correctness, are required. In terms of potential outcomes, the DiD estimator is given by

$$\tau^{\mathrm{DiD}}_{\mathrm{ATT}} = \mathrm{E}\big[Y_{i,1}(1) - Y_{i,0}(1)\big|Z_i = 1\big] - \mathrm{E}\big[Y_{i,1}(0) - Y_{i,0}(0)\big|Z_i = 1\big].$$

The difference $\mathrm{E}[Y_{i,1}(0) - Y_{i,0}(0)|Z_i = 1]$ is not observable and therefore $\mathrm{E}[Y_{i,1}(0) - Y_{i,0}(0)|Z_i = 0]$ is used as the counterfactual.

2.5 | AIPW and TMLE

AIPW (Robins et al., 1994) and TMLE (van der Laan & Rubin, 2006) are methods that combine the propensity score model and the outcome model. These are considered doubly robust because their estimation is consistent if at least one of the two models is correctly specified. Computing the AIPW and TMLE estimators requires the estimation of the propensity score and the outcome model, which are referred to as nuisance parameters because they are not the primary parameters of interest. In addition to the identifying assumptions of unconfoundedness, positivity and consistency (Baumann et al., 2021; Glynn & Quinn, 2010), statistical assumptions linked to the estimation of the nuisance parameters may be needed.

The nuisance parameters can be estimated using parametric models, such as regression, or non-parametric models, such as random forests. The nuisance parameters should be of high quality, meaning that the predictors are consistent and have convergence rates of at least $n^{-1/4}$, where n is the sample size. The convergence rate implies that the square root of the mean squared error halves when the sample size is increased by a factor of 16; additionally, the predictors should be computed in an out-of-sample manner (e.g., using V-fold cross-fitting), where the predictors for individual observations are computed without including the observation itself (Chernozhukov et al., 2018). It should be noted that non-parametric approaches have slower convergence rates compared to parametric approaches. Therefore, when using non-parametric models, doubly robust estimators need to be employed (Naimi et al., 2021).

The AIPW estimator is a one-step correction estimator based on g-computation with a correction term based on the propensity score. First, the outcome model $\mathrm{m}(Z_i, X_i) = \mathrm{E}[Y_i|Z_i, X_i]$ for $Z_i = 0$ and $Z_i = 1$ and the propensity score model $\mathrm{e}(X_i)$ need to be estimated. Information on these models are then combined to calculate

$$\hat{\tau}^{\mathrm{AIPW}}_{\mathrm{ATE}} = \frac{1}{n}\sum_i \left(\hat{\mathrm{m}}(1, X_i) - \hat{\mathrm{m}}(0, X_i) + \frac{Z_i(Y_i - \hat{\mathrm{m}}(1, X_i))}{\hat{\mathrm{e}}(X_i)} - \frac{(1 - Z_i)(Y_i - \hat{\mathrm{m}}(0, X_i))}{1 - \hat{\mathrm{e}}(X_i)}\right)$$

and

$$\hat{\tau}_{\text{ATT}}^{\text{AIPW}} = \frac{1}{n}\sum_i \left(\frac{Z_i n}{\sum_j Z_j} (Y_i - \hat{\text{m}}(0, X_i)) - \frac{n(1-Z_i)\hat{\text{e}}(X_i)}{\sum_j Z_j (1-\hat{\text{e}}(X_i))} (Y_i - \hat{\text{m}}(0, X_i)) \right).$$

To use TMLE on continuous outcomes, these need to be transformed so that they are bounded within $(0, 1)$ (Gruber & van der Laan, 2012; van der Laan & Rose, 2011). Similar to AIPW, the first step of TMLE requires estimating the outcome model and the propensity score. The adjustment procedure, however, is slightly different from that in the AIPW and includes a targeting step for the initial estimate of the expected outcome. During the targeting step, a clever covariate $\text{H}(Z_i, X_i) = \frac{Z_i}{\hat{\text{e}}(X_i)} - \frac{1-Z_i}{1-\hat{\text{e}}(X_i)}$ and the fluctuation parameter $\hat{\varepsilon}$ are calculated. The fluctuation parameter indicates how to adapt the initial outcome estimates to incorporate information about the treatment model and is computed by solving $\text{logit}(Y_i) = \text{logit}(\hat{\text{m}}(Z_i, X_i)) + \varepsilon \text{H}(Z_i, X_i)$. Finally, the initial outcome estimates are updated by $\hat{\text{m}}^*(Z_i, X_i) = \text{expit}(\text{logit}(\hat{\text{m}}(Z_i, X_i)) + \hat{\varepsilon}\text{H}(Z_i, X_i))$ and the ATE is calculated as

$$\hat{\tau}_{\text{ATE}}^{\text{TMLE}} = \frac{1}{n}\sum_i [\hat{\text{m}}^*(1, X_i) - \hat{\text{m}}^*(0, X_i)].$$

To estimate the ATT, a different clever covariate for the outcome model and an additional clever covariate for the propensity score are needed. Moreover, an iterative procedure is necessary to update the propensity score and outcome model. For further information see Chapter 8 in van der Laan and Rose (2011). AIPW and TMLE can also be used to estimate target parameters such as the odds ratio, which is also a function of the two counterfactual probabilities $\text{E}[Y_i(0)]$ and $\text{E}[Y_i(1)]$ (van der Laan & Rose, 2011; Zhong et al., 2021).

2.6 | Superlearner

As mentioned above, nuisance parameters can be estimated using regression or machine learning approaches. When deciding whether to employ a machine learning approach, it is difficult to choose among the many algorithms available —a problem that is solved by the superlearner, which combines multiple algorithms (van der Laan et al., 2007). The superlearner is a weighted ensemble of multiple baselearners that can include both parametric and non-parametric approaches, offering protection against model misspecification (Naimi et al., 2021). It performs at least as well as the best baselearner included in the ensemble (van der Laan et al., 2007) and can reduce bias and improve covariate balance in the presence of model misspecification (Pirracchio et al., 2015). The estimation procedure using the superlearner involves the following steps (Polley & van der Laan, 2010; van der Laan et al., 2007):

1. Choose a set of baselearners $\mathcal{L} = \{\Psi_k : k = 1, ..., K\}$ and fit each baselearner on the entire dataset $\mathcal{D} = \{D_i = (Y_i, X_i) : i = 1, ..., n\}$ to form the ensemble of trained baselearners $\hat{\Psi}_k$. Each baselearner is trained on the entire dataset according to its respective methodology. For example, when using regression models as baselearners, this involves estimating their regression coefficients.
2. Split the data $\mathcal{D}$ into training and validation samples according to a V-fold cross-validation scheme, such that there are V equal-sized subsets. Let $\text{V}(v)$ be the vth validation set and the remaining data $\text{T}(v) = \mathcal{D} \backslash \text{V}(v)$ be the corresponding training dataset, $v = 1, ..., V$.
3. For the vth fold, fit each baselearner in $\mathcal{L}$ on $\text{T}(v)$ and save the trained baselearners as $\hat{\Psi}_{k,T(v)}$.
4. Determine the weights α that minimize the cross-validated risk of the candidate estimator $\sum_{k=1}^K \alpha_k \hat{\Psi}_k$ as

$$\hat{\alpha} = \underset{\alpha}{\text{argmin}} \frac{1}{V}\sum_{j=1}^{V} \frac{1}{|\text{V}(v)|} \sum_{i \in \text{V}(v)} \text{L}\left(D_i, \sum_{k=1}^{K} \alpha_k \hat{\Psi}_{k,\text{T}(v)} \right)$$

where $0 \leq \alpha_k \leq 1$ and $\sum_{k=1}^K \alpha_k = 1$ and loss function L. The goal of this step is to use, in the final estimator, the baselearners that best predict the data.

5. The final estimator is given by

$$\hat{\Psi}(X) = \sum_{k=1}^{K} \hat{\alpha}_k \hat{\Psi}_k(X).$$

For a more detailed description see the step-by-step guide provided in Polley et al. (2021) and the visual guide in Hoffman (2020).

3 | PLASMODE SIMULATION

In this section, we aim to answer our research question: Which of the five causal methods performs best when estimating treatment effects using administrative health care data in various realistic scenarios with outcomes from different distributions and varying sets of covariates, and how should the nuisance parameters be estimated in each setting?

3.1 | Source data and plasmode dataset generation

In our plasmode simulation study, we constructed a data-generating process based on real-world administrative health care data. Specifically, we used data from a policy initiative that was launched with the aim of improving access to care, optimizing the use of resources and reducing costs in a deprived urban area in Germany in 2017 (Ress & Wild, 2023). The data were provided by three statutory health insurers, covering the period from 2015 to 2019. The dataset encompassed comprehensive information on health care utilization and costs, prescriptions, nursing care and the demographic characteristics of $n = 556{,}911$ individuals across $p = 3.508$ covariates, with the intervention group consisting of $n_1 = 49{,}348$ individuals. See Appendix A for an overview.

