Received: 15 July 2021 | Revised: 13 May 2022 | Accepted: 25 May 2022

DOI: 10.1002/pst.2244

MAIN PAPER

WILEY

Estimators for handling COVID-19-related intercurrent events with a hypothetical strategy

Florian Lasch[1,2] | **Lorenzo Guizzaro**[1,3]

[1]European Medicines Agency, Amsterdam, The Netherlands

[2]Hannover Medical School, Hannover, Germany

[3]Medical Statistics Unit, Università della Campania "Luigi Vanvitelli", Napoli, Italy

Correspondence
Lorenzo Guizzaro, European Medicines Agency, Amsterdam, The Netherlands.
Email: lorenzo.guizzaro@ema.europa.eu

Abstract

The COVID-19 pandemic has affected clinical trials across disease areas, raising the questions how interpretable results can be obtained from impacted studies. Applying the estimands framework, analyses may seek to estimate the treatment effect in the hypothetical absence of such impact. However, no established estimators exist. This simulation study, based on an ongoing clinical trial in patients with Tourette syndrome, compares the performance of candidate estimators for estimands including either a continuous or binary variable and applying a hypothetical strategy for COVID-19-related intercurrent events (IE). The performance is investigated in a wide range of scenarios, under the null and the alternative hypotheses, including different modeling assumptions for the effect of the IE and proportions of affected patients ranging from 10% to 80%. Bias and type I error inflation were minimal or absent for most estimators under most scenarios, with only multiple imputation- and weighting-based methods displaying a type I error inflation in some scenarios. Of more concern, all methods that discarded post-IE data displayed a sharp decrease of power proportional to the proportion of affected patients, corresponding to both a reduced precision of estimation and larger confidence intervals. The simulation study shows that de-mediation via g-estimation is a promising approach. Besides showing the best performance in our simulation study, these approaches allow to estimate the effect of the IE on the outcome and cross-compare between different studies affected by similar IEs. Importantly, the results can be extrapolated to IEs not related to COVID-19 that follow a similar causal structure.

1 | INTRODUCTION

The COVID-19 pandemic has affected clinical trials across disease areas,[1,2] raising questions on whether and how interpretable results can be obtained by ongoing studies.[3] Disruptions to trials caused by the pandemic can be considered "Intercurrent Events" (IE)[4] as defined in the Estimand framework as "Events occurring after treatment initiation that affect either the interpretation or the existence of the measurements associated with the clinical question of interest".[5]

Florian Lasch and Lorenzo Guizzaro contributed equally; order was assigned by tossing a coin.

The views expressed in this article are the personal views of the authors and may not be understood or quoted as being made on behalf of or reflecting the position of the agencies or organisations with which the authors are affiliated.

In the estimand framework, integrating with the other attributes (*treatment*, *population*, *variable of interest* and *population-level summary*), the handling of the IE defines the estimand of interest in a clinical trial and is explicitly acknowledged to reflect and shape the question that the study aims to address.[6–8]

Where the objective of a trial is to estimate the effect of a treatment in the absence of some of the impacts of the pandemic, it has been proposed that the analysis should aim to determine the effect if the COVID-related IE had not occurred.[9,10] However, the existence of reliable estimators is a precondition to implement this (hypothetical) strategy.[9] Importantly, the suitability and performance of estimators depends on the clinical context and positioning of the IE in the causal structure of the trial.

To investigate candidate estimators for an estimand that includes the hypothetical strategies for a COVID-related IE, in this article we present a simulation study, using the CANNA-TICS trial (ClinicalTrials.gov identifier NCT03087201) as an example. This multicenter, randomized, double-blind, placebo controlled, and parallel-group trial aims to demonstrate that the treatment with the cannabis extract Nabiximols is superior to placebo in reducing tics and comorbidities in adult patients with chronic tic disorders (CTD) and Gilles de la Tourette syndrome (TS). The primary endpoint is based on the Total Tic Score (TTS) of the Yale Global Tic Severity Scale (YGTSS),[11] a semi-structured interview assessing the severity of motor and phonic tics (ranges 0–50). As the primary endpoint, a dichotomization of the continuous relative change from baseline compared to 13 weeks after treatment initiation is used to measure the proportion of patients with a reduction in the YGTSS-TTS of at least 30%. More details can be found in the study protocol publication[12] and the previously published investigation of the impact of the COVID-19 pandemic on the power of the trial if a treatment-policy strategy is employed for handling the COVID-related IEs.[13] Lasch et al. showed that depending on the proportion of affected patients, substantial power losses are possible, potentially making sample size increases necessary to retain sufficient power. Additionally, the simulation study showed that by adjusting for the occurrence of the COVID-19-related IE, the power loss could be diminished to different degrees in most of the investigated scenarios.