To simplify the simulation task while still capturing the essential covariates, we employed the concept of the high-dimensional propensity score algorithm proposed by (Schneeweiss et al., 2009) and reduced the number of covariates by keeping only the most important diagnoses (based on ICD codes), procedures and diagnostics (based on OPS codes), and medications (based on ATC codes). Ultimately, this led us to select the 200 most important binary indicators for these variables as determined by the importance measure of gradient boosting (Friedman, 2001). A list of these 200 variables can be found in Appendix B.

These were then used to supplement our 32 handpicked covariates, which comprised demographic characteristics, information on participation in disease management and integrated care programs, incapacity to work and long-term care, utilization and costs of health services across different sectors, and additionally 31 binary indicators for the Elixhauser comorbidity groups (Elixhauser et al., 1998) (see Appendix A).

We used these 263 covariates together with the treatment indicator to simulate values for the outcome variables defined above while maintaining the associations between the covariates, and we generated new values for the outcome variables with a chosen effect size. The procedure to perform the plasmode simulation was as follows:

1. Estimate the association between treatment, outcome and covariates.
2. Use the estimated coefficients to predict the outcomes but modify the treatment coefficient to the desired effect size.
3. Draw J subsets of size s by resampling-with-replacement and perform steps 4 and 5 for each of those subsets.
4. Introduce noise by sampling the outcomes from suitable distributions using the simulated values from step 3 as expected values.
5. Analyze the simulated data.

We focused on three outcomes: (a) the number of outpatient physician visits (count data), (b) total health care cost (non-normally distributed continuous data), and (c) hospitalization (binary indicator). To estimate the association between treatment, outcome and covariates, we used a negative binomial regression for the observed number of outpatient physician visits as a function of treatment and covariates. To estimate the association between treatment,

covariates and total cost, we used a generalized linear model with gamma family and log-link. Finally, for hospitalization we estimated a logistic model. The goodness-of-fit for our models was quantified using Nagelkerke R^2 (Nagelkerke, 1991), yielding values of 0.380 for the outpatient physician visits, 0.167 for total cost and 0.199 for hospitalization. In all regressions, we modeled the covariates representing cost measures using natural cubic splines, which allowed us to model the non-linear effects of continuous variables while avoiding distributional assumptions (Perperoglou et al., 2019). We excluded interaction terms, maintaining only linear relationships in our model. We set the treatment coefficients to model a 5% reduction in costs and utilization and estimated the treated and untreated outcomes for each subject. Using these simulated counterfactual outcomes we calculated the true annual effects for the entire population. Thus, the true effect for the number of outpatient visits is given by $\tau_{\text{ATT}}^{\text{outpatient}} = -0.710$, the effect for total costs by $\tau_{\text{ATT}}^{\text{cost}} = -149.453$ and the effect for hospitalization by $\tau_{\text{OR}}^{\text{hospitalization}} = 0.95$. Next, we resampled $J = 1000$ subsets of size $s = 5000$ observations with replacement and drew the simulated number of outpatient visits from the Poisson distribution, the simulated total cost from the Gamma distribution, and the simulated binary indicator for hospitalization from the Bernoulli distribution using the previously estimated values as expected values.

This resulted in J subsets of size s, each containing a treatment vector Z, a matrix of covariates X and a vector containing the simulated outcome Y^* for each target measure. The data-generating mechanisms for X and Z were unknown, and any existing associations were preserved because we used the observed data without alteration. However, we created the outcomes Y^* in such a way that the data-generating mechanism, including the effect sizes for the treatment, was known. Here, the assumption of an independent and identically distributed dataset is probably not met because the simulated dataset is based on real data, and thus real-world conditions are simulated more realistically.

To simulate unobserved confounding, we dropped a subset of the confounders used to simulate the outcomes from the data used for analysis and defined these as unobserved confounders (Franklin et al., 2014). We used four different sets of covariates to vary both the level of unobserved confounding and the number of covariates used for analysis (see Appendix C for the strength of the relationships between covariate sets and outcomes). All four sets of covariates employed to control for confounding used our handpicked variables, and three of the four sets were enhanced by additional sets of covariates, as follows:

1. Handpicked covariates only (32 covariates)
2. Handpicked covariates enhanced by the indicators for the Elixhauser comorbidity groups (63 covariates)
3. Handpicked covariates enhanced by the binary indicators for the 100 most important ATC, ICD and OPS codes based on gradient boosting (132 covariates)
4. Handpicked covariates enhanced by all ICD codes among the 200 additional confounders used for simulation (204 covariates)

In the next step, we analyzed the data under different scenarios and obtained estimates of the treatment effect. Subsequently, we computed and compared the properties of the different estimation methods, as would be the case in ordinary simulation studies. We performed all computations using R version 4.1.2.

3.2 | Estimands

Our objective in this simulation study was to estimate the effect of exposure to the policy initiative (i.e., the treatment effect) based on whether or not individuals resided in the initiative's target area (i.e., a binary treatment variable). By comparing subjects who actually received the treatment with those who did not, the ATT represents an average effect on those who received the treatment and indicates what would have happened if they had not been treated, and thus may be the parameter of interest for policymakers (Heckman & Vytlacil, 2001; Wang et al., 2017). Moreover, the DiD estimator is widely used to estimate causal effects in the field of health economics and its parameter of interest is the ATT (Kreif, Grieve, et al., 2016). For these reasons, we estimated the ATT for the count data and continuous data outcomes. For the binary outcome, we estimated the marginal causal OR, which is also widely used in the field. The marginal causal odds ratio assesses the effect of a treatment across a population and thus is relevant to policymakers who are interested in uniformly applied policy decisions (Loux et al., 2017; Persson & Waernbaum, 2013). It should be

noted, however, that the use of the OR is generally discouraged due to problems of misinterpretation and comparability across studies (Norton et al., 2024).

3.3 | Simulation scenarios

We estimated the nuisance parameters, namely the propensity score and the outcome model, using two different approaches. For a baseline and comparison, we used the most common strategy, which is parametric linear regression. For the outcome models focusing on the number of outpatient physician visits and the total cost, we used OLS regression, whereas for hospitalization we used logistic regression. To estimate the propensity score, we also used logistic regression. In all models, we included only linear terms (no interactions). However, a problem when using regression, especially in settings with many covariates, is the need to specify a parametric model. To address this, we used the superlearner algorithm implemented in the *SuperLearner* package (Polley et al., 2021), which allowed us to incorporate non-parametric approaches. We included the following five algorithms as baselearners: generalized linear model with penalized maximum likelihood (*glmnet* function) (Friedman et al., 2010), random forest (*ranger* function) (Wright & Ziegler, 2017), gradient boosting (*xgboost* function) (Chen et al., 2015), support vector machines (*svm* function) (Cortes & Vapnik, 1995; Karatzoglou et al., 2006), and multivariate adaptive regression splines (*earth* function) (Friedman, 1991). All functions were used with the default parameters at the time of writing. Further information can be found in Appendix D. We used 10-fold cross-validation to estimate the baselearner weights, minimizing non-negative least squares normalized to one (van der Laan et al., 2007). We estimated the nuisance parameters out-of-sample by applying a two-fold cross-fitting procedure to estimate the target parameters on separate datasets from the nuisance models.

We then employed the five different causal methods described in the previous section to estimate causal treatment effects: propensity score matching (and subsequent DiD analysis), IPW (and subsequent DiD analysis), entropy balancing (and subsequent DiD analysis), AIPW and TMLE, see Table 1 for a summary. To estimate the ATT, we combined propensity score matching, IPW and entropy balancing with DiD, whereas we estimated the odds ratio on the matched/weighted samples directly. We matched on the propensity score using 1:3 nearest neighbor matching with the *MatchIt* function (Stuart et al., 2011). For IPW, we used the corresponding weights to calculate the ATT and ATE weights for the computation of OR (Pirracchio et al., 2012). For entropy balancing, we employed the *ebalance* function (Hainmueller, 2012) and used the ATT weights to estimate both the ATT and the odds ratio (Amusa et al., 2019). The DiD estimator was calculated using the built-in R functions *lm* and *glm*. Lastly, we performed AIPW using the *AIPW* function (Zhong et al., 2021) and TMLE using the *tmle* function (Gruber & van der Laan, 2012). Further information on the settings used for estimation can be found in Appendix D. For variance and confidence interval (CI) estimation, we used the defaults provided by the aforementioned functions.

3.4 | Performance measures

To assess the performance of the different causal estimation methods in the various scenarios described above, we compared their estimates to the true treatment effect across the population, which was known due to the simulation setup. We compared the performance with respect to bias, empirical standard error (SE) and 95% CI coverage. Bias is measured as the mean difference between the estimated and true treatment effect, reflecting the expected error of an estimation. As such, bias is a measure of the systematic deviance from the true effect and reflects the bias of the estimator. The empirical SE reflects the dispersion of the estimated effects around their mean, providing a measure of the precision of the estimator. Confidence interval coverage indicates the proportion of CIs that contain the true effect. Coverage below the confidence level reflects bias or CIs that are systemically too narrow, whereas overcoverage reflects CIs that are systemically too wide. Because the performance measures are themselves estimated, we report them with their corresponding Monte Carlo Standard errors (MCSEs) to quantify uncertainty. Formulas for the calculated performance measures and MCSEs can be found in Morris et al. (2019).

TABLE 1 Overview of simulation scenarios, as well as target estimands, nuisance parameters, and assumptions for each scenario.

Estimation method	Outcome	Target estimand	Nuisance parameter(s)	Identifying assumption(s)	Statistical assumption(s)
Matching+DiD	Number of outpatient visits, total health care cost	ATT	Propensity score	Conditional parallel trends, positivity, consistency	Modeling assumptions of nuisance parameter estimation method, statistical assumptions of linear regression model (e. g. linearity, additivity, correct model specification)
Matching	Binary indicator for hospitalization	Marginal OR	Propensity score	Unconfoundedness, positivity, consistency	Modeling assumptions of nuisance parameter estimation method
IPW+DiD	Number of outpatient visits, total health care cost	ATT	Propensity score	Conditional parallel trends, positivity, consistency	Modeling assumptions of nuisance parameter estimation method, statistical assumptions of linear regression model (e. g. linearity, additivity, correct model specification)
IPW	Binary indicator for hospitalization	Marginal OR	Propensity score	Unconfoundedness, positivity, consistency	Modeling assumptions of nuisance parameter estimation method
Entropy balancing+DiD	Number of outpatient visits, total health care cost	ATT		Conditional parallel trends, positivity, consistency	Statistical assumptions of linear regression model (e. g. linearity, additivity, correct model specification)
Entropy balancing	Binary indicator for hospitalization	Marginal OR		Unconfoundedness, positivity, consistency	
AIPW	Number of outpatient visits, total health care cost	ATT	Propensity score, outcome model	Unconfoundedness, positivity, consistency	Modeling assumptions of nuisance parameter estimation method
AIPW	Binary indicator for hospitalization	Marginal OR	Propensity score, outcome model	Unconfoundedness, positivity, consistency	Modeling assumptions of nuisance parameter estimation method
TMLE	Number of outpatient visits, total health care cost	ATT	Propensity score, outcome model	Unconfoundedness, positivity, consistency	Modeling assumptions of nuisance parameter estimation method
TMLE	Binary indicator for hospitalization	Marginal OR	Propensity score, outcome model	Unconfoundedness, positivity, consistency	Modeling assumptions of nuisance parameter estimation method

Abbreviations: AIPW, augmented inverse probability weighting; Entropy balancing(+DiD), entropy balancing (with subsequent difference-in-differences analysis); Matching(+DiD), propensity score matching (with subsequent difference-in-differences analysis); IPW(+DiD), inverse probability of treatment weighting (with subsequent difference-in-differences analysis); TMLE, targeted maximum likelihood estimation.

4 | RESULTS

MCSEs across all performance measures and scenarios were acceptable and thus allowed us to draw conclusions about the performance of the methods based on 1000 simulation iterations. All performance measures and their corresponding MCSEs can be found in Appendix E.

4.1 | Number of outpatient visits

Figure 1 presents the results for the ATT on the count-data outcome: the number of outpatient visits. First, we compared the performance of each of the five causal estimation methods while using either regression models or the superlearner to estimate the nuisance parameters. For IPW+DiD, AIPW and TMLE, we observed smaller bias and SE of the treatment effect when using the superlearner. For matching, we found that using the superlearner instead of regression models led to smaller bias, whereas the SE was larger. For entropy balancing, no nuisance parameters were

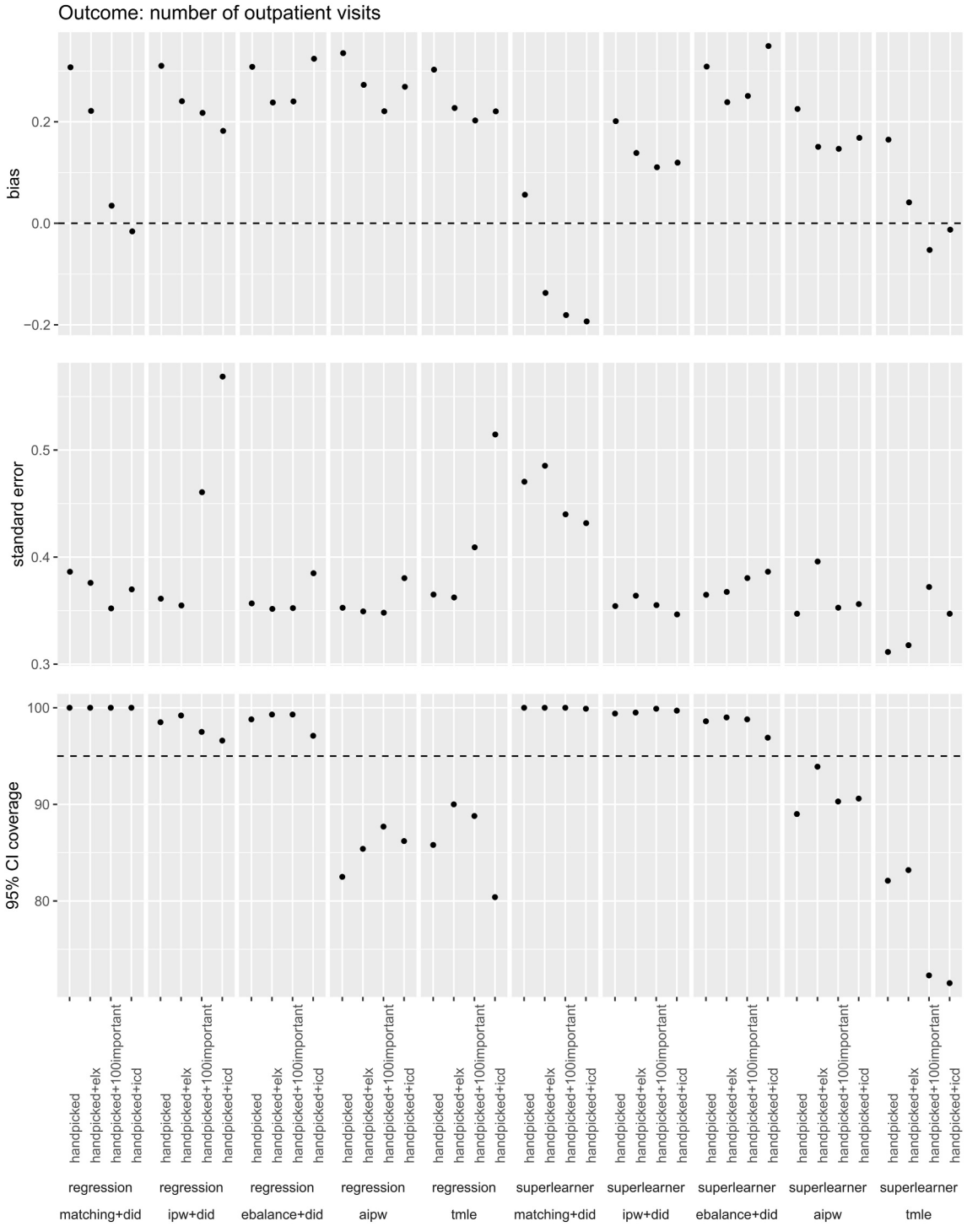

FIGURE 1 Performance measures for the ATT on the number of outpatient visits (true treatment effect $\tau_{\text{ATT}}^{\text{outpatient}} = -0.710$) for each causal method depending on the approach used for nuisance parameter estimation and the covariates used to control for confounding. The first panel shows the results for bias; the second panel the results for the empirical standard error; and the third panel the results for 95% CI coverage in percent. aipw, augmented inverse probability weighting; CI, confidence interval; ebalance+did, entropy balancing with subsequent difference-in-differences analysis; handpicked, handpicked covariates only (32 covariates); handpicked+elx, handpicked covariates enhanced by the indicators for the Elixhauser comorbidity groups (63 covariates); handpicked+icd, handpicked covariates enhanced by the ICD codes among the 200 additional confounders used for simulation (204 covariates); handpicked+100important, handpicked covariates enhanced by the binary indicators for the 100 most important ATC, ICD and OPS codes based on gradient boosting (132 confounders); ipw+did, inverse probability of treatment weighting with subsequent difference-in-differences analysis; matching+did, propensity score matching with subsequent difference-in-differences analysis; tmle, targeted maximum likelihood estimation.

estimated and thus performance measures were similar in both scenarios. Furthermore, the choice of the approach for nuisance parameter estimation did not influence CI coverage for most causal methods. However, for TMLE, we observed lower coverage when combined with the superlearner compared to regression. For AIPW CI coverage increased when using the superlearner.