For the CANNA-TICS trial, the following elements are examples for COVID-19-related IEs:

1. The implementation of social distancing measures such as social distancing or complete lockdowns can affect mental health and increase anxiety. This has been shown to increase the severity of tics of patients with CTD and TS.[14] In a clinical trial, this would imply a potential increase in the YGTSS-TTS due to the COVID-19 pandemic in patients recruited before the start of the COVID-19 pandemic and assessed at 13 weeks during the pandemic.
2. The ascertainment method for the YGTSS-TSS has been changed from in-person to remote due to the change in access rules by participating hospitals. This change in the assessment method from baseline to primary endpoint measurement might influence the interpretability of the primary endpoint, for example due to the different psychological state of the subjects undergoing a visit in a clinic versus connecting from their own house, leading to a different assessment of their own symptoms.[15] While no specific evidence comparing in-person versus remote assessment of the YGTSS exists, conflicting evidence can be found from other neurological and psychiatric conditions. On the one hand, there is some support to the interchangeability of in person and remote assessment for some of the most widely used scales for depression[15] and remote assessment of the YGTSS is used as a pre-defined standard procedure in the EMTICS trial.[16] On the other hand, there is some evidence suggesting that the physical distance from the clinician can decrease anxiety,[17] or that patients might be more or less prone to admit severity of symptoms depending on the modality of assessment.[18]

Note that addressing the IE with a hypothetical strategy does not directly translate to estimating the effect of the treatment in a "post-COVID-19 world" or "in line with the initial trial objective."[19] While handling these two IEs with a hypothetical strategy is in principle compatible with these scenarios, additional IEs like COVID-19 infections would need to be considered in addition to fully reflect the scenario of interest. Additionally, different stakeholders (pharmaceutical industry, regulators, Health Technology Assessment bodies, prescribers, and patients) might be interested in slightly different questions, which would translate to different estimands of interest. The above-described IEs are considered relevant examples in the investigated disease and trial setting, but they do not aim to represent all relevant IE that are related to Covid-19. Consequently, the focus of our article is the comparison of the performance of candidate estimators for an estimand that handles (some) Covid-19-related IE with a hypothetical strategy. The findings are applicable for different IEs as long as data-generating model and the causal structure are the same as one of those simulated.

The impact of the IE on the trial and on the performance of different estimators depends not only on the characteristics of the IE per se, but also on the proportion of patients affected and the absolute number of patients. For example, the operating characteristics of some estimators might be sensitive to the absolute number of affected and unaffected

patients available for modeling. While the proportion of affected patients cannot be known at the planning stage, it can be determined at the blind review stage before the final analysis. The performance of the estimators is also influenced by factors that are usually not known a priori such as the magnitude of the effect of the IE and the relationship between the magnitude of this effect and the original value of the outcome variable.

In the following sections, first we briefly introduce the CANNA-TICS trial as a motivating example and outline the simulation model and the simulated scenarios—implementing different modeling assumptions for the real relationship between IE and outcome under both the presence and absence of a treatment effect. In this article, we assume that only the IE influences the outcome of the patient and not vice versa. This assumption would not hold in case the occurrence of the IE was confounded by the patient outcome: for example, if patients could choose between assessment methods, patients with a high YGTSS-TTS might be less/more likely to choose a clinical assessment. Subsequently, we describe the candidate estimators for both the responder criterion, dichotomizing the relative change in YGTSS-TTS, and the continuous relative change in YGTSS-TTS.

Finally, we present the performance results for the simulated scenarios, for both the continuous and the dichotomized endpoint and discuss the consequences of our findings.

2 | OBJECTIVES

This article discusses estimators for an estimand applying a hypothetical strategy for COVID-19-related IE for both a dichotomized and a continuous endpoint in trials with a small sample size. We aim to compare different estimators for the continuous analysis of the underlying scale and the binary endpoint regarding bias, mean squared error, type I error control, and power.

3 | METHODS

To compare the performance of different estimators for the estimand applying a hypothetical strategy for COVID-19-related IEs, we designed a simulation study motivated by the ongoing CANNA-TICS trial. For better identification of the impact of the IEs, we simplified some aspects of the trial for the simulation study as outlined below.

4 | DATA-GENERATING MODEL

As the CANNA-TICS trial and the modeling of the IEs have already been extensively described in the context of the estimation of a treatment-policy effect,[13] we will provide a concise description here.

In the CANNA-TICS trial patients were randomized in a 2:1 ratio to Nabiximols or placebo. The assignment is represented here as a random variable Z taking values of 1 for patients assigned to Nabiximols and 0 for patients assigned to placebo. The outcome, the YGTSS-TTS, is measured twice, at baseline (Y_0) and after 13 weeks (Y_1^*). Some of the patients (depending on time of recruitment and region) will be affected by the pandemic ($P = 1$), which is the intercurrent event of interest. In this case, their outcome value at week 13 (Y_1^*) will be a function of their (unmeasured) underlying severity value Y_1 (i.e., of the severity value that would have been observed in absence of the IE) and of the effect of the IE. For all other patients, $Y_1^* = Y_1$. The YGTSS-TTS at week 13 (Y_1) is also affected in all scenarios (Figure 1 and

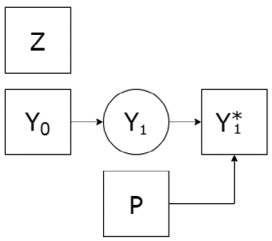

FIGURE 1 Data-generating mechanism in absence of an effect of treatment on the outcome. Observed variables are represented as squares, unobserved variables as circles

Figure 2) by the baseline value (Y_0), and only in the scenarios simulated under the alternative hypothesis (Figure 2) by the treatment assigned (Z). We are here assuming that the study was fully recruited (or recruitment was halted) at the time the first measures were implemented. Hence, no patients are affected by the pandemic at baseline. In case recruitment was still ongoing at the onset of the pandemic, more granular definitions of the IEs might be needed. For the change in assessment method, the IE "change in assessment method" could be split in two separate IEs (i) "change in assessment method from clinical to remote" and (ii) "change in assessment method from remote to clinical." Both IEs could be handled with a hypothetical strategy, but estimation would be more complex as compared to the simulation study as two IEs would need to be considered simultaneously.

For a fixed sample size $n = 75$ and fixed proportions of patients affected by the IE p_P, the number of patients (not) affected was calculated as follows:

$$\mathrm{n_{unaffected}} = \mathrm{round}\left(\mathrm{n} * (1 - \mathrm{p_P})\right)$$

and

$$n_{affected} = n - n_{unaffected}.$$

The variable $P_i = \begin{cases} 0 & i \in \{1, \ldots, n_{unaffected}\} \\ 1 & i \in \{n_{unaffected} + 1, \ldots n\} \end{cases}$.

denotes whether a patient is affected by the IE in question. Note that in the causal DAGs assumed for this simulation study, the occurrence of the IE is independent on the outcome of the patient. The treatment allocation $Z_i \tilde{Ber}(p_{treat})$ for each patient $i \in \{1, \ldots, n\}$ was simulated as Bernoulli distributed variable with fixed probability $ptreat = \frac{2}{3}$ to receive Nabiximols. For each patient, a baseline value $Y_{0,i} \sim N(25, 6.5)_{[14,50]}$ from a truncated normal distribution was simulated,[22] where the lower boundary of 14 reflects a simplified inclusion criterion and the upper boundary of 50 reflects the maximal value of the YGTSS-TTS. We simplify the assumptions of the Cannatics trial by assuming equal variances of the relative change in both treatment groups. The relative change from baseline to week 13 was assumed to be approximately normal distributed with

$$\mathrm{change_i}\Big|_{\mathrm{treat_i}=0} = \frac{\mathrm{Y_{1,i}} - \mathrm{Y_{0,i}}}{\mathrm{Y_{0,i}}}\Big|_{\mathrm{Z_i}=0} \sim \mathrm{N}(-0.025, 0.12)$$

for placebo patients. For patients receiving Nabiximols, we chose two different distributions for investigating candidate estimators for the binary responder criterion and candidate estimators for the relative change as a continuous endpoint. This ensures that—in both the binary and the continuous cases—an analysis based on the unaffected values has a nominal power of 90% and thus is sensitive to detect differences in power between candidate estimators. We have chosen the relative changes in order to achieve such power. Consequently, for a comparison of estimators for the binary responder criterion, we modeled the relative change via

$$\mathrm{change_i}\Big|_{\mathrm{treat_i}=1} = \frac{\mathrm{Y_{1,i}} - \mathrm{Y_{0,i}}}{\mathrm{Y_{0,i}}}\Big|_{\mathrm{Z_i}=1} \sim \mathrm{N}(-0.234, 0.12)$$

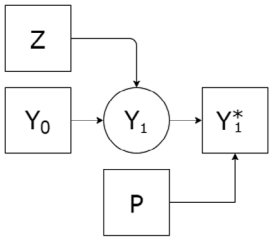

FIGURE 2 Data-generating mechanism in presence of an effect of treatment on the outcome. Observed variables are represented as squares, unobserved variables as circles

and for a comparison of estimators for the continuous endpoint, we simulated the relative change via

$$change_i \Big|_{treat_i=1} = \frac{Y_{1,i} - Y_{0,i}}{Y_{0,i}} \Big|_{Z_i=1} \sim N(-0.122, 0.12).$$

Next, the underlying, potentially unobserved YGTSS-TTS scores after 13 weeks, unaffected by IEs, were calculated as follows:

$$Y_{1,i} = Y_{0,i} + change_i^* Y_{0,i} = Y_{0,i} + \frac{Y_{1,i} - Y_{0,i}^*}{Y_{0,i}} Y_{0,i}.$$

For modeling the impact of the intercurrent event on the score at week 13, we implemented both (i) an additive model and (ii) a multiplicative model influencing the score that would have been observed in absence of the IE.

Under both the additive and multiplicative models, we assume that the IE affects the patients' outcomes heterogeneously.

For the additive scenarios, a random term from a normal distribution was simulated for each patient, $C_i \sim N(2,1)$. Using this term, the observed scores were calculated as

$$Y_{1,i}^* = Y_{1,i} + C_i * P_i,$$

making sure that for unaffected patients, the observed score equals the underlying score. This corresponds to an average increase of the YGTSS-TTS by the occurrence of the IE.

For the multiplicative scenarios, a random factor from a normal distribution truncated at 0 and 2 was simulated for each patient, $C_i \sim N(1.5, 0.1)$. Using this factor, the observed scores were calculated as $Y_{1,i}^* = Y_{1,i} * C_i^{P_i}$, making sure that for unaffected patients, the observed score equals the underlying score. On average, this corresponds to an increase of the YGTSS-TTS by the occurrence of the IE.