Second, we compared the performance of each causal estimation method depending on the covariate set used to control for confounding. With regard to bias and CI coverage, there were no consistent relationships with the size of the covariate set used. For all causal estimation methods combined with regression except for propensity score matching, the SE increased considerably for the largest covariate set.

Overall, TMLE combined with the superlearner showed the smallest bias and SE. For TMLE and AIPW CI coverage was generally poor and below 95% in all scenarios, whereas CI coverage for the other methods was always above 95%. In some scenarios, other methods achieved the level of performance of TMLE in terms of bias and SE, such as propensity score matching+DiD with regression or IPW+DiD with the superlearner on big covariate sets, while exceeding the performance of TMLE with regard to CI coverage.

4.2 | Total health care cost

Subsequently, we examined the ATT on the non-normally distributed continuous outcome: total health care cost. The results are presented in Figure 2. We compared the performance of the five causal estimation methods based on the approach for estimating the nuisance parameters. For AIPW and TMLE, we observed smaller bias and SE when using the superlearner instead of regression models. When using the superlearner instead of regression models, CI coverage increased for AIPW and decreased for TMLE but remained below 95% in all scenarios. In the case of propensity score matching+DiD, we found that using the superlearner led to larger bias and SE. Regarding IPW+DiD, there was no clear relationship between SE and the choice of the estimation approach for the nuisance parameters, whereas bias was smaller when combined with the superlearner. Propensity score matching+DiD and IPW+DiD showed similar performance in terms of CI coverage regardless of whether regression models or the superlearner were used.

When comparing the performance of the five causal estimation methods with regard to the covariate set used to control for confounding, we observed that using the superlearner for nuisance parameter estimation tended to result in smaller bias and SE when controlling for larger covariate sets. Conversely, when using regression models, the opposite trend was observed, with SE tending to increase or stay unchanged as the size of the covariate set increased for all methods except propensity score matching+DiD. With regard to bias, no consistent relationship between covariate set size and performance was observed for methods used in combination with regression for nuisance parameter estimation. The choice of the covariate set did not consistently affect CI coverage across the different causal estimation methods.

Among the different methods, TMLE in combination with the superlearner demonstrated the smallest bias and SE when performed on large covariate sets. However, for TMLE and AIPW, CI coverage was below the 95% threshold, with especially low values seen for TMLE, where it ranged from 70.2% to 75.6%. In contrast, the CI coverage for propensity score matching+DiD, IPW+DiD and entropy balancing+DiD was greater than 95% in all scenarios, and performance regarding bias and SE was only slightly worse in some scenarios compared to the TMLE/superlearner combination. It is important to note that TMLE exhibited the worst performance among the five methods with regard to bias and SE when combined with regression models for nuisance parameter estimation instead of the superlearner.

4.3 | Binary indicator for hospitalization

Lastly, we examined the OR for the binary outcome: hospitalization. The results are presented in Figure 3. The performance of the different casual methods varied depending on the choice of approach taken to estimate the nuisance parameters. When we used the superlearner instead of regression models, AIPW exhibited smaller bias and SE. The same was observed for IPW and TMLE with regard to SE. However, for propensity score matching, the use of superlearner led to larger bias and SE. For IPW CI coverage was comparable regardless of the choice between regression models and the superlearner. For propensity score matching, AIPW and TMLE, CI coverage was lower when using the superlearner.

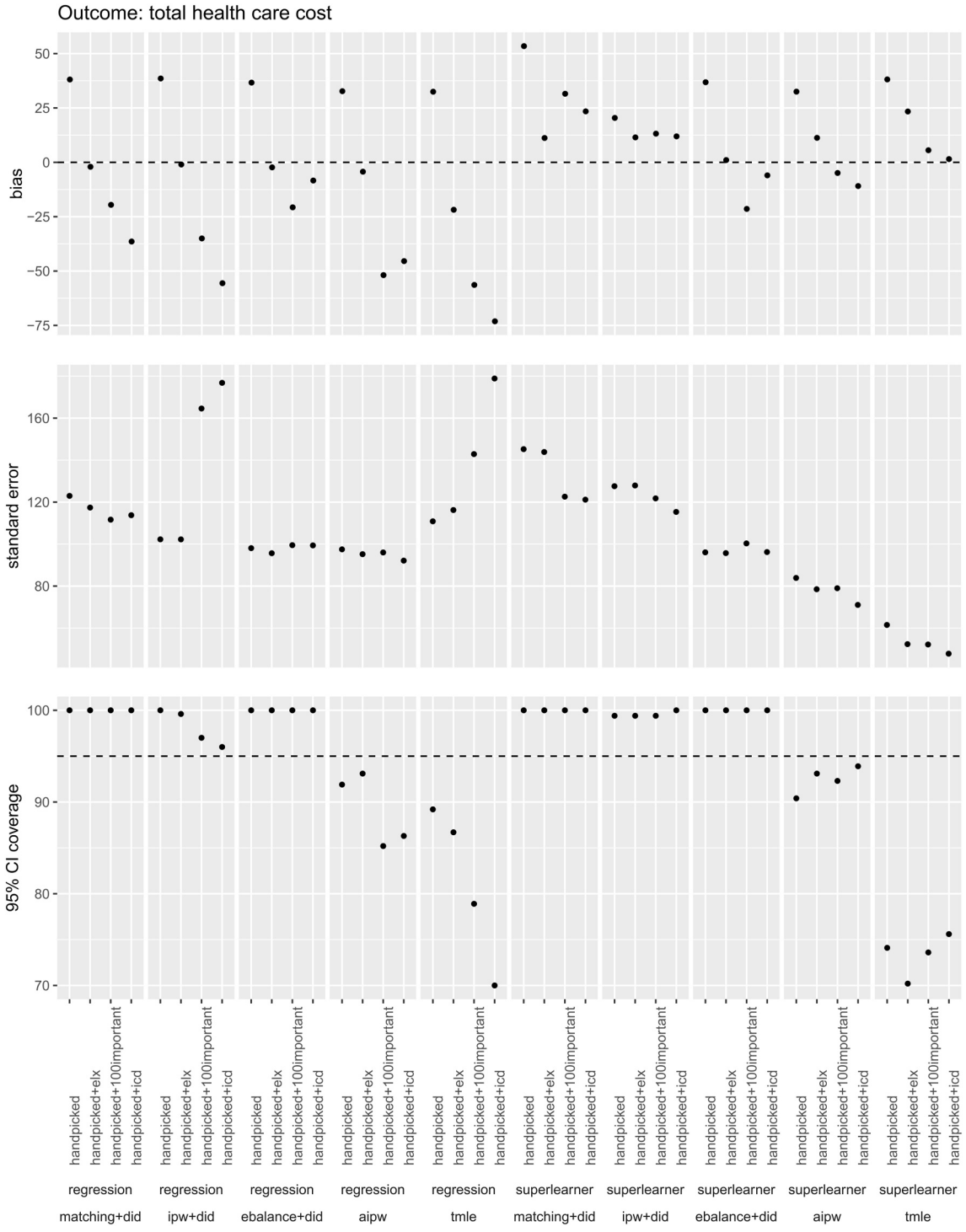

FIGURE 2 Performance measures for the ATT on total health care cost (true treatment effect $\tau_{\text{ATT}}^{\text{cost}} = -149.453$) for each causal estimation method depending on the approach used for nuisance parameter estimation and the covariates used to control for confounding. The first panel shows the results for bias; the second panel the results for the empirical standard error; and the third panel the results for 95% CI coverage in percent. aipw, augmented inverse probability weighting; CI, confidence interval; ebalance+did, entropy balancing with subsequent difference-in-differences analysis; handpicked, handpicked covariates only (32 covariates); handpicked+elx, handpicked covariates enhanced by the indicators for the Elixhauser comorbidity groups (63 covariates); handpicked+icd, handpicked covariates enhanced by the ICD codes among the 200 additional confounders used for simulation (204 covariates); handpicked+100important, handpicked covariates enhanced by the binary indicators for the 100 most important ATC, ICD and OPS codes based on gradient boosting (132 covariates); ipw+did, inverse probability of treatment weighting with subsequent difference-in-differences analysis; matching+did, propensity score matching with subsequent difference-in-differences analysis; tmle, targeted maximum likelihood estimation.