Under both models, the impact of the pandemic, C_i, does not only increase the YGTSS-TTS score $Y_{1,i}^*$, but also increase the variance of $Y_{1,i}^*$, which in turn influences the expectation of the dichotomized responder criterion. Based on the observed value$Y_{1,i}^*$, the observed relative change is calculated as $change_i^* = \frac{Y_{1,i}^* - Y_{0,i}}{Y_{0,i}}$, and the responder criterion is calculated as follows:

$$\text{responder}_i^* = \begin{cases} 0 & \text{change}_i^* > -0.3 \\ 1 & \text{change}_i^* \leq -0.3 \end{cases}$$

5 | SIMULATION SCENARIOS

The performance of the candidate estimators is investigated under the null hypothesis and under the alternative hypothesis. Under both hypotheses, we investigate both and additive impact of the IE and a multiplicative impact of the IE on the YGTSS-TTS after 13 weeks of treatment, Y_1. As described in the previous section, under the alternative hypothesis, different distributions for $change_i|_{treat_i=1}$ were used depending on whether estimators using the binary responder criterion or the continuous variable *relative change* were compared. To investigate the impact of the proportion of patients affected by the pandemic on the performance of the candidate estimators, for each of the above scenarios, we investigate proportions $p_P \in \{10\%20\%, ..., 80\%\}$ of affected patients. As the methods reported are proposed to handle the intercurrent event, scenarios with 0% affected patients were not simulated. In total, this results in $2*8=16$ simulation scenarios under the null hypothesis and $2*2*8=32$ simulation scenarios under the alternative hypothesis, each with 10,000 simulation runs (see Table 1). The number of simulation runs was chosen to achieve a standard error for the estimated empirical power/empirical type I error of at most 0.005, resulting in a half-width of the respective 95% confidence interval of at most 0.01, which we considered resulting in negligible standard errors.

TABLE 1 Overview of simulation scenarios

Hypothesis	Impact of IE	Analysis focus	Distribution of $\text{change}_i \mid_{\text{treat}_i=1}$	Proportion of affected patients
Null	Additive $C_i \sim N(2,1)$	binary + continuous	$N(-0.025, 0.12)$	$p_P \in \{10\%20\%, ..., 80\%\}$
Null	Multiplicative $C_i \sim N(1.5, 0.1)$	binary + continuous	$N(-0.025, 0.12)$	$p_P \in \{10\%20\%, ..., 80\%\}$
Alternative	Additive $C_i \sim N(2,1)$	binary	$N(-0.234, 0.12)$	$p_P \in \{10\%20\%, ..., 80\%\}$
Alternative	Additive $C_i \sim N(2,1)$	continuous	$N(-0.122, 0.12)$	$p_P \in \{10\%20\%, ..., 80\%\}$
Alternative	Multiplicative $C_i \sim N(1.5, 0.1)$	binary	$N(-0.234, 0.12)$	$p_P \in \{10\%20\%, ..., 80\%\}$
Alternative	Multiplicative $C_i \sim N(1.5, 0.1)$	continuous	$N(-0.122, 0.12)$	$p_P \in \{10\%20\%, ..., 80\%\}$

6 | ESTIMANDS AND ESTIMATORS

6.1 | Formal definition of the Estimands

The estimands of interest for (a) the continuous and (b) the binary endpoint are defined as follows, where IE, the corresponding strategies, and the estimand attributes of treatment condition and population are the same for both (and they only differ for variable and summary measure):

Treatment conditions: Investigational treatment with cannabis extracts Nabiximols for a 4-week titration period followed by 9 weeks of stable treatment at the individually selected dose versus matching placebo;

Population: Patients with Chronic Tic Disorder or Gilles de la Tourette syndrome;

Variable:

a. Relative change of YGTSS-TTS from baseline to 13 weeks ([13 weeks—baseline] / baseline);
b. Responder status (responder/non-responder), where a responder is defined as having a reduction of YGTSS-TTS from baseline to 13 weeks of at least 30% (*relative change* $\leq -30\%$).

6.2 | Summary measure

a. Difference in mean relative change of YGTSS-TTS from baseline to 13 weeks;
b. Difference in responder proportions.

IE and strategies:

1. Implementation of social distancing measures—hypothetical strategy (as if the IE—that is, the implementation of social distancing measures—had not taken place)
2. Change in assessment method for the YGTSS-TTS—hypothetical strategy (as if the IE—that is, the implementation of social distancing measures—had not taken place)

Note that the estimands attribute "population" defines the target population and does not equal the analysis set used for the estimation. In our example, we did not restrict the target population regarding COVID-19 (e.g., by targeting only Patients with Chronic Tic Disorder or Gilles de la Tourette syndrome that are not affected by social distancing measures). On the estimator level however, we investigate candidate estimators that use as analysis set either (i) all trial patients irrespective of the occurrence of COVID-19-related IEs or (ii) only patients that finished study procedures prior to the outbreak of the pandemic. Importantly, even if the target population had been restricted to patients that are not

affected by social distancing measures, on the estimator level it would still be an option to use all trial patients in the analysis.

In formal notation, the estimand in our simulation study can be defined as follows. Although the IE is not a mediator in the causal structure assumed for this simulation study, we can use the counterfactual concept from mediation analysis for a formal definition described by Richiardi et al.[20] The estimand is the effect of Z on $change_1^*$ in the hypothetical scenario that no COVID-19-related IE would have occurred, a controlled direct effect that can be formalized as follows:

$$\mathrm{E}[\mathrm{change}_1 \mid \mathrm{Z}=1] - \mathrm{E}[\mathrm{change}_1 \mid \mathrm{Z}=0]$$

In the given DAGs, this is equivalent to defining the estimand of interest as

$$\mathrm{E}\left[\mathrm{change}_1^* \mid \mathrm{Z}=1, \mathrm{P}=0\right] - \mathrm{E}\left[\mathrm{change}_1^* \mid \mathrm{Z}=0, \mathrm{P}=0\right]$$

because the unobserved variable $change_1$ equals $change_1^* \mid P=0$, and conditioning on P does not open any backdoor path between Z and $change_1$.

7 | ESTIMATORS

7.1 | Reference estimators for the true value of the estimands

We define the true value of the estimand (a) as the treatment effect estimate derived from a linear regression model with treatment and baseline severity as linear predictors, using the underlying values of the outcome before applying the effect of the COVID-19-related IE (Y_1). This estimator cannot be obtained in practice as Y_1 is unobserved but is available to us given the simulation nature of our experiment.

This will be our estimator of reference for (a) that other estimators will be compared against.

We define the true value of the estimand (b) as the difference in responder proportions based on the underlying values of the outcome before applying the effect of the COVID-19-related IE (i.e., on Y_1). Due to the small sample size and small expected responder proportion under the null-hypothesis, Fisher's exact test is used for the responder analysis.

This will be our estimator of reference for (b) that other estimators will be compared against.

In the following section, we briefly describe the candidate estimators investigated for the continuous YGTSS-TTS and for the dichotomized responder endpoint. We first describe the continuous analysis methods and thereafter introduce analogous approaches for the binary analysis, by modifying or building on the continuous estimators. Candidate estimators have been chosen from different classes of approaches including different regression approaches, multiple imputation, weighting-based methods and de-mediation approaches to cover a wide range of possible candidates. Methods have been grouped according to the use of values collected after the occurrence of COVID-19-related IEs, as those represent two radically different approaches.

7.2 | Methods using the values collected after occurrence of COVID-19-related IE

7.2.1 | Observed values analysis

For the continuous analysis, this approach uses a linear regression model with the observed relative change $change_1^*$ as dependent variable irrespective of the patient being affected by the IE, and baseline YGTSS-TTS Y_0 and treatment Z as independent variables. The linear regression is implemented in R using the *glm* function.

For the binary analysis, both estimation and significance testing are as per the reference estimator but using the responder proportions calculated from the observed outcome Y_1^*.

This method is expected to be unbiased under the null hypothesis, as the IE would affect both arms equally by a parallel shift in the continuous score with the same starting point and hence would disappear by using the difference in

between arms as the effect estimate both in the binary and continuous analysis. Under the alternative hypothesis in the multiplicative scenario, the IE would impact the two treatment arms differently, and effect estimate derived from the linear regression model would be biased as well as the difference in responder proportions. Under the alternative hypothesis in the additive scenario, the continuous analysis is expected to be unbiased again, as the impact of the IE would be equal on both treatment arms. However, the binary analysis is expected to be biased, as the parallel shift due to the IE in both treatment arms would have different starting points (since the treated patients are expected to have a lower score) and thus affect the responder proportions after the dichotomization of the continuous score (see[13] for a graphical representation).

7.2.2 | COVID-19 covariate

For the continuous analysis only, this approach uses a linear regression model with the observed relative change $change_1^*$ as dependent variable irrespective of the patient being affected by the IE, and baseline YGTSS-TTS Y_0, Z and P, a categorical variable indicating whether the patient was affected by the IE, as independent variables. The linear regression is implemented in R using the *glm* function.

The performance of this method is expected to be similar to the "observed values analysis" described above with a gain in precision due to the adjustment for the occurrence of the IE.

7.2.3 | Loh's g-estimation—de-mediation for Y_1^*

Loh and colleagues[21] propose a g-estimation method for estimating controlled direct effects in randomized studies, where mediators are present. By extension, we propose that this allows comparing the potential outcomes under fixed values of the treatment and of other events affecting the outcome (in our case, not occurrence of COVID-19-related IEs). Loh's method was applied to the simulated trials by the following steps:

i. fit a logistic model for the expected value of P (i.e., being affected or not by the COVID-19-related IE) using the assigned treatment Z and Y_0 as predictors;
ii. predict the probability p_P (i.e., being affected or not by the COVID-19-related IE) by the logistic model for all patients;
iii. fit a linear model for the expected Y_1^* using as predictors p_P, P, Z and Y_0;
iv. compute a transformed response variable $R_{0,j} = Y_{1,j}^* - coeff(P) * P_j$ using the coefficient from step (iii) to remove the effect of the potential the occurrence of the COVID-19-related IE on the outcome, and calculate the relative change as $R_{1,j} = \frac{R_{0,j} - Y_0}{Y_0}$
v. run a linear regression on the transformed response variable R_1 with Z and Y_0 as linear predictors to estimate the controlled direct effect of the treatment.

We used the model-derived standard error from step (iv) for deriving confidence intervals and statistical inference. In a reduced set of simulations, this showed similar performance to the more computationally intensive bootstrapping approach recommended by Loh et al.

For the binary analysis, we follow the steps (i) to (iv) as outlined above and use R_j for deriving the responder status $respond_j$ before calculating the difference in responder proportions and applying an exact Fisher test.