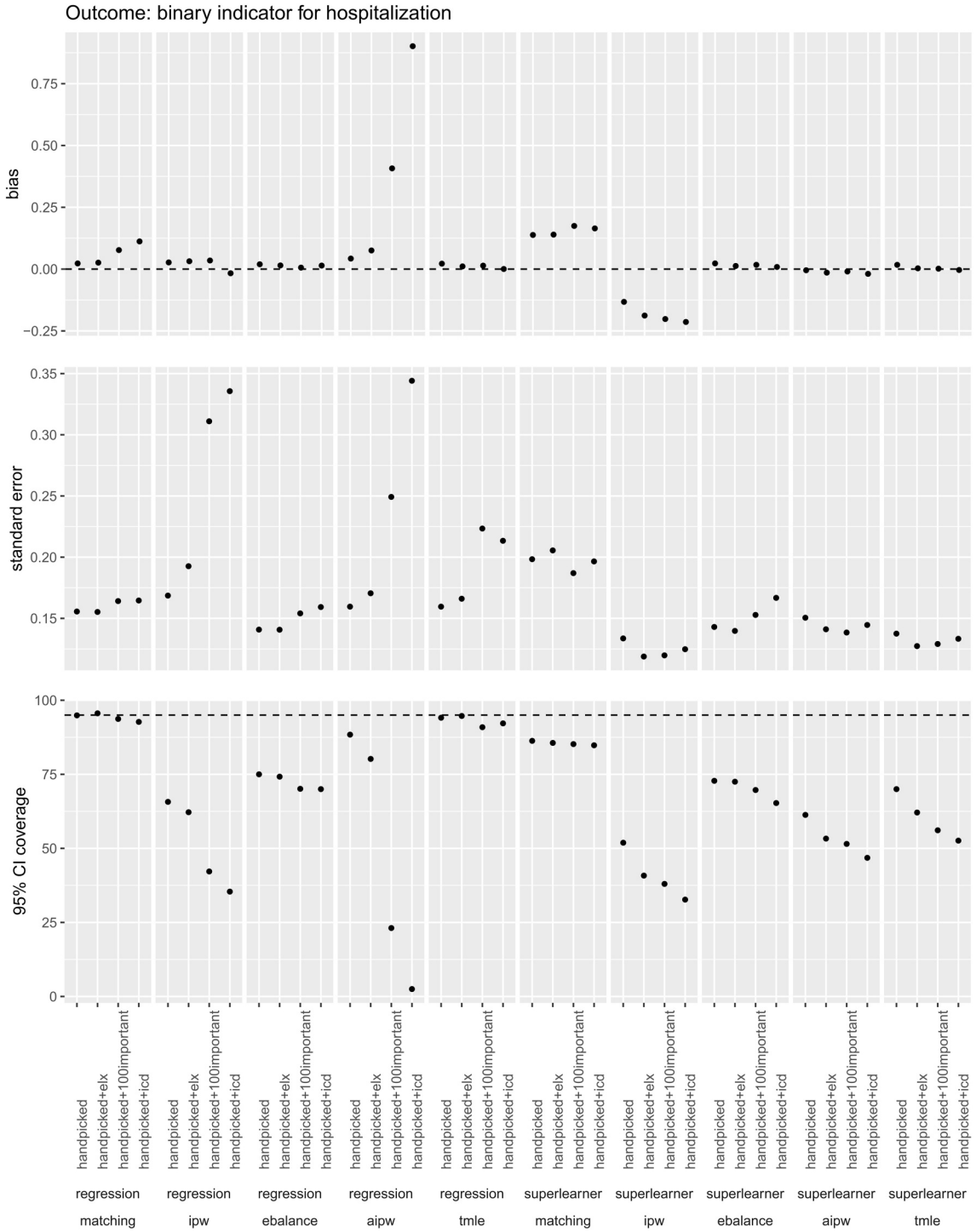

FIGURE 3 Performance measures for the odds ratios (OR) estimates of hospitalization (true treatment effect $\tau_{OR}^{hospitalization} = 0.95$) for each causal estimation method depending on the approach used for nuisance parameter estimation and the covariates used to control for confounding. The first panel shows the results for bias; the second panel the results for the empirical standard error; and the third panel the results for 95% CI coverage in percent. aipw, augmented inverse probability weighting; CI, confidence interval; ebalance, entropy balancing; handpicked, handpicked covariates only (32 covariates); handpicked+elx, handpicked covariates enhanced by the indicators for the Elixhauser comorbidity groups (63 covariates); handpicked+icd, handpicked covariates enhanced by the ICD codes among the 200 additional confounders used for simulation (204 covariates); handpicked+100important, handpicked covariates enhanced by the binary indicators for the 100 most important ATC, ICD and OPS codes based on gradient boosting (132 covariates); ipw, inverse probability of treatment weighting; matching, propensity score matching; tmle, targeted maximum likelihood estimation.

Regarding the choice of the covariate set used to control for confounding, bias tended to be larger for larger covariate sets for propensity score matching and IPW when combined with the superlearner, while performance did not vary much for the other methods. When using regression models to estimate the nuisance parameters, the SE was larger for larger covariate sets. However, when using the superlearner, there was no relationship between the size of the covariate set and the SE. Moreover, CI coverage tended to be lower for larger covariate sets across all scenarios, and was especially low for AIPW combined with regression for large covariate sets (23.1% and 2.5%).

Bias was smallest for entropy balancing, TMLE in combination with either regression models or the superlearner, IPW in combination with regression, and AIPW in combination with the superlearner. However, AIPW combined with regression models exhibited very large bias and SE. IPW combined with regression also showed large SE. On the other hand, IPW combined with the superlearner, had the smallest SE among all methods. Confidence interval coverage was below the 95% threshold for all methods except propensity score matching in combination with regression on the covariate set containing handpicked variables and Elixhauser groups.

5 | DISCUSSION

5.1 | Review of findings

We conducted a plasmode simulation study based on real-world administrative health care data, allowing us to evaluate the performance of five causal effect estimation methods under conditions closer to real-world conditions: propensity score matching (and subsequent DiD analysis), IPW (and subsequent DiD analysis), entropy balancing (and subsequent DiD analysis), AIPW and TMLE. We expanded upon previous plasmode simulation studies by focusing on count data (number of physician visits), a non-normally distributed continuous outcome (total health care costs), and a binary outcome (indicator for hospitalization). We varied the set of covariates used for analysis and the approach used to estimate the nuisance parameters.

Our simulation results indicate that using the superlearner instead of regression models to estimate the nuisance parameters leads to less biased estimates with smaller SE for AIPW and TMLE across all outcomes, and, in most cases, for IPW—with subsequent DiD analysis for estimating the ATT. The strong performance of doubly robust methods, especially TMLE, in terms of bias has been observed in several previous simulation studies (Bahamyirou et al., 2019; Meng & Huang, 2021; Naimi et al., 2021; Schuler & Rose, 2017; Zivich & Breskin, 2021) and aligns with the theoretical properties of doubly robust approaches (van der Laan & Rubin, 2006). Zivich and Breskin (2021) found that AIPW and TMLE, when combined with the superlearner, outperformed singly robust approaches using either the superlearner or regression models for nuisance parameter estimation. They also found that AIPW and TMLE combined with the superlearner outperformed singly robust approaches when combined with regression for nuisance parameter estimation (Zivich & Breskin, 2021). Similarly, Naimi et al. (2021) observed that using the superlearner in scenarios where nuisance parameter models are misspecified reduced bias for AIPW and TMLE. They further noted that while using the superlearner with singly robust estimators can result in biased estimators, performance improves when the superlearner is combined with doubly robust estimators (Naimi et al., 2021). This observation is both in line with theory (Naimi et al., 2021) and our observations. However, an important consideration arises when using the superlearner in conjunction with AIPW and TMLE: the coverage falls below the nominal level. This property, observed in our study, is consistent with the results of previous studies (Meng & Huang, 2021; Naimi et al., 2021; Zivich & Breskin, 2021).

For propensity score matching and IPW, previous studies have reported mixed results with regard to the choice between superlearner and regression for propensity score estimation, which was also observed in our study. Pirracchio et al. (2015) found that propensity score estimation using logistic regression outperformed the superlearner when the propensity score was either non-linear or non-additive. Conversely, if both conditions were met, the superlearner performed better. However, Alam et al. (2019) did not identify a consistently dominant approach and concluded that the superlearner was not superior to regression for estimating the propensity score when matching or IPW was subsequently performed. We found that CI coverage was similar regardless of the choice of approach for nuisance parameter estimation for propensity score matching, IPW or entropy balancing and subsequent DiD analysis in the case of ATT estimation.