7.2.4 | Loh's g-estimation 1—de-mediation for $change_1^*$

As an alternative approach to de-mediating the impact of the intercurrent event on the dependent variable relative change directly, we modified step (iii) and (iv) of Loh's g-estimation. In step (iii), a model for the expected relative

change is build using as predictors p_P, P, Z and Y_0. In step (iv), the impact is de-mediated via $R_{1j} = \text{change}^*_{1j} - \text{coeff}(P) * P_j$ directly, without the need to calculate R_{0j} in an intermediate step. As for the un-modified approach, statistical inference is based on model-derived standard errors.

7.2.5 | Loh's g-estimation 2—de-mediation for $\log(Y^*_1)$

To allow de-mediation in a case where the IE impacts the YGTSS-TTS multiplicatively, we modified steps (iii) and (iv) of Loh's g-estimation by using a logistic function for de-mediating. In step (iii), a model for the expected $\log(Y_1)$ is build using as predictors p_P, P, Z and Y_0.In step (iv), the impact is de-mediated via $R_{0j} = \exp\left(Y^*_{1j} - \text{coeff}(P) * P_j\right)$ and the relative change is calculated as $R_{1j} = \frac{R_{0j} - Y_0}{Y_0}$. As for the un-modified approach, statistical inference is based on model-derived standard errors.

7.2.6 | Adaptive de-mediation g-estimation

Investigating a data-driven approach to de-mediation, we combined the additive de-mediation approaches for $change^*_1$ and Y^*_1 and the multiplicative de-mediation approach outlined above to a candidate estimator using R^2 for selecting the de-mediation approach: based on the model estimating the de-mediation effect in step (iii) of the (modified) Loh's g-estimation, we selected the de-mediation approach based on the approach with the higher value of the R^2 for estimating the de-mediation effect Note that due to the same number of predictors in each de-mediation model, this is equivalent to using adjusted R^2. Other criteria for model selection (like AIC or BIC) are not suited to our specific situation, as the de-mediation models are based on different outcome scales (absolute values of YGTSS-TTS, log transformed absolute values of YGTSS-TTS or relative change), and therefore no meaningful comparison of the AICs can be made, but could be candidates in case all de-mediation models use the same outcome.

As for the individual approaches, for statistical inference we used the model-derived standard error of the selected model separately.

7.2.7 | Sequential g-estimation

G-estimation aims at estimating the controlled direct effect[22] based on the variance estimation from Acharya et al.[23] We applied the method using the *sequential_g* function as implemented in the R package *DirectEffects*,[24] specifying $change^*_1$ as the outcome, treatment allocation Z and Y_0 as baseline variables, Z and Y_0 as intermediate variables and P as the only variable in the demediation function. The sequential g-estimation approach implemented in the *sequential_g* function in principle follows the same approach as Loh's sequential g-estimation. But in contrast to Loh's g-estimation described above, steps (i) and (ii) are omitted and the predicted probability p_P is omitted from the de-mediation model in step (iii). Loh and colleagues showed that the estimator for the effect of the mediator derived in step (iii) from Loh's g-estimation is consistent even if the outcome model in step (iii) is incorrectly specified. In contrast, the estimator for the effect of the mediator on the outcome derived from the sequential g-estimation approach is only consistent if the outcome model is correctly specified.[21]

The expectation for the performance of the g-estimation approaches are as follows. The two key assumptions on this method are that (1) no unobserved confounder between treatment and outcome and (2) no unobserved confounders between the intercurrent event and the outcome exist. Both (1) and (2) are met in all scenarios. Due to the causal structure used in the DGM for this simulation study, both assumptions are fulfilled. Additionally, the (mis)specification of the de-mediation model influences the performance of the estimators.

For all additive scenarios, the sequential g-estimation method and Loh's g-estimation de-mediating for Y^*_1 are expected to be unbiased (that is, to target the estimand of interest unbiasedly). In the multiplicative scenarios, Loh's g-estimation de-mediating for $\log(Y^*_1)$ is expected to be unbiased. Additionally, for the same reasons given above for the 'observed case analysis', under the Null hypothesis, all g-estimation methods are expected to be unbiased. In all other cases, the respective methods are expected to show a bias with unknown magnitude.

7.3 | Methods not using the values collected after occurrence of Covid-19-related IE

All method not using the values collected after the occurrence of the IE should be unbiased estimators for the target estimand both in the additive and multiplicative scenario, as the occurrence of the IE is completely random in the DGM. Therefore, excluding affected patients does not introduce bias, but increase variability. The increase in variability is expected to be largest for the 'unaffected case analysis', and to be compensated to some degree by the weighting-based and imputation-based methods. Hence, in principle no increase in type I error is expected, but a decrease in power is expected proportional to the proportion of affected patients to different degrees.

However, due to the small sample size, computational problems might affect the below methods differentially as some methods might be sensitive to the absolute number of observations available for modeling/analysis.

For example, both inverse probability weighting methods (normal and doubly robust), should be unbiased, if the following assumptions as outlined by Hernán, are fulfilled[25]:

- the average outcome in the individuals not being affected by the IE must equal the unobserved average outcome in the individuals affected by the IE with the same values for treatment Z, Y_0 and Y_1^* (exchangeability),
- all conditional probabilities of not being affected by the IE, $P(P=0|Z=z, Y_1^*=y_1^*, Y_0=y_0)>0)$ given realizations of the variables Z, Y_0 and Y_1^* must be greater than zero (positivity), and
- a well-defined treatment (consistency).

Due to the causal structure depicted in Figures 1 and 2 used for our simulation study, exchangeability and consistency are given. In principle (that is, in expectation) also positivity is given, since occurrence of an IE is always random in our data-generating model (that is, it does not depend on any measured variable), and hence, the expected conditional probability for not being affected by a COVID-19-related IE is larger than 0 for all realizations of variables Z , Y_0 and Y_1^*. Thus, we do not have a "structural violation" of the positivity assumption.[25] However, in any simulated study, by chance there might be a threshold $a>0$, so that (1) there are patients with $Y_1^*>a$ and (2) all patients with $Y_1^*>a$ are affected by an IE. Consequently, the conditional probability for not being affected by the IE is 0 for a subset of patients with $P(Y_1^*>a)>0$. This "random violation" of the positivity assumption could lead to biased estimates. This problem has also been highlighted Mallinckrodt and colleagues,[26] but the magnitude of potential bias cannot be anticipated.

7.3.1 | Unaffected cases analysis

This approach includes only patients unaffected by the IE. For the continuous analysis, a linear regression model with the observed relative change $change_1^*$ as dependent variable, and baseline YGTSS-TTS Y_0 and Z as independent variables. The linear regression is implemented in R using the *glm* function.

The responder proportions are compared based on Fisher's exact test and a risk difference is computed.

Given that the measurements of interest are those under unaffected conditions, this method is equivalent to using a complete case analysis considering all data from affected patients as missing. As the structure of the (artificial) missing data problem is MCAR (see DAGs in Figures 1 and 2), it can be expected that the method will not be biased, but lose precision and power due to not using the information from the affected patients.[27]

7.3.2 | Predictive mean matching

Predictive mean matching was implemented in the following steps:

i. missing data are imputed through predictive mean matching[28] using the mice function implemented in the R package *Mice*.[29] Values after the occurrence of the IE are set to missing and imputed using Y_0 and Z in the prediction matrix. The mice function imputes missing values by first building a linear regression on all complete cases based on Y_0 and Z generating regression parameter estimates $\widehat{\beta}$. Secondly, regression parameters $\widetilde{\beta}$ are randomly drawn from the posterior predictive distribution of $\widehat{\beta}$. Thirdly, the outcome for all cases is predicted based on $\widetilde{\beta}$ to identify in a fourth step which complete cases are most similar in the predicted values to the predicted value of the missing value. Lastly, the observed outcome value for one of the closest five neighbor cases is randomly chosen to replace

each missing value. While Schafer[30] proposed to impute five datasets, an a priori investigation showed a better performance (equal bias and T1E, but slightly larger power) for 50 imputed datasets. Consequently, for all multiple imputation methods, 50 datasets were imputed.

ii. For the continuous analysis, linear regression analyses with Y_0 and Z as predictors are performed, as for the reference estimators on the imputed datasets and results is pooled.[31]

iii. For the binary analysis, the responder status $respond_i$ is determined based on the imputed data sets. Next, risk differences are calculated, and Fisher's test is applied to the 5 sets of responder strata, and the final risk difference is calculated as the average of the imputed dataset-specific risk differences. Additionally, a p-value is derived by taking the median value of the p-values calculated for the imputed data sets.[32]

We investigated both, the predictive mean matching implementation using meth="pmm," and the Midas touch algorithm developed by Gaffert et al.[33] by specifying meth="midas."

7.3.3 | Multiple Imputation—normal

Parametric multiple imputation was applied in line with Bayesian linear regression as outlined in Rubin[34] by the following steps:

i. missing data are imputed using the *mice* function implemented in the R package *Mice*[29] by specifying meth="normal." Values after the occurrence of the IE are set to missing and imputed using Y_0 and Z in the prediction matrix. The *mice* function based on Bayesian linear regression imputes missing values in line with Rubin[31] by first building a linear regression model on all complete cases based on Y_0 and Z generating regression parameter estimates $\widehat{\beta}$ and estimated variances $\widehat{\sigma}^2$. Secondly, regression parameters $\widetilde{\beta}$ and a variance $\widetilde{\sigma}^2$ are randomly drawn from the posterior predictive distribution of $\widehat{\beta}$ and $\widehat{\sigma}^2$. Lastly, Y_1 is imputed as a random draw from the predictive distribution based on $\widetilde{\beta}$ and $\widetilde{\sigma}^2$.

Subsequently, analyzing the imputed data sets and pooling the results was conducted exactly as step (ii) for predictive mean matching for the continuous and binary variable.

7.3.4 | Multiple Imputation based on logistic regression

For the responder analysis only, we investigate also the approach of directly imputing a binary variable instead of imputing on the continuous scale and dichotomizing the imputed value as outlined for the MI methods above. Using the *mice* function with specification meth = "logreg," we implemented imputation by logistic regression, which follows the steps outlined by Rubin[31]: Values after the occurrence of the IE are set to missing and imputed using Y_0 and Z in the prediction matrix. Based on the non-affected values, a logistic regression with logit link function with the responder status as dependent and Y_0 and Z as independent variables is estimated, generating logistic regression parameter estimates $\widehat{\beta}$. Secondly, regression parameters $\widetilde{\beta}$ are randomly drawn from the posterior predictive distribution of $\widehat{\beta}$. To generate the responder status, random uniform variables are compared with the inverse logit of the predicted value. Missing values are imputed as responders, if the random uniform variable is smaller than the inverse logit of the predicted value.