When we compared the performance of the causal estimation methods with regard to the set of covariates used to control for confounding, a distinction arose depending on the choice of approach for nuisance parameter estimation. Using the superlearner tended to result in decreased bias when controlling for larger sets of covariates. In contrast,

when regression was used to estimate nuisance parameters, controlling for larger covariate sets did not consistently decrease bias. This observation aligns with other studies, which have found that covariate set size has no influence on performance measures (Amusa et al., 2022; Schuler & Rose, 2017) or have observed reduced bias and SE for larger covariate sets, suggesting that researchers should adjust for numerous variables (Karim et al., 2018; Pang et al., 2016). However, findings by Ripollone et al. (2020) for propensity score matching and Wyss et al. (2018) for propensity score matching, IPW and TMLW, suggest that bias increases with covariate set size. With regard to SE and CI coverage, we observed no consistent patterns related to covariate set size across outcomes.

When evaluating estimation methods, it is important to recognize that different performance measures reflect distinct properties of these methods and the relative importance of each measure can vary according to one's research objectives. For instance, researchers may prioritize minimizing bias or maximizing precision based on the goals of their study. It is also important to recognize that the methods we compare in this simulation study have different underlying causal assumptions. For example, methods like AIPW and TMLE presuppose the absence of unobserved confounding. In contrast, the DiD approach, while not requiring this assumption, does rely on the assumption of parallel trends. Despite these differences, we chose to compare these and the other methods due to their frequent application in practice and their endorsement in previous simulation studies. Our object in doing so was to offer practical guidance for applied researchers selecting an estimation strategy in scenarios similar to that presented in our study.

Thus, to conclude with a summary of our results across the employed performance measures, we ranked each method in terms of its performance with regard to bias, SE and CI coverage, and calculated their average ranks as a possible summary indicator. The ranking can be found in Appendix F. This approach showed that matching (with subsequent DiD analysis for ATT estimation) in combination with regression for nuisance parameter estimation performed well for all outcomes, according to the average rankings. Furthermore, entropy balancing combined with DiD analysis performed well for the outcome of total health care costs. Employing the superlearner for estimating nuisance parameters, in combination with treatment effect estimation using IPW or matching (followed by DiD analysis) or AIPW, performed well for estimating the treatment effect on the number of outpatient visits and total health care costs.

5.2 | Limitations

Several important limitations need to be considered when interpreting our results. First, while the plasmode framework captures the unknown and complex relationships among covariates, the outcome model in our study is still parametrically specified. The simulation process uses a relatively small number of covariates to generate outcomes, whereas real observed outcomes may be influenced by a much larger set of measured and unmeasured confounders, leading to increased complexity. However, it is worth noting that the associations with the included covariates may partially reflect the influence of additional covariates. This is a natural limitation of simulation studies, which arises from the challenge of generating a user-specified true parameter while preserving the structure of real-world data structure. As a result, our simulation design may favor approaches that are more closely related to the underlying data-generating process (Meng & Huang, 2021). Second, the strength of our plasmode simulation lies in its use of realistic data, but this also presents a challenge: the lack of knowledge about the underlying structure of these data. This also implies that we cannot be certain whether the model assumptions are met. For example, in our application of doubly robust methods, we rely on the relationships between covariates and treatment as they appear in the observed data, without alteration. This mirrors real-world applications of the method, where the accuracy of the treatment model is unknown and thus we cannot know whether the method relies on its double robustness property. As a consequence, our results primarily offer insights into performance in settings similar to those in our study, limiting generalizability to different contexts (Strobl & Leisch, 2022). Third, although we simulated outcomes for three different distributions and considered multiple levels of confounding, we were not able to test for other important features of the data-generating process, such as instrumental variables, varying treatment prevalence or missing data, nor did we use further data sources to generate the plasmode dataset. Moreover, the findings of simulation studies are inherently limited by the estimation strategies that are chosen. In our study, we did not, for example, consider different approaches for matching on the propensity score or further approaches to nuisance parameter estimation, such as extensive baselearner sets for the superlearner, nor did we perform hyperparameter tuning. Additionally, variations and extensions of the estimation strategies compared in our simulation—for example, staggered DiD for settings in which subjects are treated during different periods (Faghani Dermi & VanOmmeren, 2024; Wing et al., 2024)—may yield different results. However,

expanding the set of estimation strategies would have substantially increased computation time, whereas our objective was to provide applied researchers with initial guidance on choosing an estimation strategy in general. After choosing a broad strategy based on our findings, applied researchers can further refine their approach based on the findings of other simulation studies, such as comparing the performance of different propensity score methods (Franklin et al., 2017).

5.3 | Conclusion

In our plasmode simulation study using real-world administrative health care data, we evaluated and compared the performance of five methods for estimating causal treatment effects. We found that TMLE combined with the superlearner performed best in terms of bias and SE, but exhibited shortcomings in terms of CI coverage. When considering all performance measures and outcomes, the combination of matching and subsequent DiD analysis in conjunction with regression for nuisance parameter estimation performed best. For the individual outcomes and scenarios, other approaches showed similar performance to that of the matching and regression combination.

Based on our findings, we observed the following general trends that researchers working with administrative health care data may wish to consider when choosing strategies for estimating treatment effects in causal analysis:

1. When aiming to control for a large covariate set, consider using the superlearner to estimate nuisance parameters.
2. When employing the superlearner to estimate nuisance parameters, consider using doubly robust estimation approaches, such as AIPW and TMLE.
3. When faced with a small covariate set, consider using regression to estimate nuisance parameters.
4. When employing regression to estimate nuisance parameters, consider using singly robust estimation approaches, such as propensity score matching or IPW.

ACKNOWLEDGMENTS

This work was supported by the Innovation Fund of the German Federal Joint Committee, P.O. Box 12 06 06, Berlin, Germany [NVF2_2016-042].

Open Access funding enabled and organized by Projekt DEAL.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

The data are owned by the German statutory health insurers DAK, BARMER and AOK-RH. To fulfill the legal requirements to obtain the data, researchers must obtain permission for a specific research question from the German Federal (Social) Insurance Office.

ORCID

Vanessa Ress https://orcid.org/0000-0003-2989-8638
Eva-Maria Wild https://orcid.org/0000-0001-7243-5984

REFERENCES

Abadie, A. (2005). Semiparametric difference-in-differences estimators. *The Review of Economic Studies*, *72*, 1–19. https://doi.org/10.1111/0034-6527.00321

Alam, S., Moodie, E. E. M., & Stephens, D. A. (2019). Should a propensity score model be super? The utility of ensemble procedures for causal adjustment. *Statistics in Medicine*, *38*(9), 1690–1702. https://doi.org/10.1002/sim.8075

Amusa, L., Zewotir, T., & North, D. (2019). Examination of entropy balancing technique for estimating some standard measures of treatment effects: A simulation study. *Electronic Journal of Applied Statistical Analysis*, *12*, 491–507.

Amusa, L., Zewotir, T., & North, D. (2022). The impact of unmeasured confounding on causal inferencein observational studies: A plasmode simulation study of targeted maximum likelihood estimation. *Songklanakarin Journal of Science and Technology*, *44*(2), 474–480. https://doi.org/10.14456/SJST-PSU.2022.65

Aranda, R., Darden, M., & Rose, D. (2021). Measuring the impact of calorie labeling: The mechanisms behind changes in obesity. *Health Economics*, *30*(11), 2858–2878. https://doi.org/10.1002/hec.4415

Bahamyirou, A., Blais, L., Forget, A., & Schnitzer, M. E. (2019). Understanding and diagnosing the potential for bias when using machine learning methods with doubly robust causal estimators. *Statistical Methods in Medical Research*, *28*(6), 1637–1650. https://doi.org/10.1177/0962280218772065

Baumann, P. F. M., Schomaker, M., & Rossi, E. (2021). Estimating the effect of central bank independence on inflation using longitudinal targeted maximum likelihood estimation. *Journal of Causal Inference*, *9*(1), 109–146. https://doi.org/10.1515/jci-2020-0016

Bäuml, M., Marcus, J., & Siedler, T. (2023). Health effects of a ban on late-night alcohol sales. *Health Economics*, *32*(1), 65–89. https://doi.org/10.1002/hec.4610

Baxter, S., Johnson, M., Chambers, D., Sutton, A., Goyder, E., & Booth, A. (2018). The effects of integrated care: A systematic review of UK and international evidence. *BMC Health Services Research*, *18*(1), 350. https://doi.org/10.1186/s12913-018-3161-3

Bijwaard, G. E. (2022). Educational differences in mortality and hospitalisation for cardiovascular diseases. *Journal of Health Economics*, *81*, 102565. https://doi.org/10.1016/j.jhealeco.2021.102565

Caliendo, M., & Kopeinig, S. (2008). Some practical guidance for the implementation of propensity score matching. *Journal of Economic Surveys*, *22*(1), 31–72. https://doi.org/10.1111/j.1467-6419.2007.00527.x

Card, D., & Krueger, A. (1993). Minimum wages and employment: A case study of the fast food industry in New Jersey and Pennsylvania.

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., & Chen, K. (2015). Xgboost: Extreme gradient boosting. In *R package version 0.4-2 1* (pp. 1–4).

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, *21*(1), C1–C68. https://doi.org/10.1111/ectj.12097

Clarke, G. M., Conti, S., Wolters, A. T., & Steventon, A. (2019). Evaluating the impact of healthcare interventions using routine data. *BMJ*, *365*, l2239. https://doi.org/10.1136/bmj.l2239

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297. https://doi.org/10.1007/bf00994018

Crown, W. H. (2019). Real-world evidence, causal inference, and machine learning. *Value in Health : the journal of the International Society for Pharmacoeconomics and Outcomes Research*, *22*(5), 587–592. https://doi.org/10.1016/j.jval.2019.03.001

Elixhauser, A., Steiner, C., Harris, D. R., & Coffey, R. M. (1998). Comorbidity measures for use with administrative data. *Medical Care*, *36*(1), 8–27. https://doi.org/10.1097/00005650-199801000-00004

Faghani Dermi, H., & VanOmmeren, S. (2024). Staggered difference-in-differences estimation for antitrust analysis: A review of new literature and recommendations for practitioners. *SSRN Electronic Journal*. Steven. https://doi.org/10.2139/ssrn.4812196

Flawinne, X., Lefebvre, M., Perelman, S., Pestieau, P., & Schoenmaeckers, J. (2023). Nursing homes and mortality in Europe: Uncertain causality. *Health Economics*, *32*(1), 134–154. https://doi.org/10.1002/hec.4613

Franklin, J. M., Eddings, W., Austin, P. C., Stuart, E. A., & Schneeweiss, S. (2017). Comparing the performance of propensity score methods in healthcare database studies with rare outcomes. *Statistics in Medicine*, *36*, 1946–1963.

Franklin, J. M., Schneeweiss, S., Polinski, J. M., & Rassen, J. A. (2014). Plasmode simulation for the evaluation of pharmacoepidemiologic methods in complex healthcare databases. *Computational Statistics and Data Analysis*, *72*, 219–226. https://doi.org/10.1016/j.csda.2013.10.018

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, *33*, 1–22. https://doi.org/10.18637/jss.v033.i01

Friedman, J. H. (1991). Multivariate adaptive regression splines. *Annals of Statistics*, *19*(1). https://doi.org/10.1214/aos/1176347963

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, *29*(5), 1189–1232. https://doi.org/10.1214/aos/1013203451

Fu, R., Noguchi, H., Kawamura, A., Takahashi, H., & Tamiya, N. (2017). Spillover effect of Japanese long-term care insurance as an employment promotion policy for family caregivers. *Journal of Health Economics*, *56*, 103–112. https://doi.org/10.1016/j.jhealeco.2017.09.011

Glynn, A. N., & Quinn, K. M. (2010). An introduction to the augmented inverse propensity weighted estimator. *Political Analysis*, *18*(1), 36–56. https://doi.org/10.1093/pan/mpp036

Gruber, S., & van der Laan, M. J. (2012). Tmle: An R package for targeted maximum likelihood estimation. *Journal of Statistical Software*, *51*(13), 1–35. https://doi.org/10.18637/jss.v051.i13

Hainmueller, J. (2012). Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. Polit. *anal.*, *20*(1), 25–46. https://doi.org/10.1093/pan/mpr025

Heckman, J. J., & Vytlacil, E. (2001). Policy-relevant treatment effects. *The American Economic Review*, *91*(2), 107–111. https://doi.org/10.1257/aer.91.2.107

Hoffman, K. (2020). Become a superlearner! An illustrated guide to superlearning. https://www.khstats.com/blog/sl/superlearning. Accessed 15 November 2023.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, *81*(396), 945–960. https://doi.org/10.2307/2289064

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

Husereau, D., Drummond, M., Augustovski, F., Bekker-Grob, E. de, Briggs, A. H., Carswell, C., Caulley, L., Chaiyakunapruk, N., Greenberg, D., Loder, E., Mauskopf, J., Mullins, C. D., Petrou, S., Pwu, R.-F., & Staniszewska, S. (2022). Consolidated health economic evaluation reporting standards 2022 (CHEERS 2022) statement: Updated reporting guidance for health economic evaluations. *International Journal of Technology Assessment in Health Care*, *38*, e13. https://doi.org/10.1016/j.hpopen.2021.100063

Hwang, J.-S., Hu, T.-H., Lee, L. J.-H., & Wang, J.-D. (2017). Estimating lifetime medical costs from censored claims data. *Health Economics*, *26*(12), e332–e344. https://doi.org/10.1002/hec.3512

Jones, A. M., Lomas, J., & Rice, N. (2015). Healthcare cost regressions: Going beyond the mean to estimate the full distribution. *Health Economics*, *24*(9), 1192–1212. https://doi.org/10.1002/hec.3178

Karatzoglou, A., Meyer, D., & Hornik, K. (2006). Support vector machines. *R. J. Stat. Soft.*, *15*(9), 1–28. https://doi.org/10.18637/jss.v015.i09

Karim, M. E., Pang, M., & Platt, R. W. (2018). Can we train machine learning methods to outperform the high-dimensional propensity score algorithm? *Epidemiology*, *29*(2), 191–198. https://doi.org/10.1097/ede.0000000000000787

Kreif, N., Grieve, R., Hangartner, D., Turner, A. J., Nikolova, S., & Sutton, M. (2016). Examination of the synthetic control method for evaluating health policies with multiple treated units. *Health Economics*, *25*(12), 1514–1528. https://doi.org/10.1002/hec.3258

Kreif, N., Grieve, R., Radice, R., & Sekhon, J. S. (2013). Regression-adjusted matching and double-robust methods for estimating average treatment effects in health economic evaluation. *Health Services & Outcomes Research Methodology*, *13*(2–4), 174–202. https://doi.org/10.1007/s10742-013-0109-2

Kreif, N., Gruber, S., Radice, R., Grieve, R., & Sekhon, J. S. (2016). Evaluating treatment effectiveness under model misspecification: A comparison of targeted maximum likelihood estimation with bias-corrected matching. *Statistical Methods in Medical Research*, *25*(5), 2315–2336. https://doi.org/10.1177/0962280214521341

Loux, T. M., Drake, C., & Smith-Gagen, J. (2017). A comparison of marginal odds ratio estimators. *Statistical Methods in Medical Research*, *26*(1), 155–175. https://doi.org/10.1177/0962280214541995

Luyten, J., Verbeke, E., & Schokkaert, E. (2022). To be or not to be: Future lives in economic evaluation. *Health Economics*, *31*(1), 258–265. https://doi.org/10.1002/hec.4454

Marcus, J. (2013). The effect of unemployment on the mental health of spouses - evidence from plant closures in Germany. *Journal of Health Economics*, *32*(3), 546–558. https://doi.org/10.1016/j.jhealeco.2013.02.004

Meng, X., & Huang, J. (2021). Doubly robust, machine learning effect estimation in real-world clinical sciences: A practical evaluation of performance in molecular epidemiology cohort settings. *arXiv e-prints*. arXiv: 2105.13148.

Morgan, S. L., & Todd, J. J. (2008). Diagnostic routine for the detection of consequential heterogeneity of causal effects. *Sociological Methodology*, *38*(1), 231–282. https://doi.org/10.1111/j.1467-9531.2008.00204.x

Morris, T. P., White, I. R., & Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, *38*(11), 2074–2102. https://doi.org/10.1002/sim.8086

Nagelkerke, N. J. D. (1991). A note on a general definition of the coefficient of determination. *Biometrika*, *78*(3), 691–692. https://doi.org/10.2307/2337038

Naimi, A. I., Mishler, A. E., & Kennedy, E. H. (2021). Challenges in obtaining valid causal effect estimates with machine learning algorithms. *American Journal of Epidemiology*, *192*(9), 1536–1544. https://doi.org/10.1093/aje/kwab201

Nasseh, K., Vujicic, M., & Glick, M. (2017). The relationship between periodontal interventions and healthcare costs and utilization. Evidence from an integrated dental, medical, and pharmacy commercial claims database. *Health Economics*, *26*(4), 519–527. https://doi.org/10.1002/hec.3316

Norton, E. C., Dowd, B. E., Garrido, M. M., & Maciejewski, M. L. (2024). Requiem for odds ratios. *Health Services Research*, *59*(4). https://doi.org/10.1111/1475-6773.14337

O'Neill, S., Kreif, N., Grieve, R., Sutton, M., & Sekhon, J. S. (2016). Estimating causal effects: Considering three alternatives to difference-in-differences estimation. *Health Services & Outcomes Research Methodology*, *16*, 1–21.

Pang, M., Schuster, T., Filion, K. B., Schnitzer, M. E., Eberg, M., & Platt, R. W. (2016). Effect estimation in point-exposure studies with binary outcomes and high-dimensional covariate data - a comparison of targeted maximum likelihood estimation and inverse probability of treatment weighting. *International Journal of Biostatistics*, *12*(2). https://doi.org/10.1515/ijb-2015-0034

Pearl, J. (2009). Causal inference in statistics: An overview. *Statistics Surveys*, *3*(none). https://doi.org/10.1214/09-ss057

Perperoglou, A., Sauerbrei, W., Abrahamowicz, M., & Schmid, M. (2019). A review of spline function procedures in R. *BMC Medical Research Methodology*, *19*(1), 46. https://doi.org/10.1186/s12874-019-0666-3

Persson, E., & Waernbaum, I. (2013). Estimating a marginal causal odds ratio in a case-control design: Analyzing the effect of low birth weight on the risk of type 1 diabetes mellitus. *Statistics in Medicine*, *32*(14), 2500–2512. https://doi.org/10.1002/sim.5826

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*(2), 108–119. https://doi.org/10.1093/aje/kwu253

Pirracchio, R., Resche-Rigon, M., & Chevret, S. (2012). Evaluation of the propensity score methods for estimating marginal odds ratios in case of small sample size. *BMC Medical Research Methodology*, *12*(1), 70. https://doi.org/10.1186/1471-2288-12-70

Polley, E., LeDell, E., Kennedy, C., & van der Laan, M. (2021). Super learner: Super learner prediction. https://cran.r-project.org/web/packages/SuperLearner/index.html

Polley, E., & van der Laan, M. (2010). Super learner in prediction. *U.C. Berkeley Division of Biostatistics Working Paper Series*.

Radice, R., Ramsahai, R., Grieve, R., Kreif, N., Sadique, Z., & Sekhon, J. S. (2012). Evaluating treatment effectiveness in patient subgroups: A comparison of propensity score methods with an automated matching approach. *International Journal of Biostatistics*, *8*(1). https://doi.org/10.1515/1557-4679.1382

Ress, V., & Wild, E.-M. (2023). The impact of integrated care on health care utilization and costs in a socially deprived urban area in Germany: A difference-in-differences approach within an event-study framework. *Health Economics*, *33*(2), 229–247. https://doi.org/10.1002/hec.4771

Ripollone, J. E., Huybrechts, K. F., Rothman, K. J., Ferguson, R. E., & Franklin, J. M. (2020). Evaluating the utility of coarsened exact matching for pharmacoepidemiology using real and simulated claims data. *American Journal of Epidemiology*, *189*(6), 613–622. https://doi.org/10.1093/aje/kwz268

Robins, J. M., Hernán, M. A., & Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology*, *11*(5), 550–560. https://doi.org/10.1097/00001648-200009000-00011

Robins, J. M., Rotnitzky, A., & Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, *89*(427), 846–866. https://doi.org/10.2307/2290910

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. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66*(5), 688–701. https://doi.org/10.1037/h0037350

Sarma, S., Mehta, N., Devlin, R. A., Kpelitse, K. A., & Li, L. (2018). Family physician remuneration schemes and specialist referrals: Quasi-experimental evidence from Ontario, Canada. *Health Economics*, *27*(10), 1533–1549. https://doi.org/10.1002/hec.3783

Schneeweiss, S., Rassen, J. A., Glynn, R. J., Avorn, J., Mogun, H., & Brookhart, M. A. (2009). High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. *Epidemiology*, *20*(4), 512–522. https://doi.org/10.1097/ede.0b013e3181a663cc

Schreyögg, J., Stargardt, T., & Tiemann, O. (2011). Costs and quality of hospitals in different health care systems: A multi-level approach with propensity score matching. *Health Economics*, *20*(1), 85–100. https://doi.org/10.1002/hec.1568

Schuler, M. S., & Rose, S. (2017). Targeted maximum likelihood estimation for causal inference in observational studies. *American Journal of Epidemiology*, *185*(1), 65–73. https://doi.org/10.1093/aje/kww165

Sekhon, J. S., & Grieve, R. D. (2012). A matching method for improving covariate balance in cost-effectiveness analyses. *Health Economics*, *21*(6), 695–714. https://doi.org/10.1002/hec.1748

Somé, N. H., Devlin, R. A., Mehta, N., Zaric, G., Li, L., Shariff, S., Belhadji, B., Thind, A., Garg, A., & Sarma, S. (2019). Production of physician services under fee-for-service and blended fee-for-service: Evidence from Ontario, Canada. *Health Economics*, *28*(12), 1418–1434. https://doi.org/10.1002/hec.3951

Strobl, C., & Leisch, F. (2022). Against the "one method fits all data sets" philosophy for comparison studies in methodological research. *Biometrical Journal*, *66*(1). https://doi.org/10.1002/bimj.202200104

Strumpf, E., Ammi, M., Diop, M., Fiset-Laniel, J., & Tousignant, P. (2017). The impact of team-based primary care on health care services utilization and costs: Quebec's family medicine groups. *Journal of Health Economics*, *55*, 76–94. https://doi.org/10.1016/j.jhealeco.2017.06.009

Stuart, E. A., King, G., Imai, K., & Ho, D. (2011). MatchIt: Nonparametric preprocessing for parametric causal inference. *Journal of Statistical Software*, *42*(8). https://doi.org/10.18637/jss.v042.i08

Urwin, S., Lau, Y.-S., Grande, G., & Sutton, M. (2023). Informal caregiving, time use and experienced wellbeing. *Health Economics*, *32*(2), 356–374. https://doi.org/10.1002/hec.4624

Urwin, S., Mason, T., & Whittaker, W. (2021). Do different means of recording sexual orientation affect its relationship with health and wellbeing? *Health Economics*, *30*(12), 3106–3122. https://doi.org/10.1002/hec.4422

van der Laan, M. J., Polley, E. C., & Hubbard, A. E. (2007). Super learner. *Statistical Applications in Genetics and Molecular Biology*, *6*(1), Article25. https://doi.org/10.2202/1544-6115.1309

van der Laan, M. J., & Rose, S. (2011). *Targeted learning: Causal inference for observational and experimental data*. Springer.

van der Laan, M. J., & Rubin, D. (2006). Targeted maximum likelihood learning. *International Journal of Biostatistics*, *2*(1). https://doi.org/10.2202/1557-4679.1043

Varga, A. N., Guevara Morel, A. E., Lokkerbol, J., van Dongen, J. M., van Tulder, M. W., & Bosmans, J. E. (2023). Dealing with confounding in observational studies: A scoping review of methods evaluated in simulation studies with single-point exposure. *Statistics in Medicine*, *42*(4), 487–516. https://doi.org/10.1002/sim.9628

Vaughan, L. K., Divers, J., Padilla, M., Redden, D. T., Tiwari, H. K., Pomp, D., & Allison, D. B. (2009). The use of plasmodes as a supplement to simulations: A simple example evaluating individual admixture estimation methodologies. *Computational Statistics and Data Analysis*, *53*(5), 1755–1766. https://doi.org/10.1016/j.csda.2008.02.032

Wang, A., Nianogo, R. A., & Arah, O. A. (2017). G-computation of average treatment effects on the treated and the untreated. *BMC Medical Research Methodology*, *17*(1), 3. https://doi.org/10.1186/s12874-016-0282-4

Wing, C., Freedman, S. M., & Hollingsworth, A. (2024). *Stacked difference-in-differences*. National Bureau of Economic Research.

Wright, M. N., & Ziegler, A. (2017). Ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software*, *77*(1). https://doi.org/10.18637/jss.v077.i01

Wyss, R., Schneeweiss, S., van der Laan, M., Lendle, S. D., Ju, C., & Franklin, J. M. (2018). Using super learner prediction modeling to improve high-dimensional propensity score estimation. *Epidemiology*, *29*(1), 96–106. https://doi.org/10.1097/ede.0000000000000762

Zhong, Y., Kennedy, E. H., Bodnar, L. M., & Naimi, A. I. (2021). Aipw: An R package for augmented inverse probability-weighted estimation of average causal effects. *American Journal of Epidemiology*, *190*(12), 2690–2699. https://doi.org/10.1093/aje/kwab207

Zivich, P. N., & Breskin, A. (2021). Machine learning for causal inference: On the use of cross-fit estimators. *Epidemiology*, *32*(3), 393–401. https://doi.org/10.1097/ede.0000000000001332

SUPPORTING INFORMATION

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

How to cite this article: Ress, V., & Wild, E.-M. (2024). Comparing methods for estimating causal treatment effects of administrative health data: A plasmode simulation study. *Health Economics*, *33*(12), 2757–2777. https://doi.org/10.1002/hec.4891