Subsequently, analyzing the imputed datasets and pooling the results was conducted exactly as step (ii) for predictive mean matching for the binary endpoint only.

7.3.5 | Multiple imputation by Classification and Regression Trees (CART)

As for the above, this method only applies to the responder analysis. Values after the occurrence of the IE are set to missing and the responder status is imputed using Y_0 and Z in the prediction matrix. The method has been employed as implemented in the R package *mice*. Based on complete cases, a classification tree is created with splits based on

values of the variables included in the prediction matrix.[35,36] Trees are built so that any terminal "node" contains at least five observations. Each case with missing outcome (i.e., each patient with the IE) is classified based on the tree and a random value from those in the same "leaf" is assigned as imputed value. This process is used to build five imputed datasets and the results from the analysis in the five datasets imputed were pooled as described above.

7.3.6 | Inverse probability weighting

Inverse probability weighting was carried out for the continuous analysis only in the following steps[37]:

i. a logistic regression model with $1-P$ as dependent variable is fitted using Y_0 and Z as independent variables;
ii. this model is used to predict probability of the IE not occurring;

iii. the weights are computed as the inverse of the probability computed above. Extreme weights are trimmed—bottom and top 2.5% weights are replaced by the 2.5% and 97.5% weight percentiles;

iv. a weighted linear regression model with Y_0 and Z as independent variables on $change_1$ is performed, using data only from patients without occurrence of the IE; and
v. standard errors are computed by stratified (by $Z*P$) bootstrapping using the *boot* function from the R package *boot*.[38,39] Stratified boostrap (in contrast to unstratified boostrap) was chosen to avoid computational problems that would occur due to the overall small sample size for the unstratified boostrap for higher proportions of affected patients because in the sampled data all unaffected patients might have the same treatment status, and hence, the regression model in step (iv) with Z as independent variable could not be calculated.

7.4 | Performance criteria

The following performance criteria were used in evaluating the performance of the estimators in the scenarios tested[40]:

- bias, expressed on the scale of the YGTSS-TSS and estimated as $\frac{1}{n_{sim}}\sum\widehat{\theta}_k-\theta$ where θ is the true value if the estimand, n_{sim} is the number of simulations, $\widehat{\theta}_k$ is the value of the estimand estimated by each method in a specific realization and θ is the true value of the estimand in the respective scenario);
- length of the 95% confidence interval estimates of $\widehat{\theta}_k$. This is estimated in each run as the difference between upper and lower bound, and averaged thereafter;
- Coverage, estimated as the % of times θ falls within the 95%-confidence interval for $\widehat{\theta}_k$;
- mean squared error, *MSE*, estimated as $\frac{1}{n_{sim}}\sum(\widehat{\theta_k}-\theta)^2$
- square root mean squared error, $\sqrt{MSE}$, expressed on the scale of the relative change and estimated as $=\sqrt{\frac{1}{n_{sim}}\sum(\widehat{\theta_k}-\theta)^2}$;
- empirical alpha, measured as the proportion of cases where the null hypothesis was rejected. This will be reported and interpreted in different ways depending on whether the scenarios were simulated under null or alternative hypothesis. Under the null hypothesis, this measure will be reported as empirical type I error rate, and under the alternative hypothesis as empirical power.

For all performance criteria, Monte Carlo estimates of the Standard Errors are computed.[40] For the binary analysis, only bias, mean squared error, square root mean squared error and the empirical alpha are calculated.

8 | RESULTS

The following sections display the performance of the candidate estimators under the above-described data-generating mechanisms for Null hypothesis and alternative hypothesis depending on the proportion of affected patients. Tables include all investigated methods while the figures include a selection only. For methods from the same class

following a similar approach (e.g., predictive mean matching using the *midastouch* or the *pmm* algorithm), only one representing method was selected for the figures if they showed similar performance throughout all investigated scenarios.

8.1 | Null hypothesis

All methods investigated display a negligible bias and imprecision under the null hypothesis, regardless of the modeling assumption for the effect of the IE, both for the continuous (Figures S1 and S2 and Tables S1–S4) and for the binary analysis (Figures S3 and S4, lower panels and Tables S5–S8). Coverage is close to the nominal level (Tables S9 and S10), except—at higher proportions of patients affected—for Multiple imputation via the Normal model and for the weighting-based methods.

The type I error is controlled below the value of one-sided 0.025 for most methods. However, for the continuous analysis (Figures 3 and S5, S11, and S12), the weighting-based estimators have an increased type I error rate for scenarios with proportions of affected patients ≥70%. For the binary analysis (Figures 4 and S6 and Tables S13 and S14), an inflation of type I error rates can be observed for the Multiple Imputation with the Logistic model.

The length of the confidence interval markedly increases for the imputation and weighting-based methods with proportions of patients increased above 60% (Tables S15 and S16).

8.2 | Alternative hypothesis

The scenarios under the alternative hypothesis have been simulated with a real value of the estimands of a difference between mean relative change in YGTSS-TTS between treatment groups of $-0.122 - (-0.025) = 0.097$ for the continuous analysis and of a difference between responder proportions of $29\%–1\% = -28\%$ for the binary analysis.

The methods tested mostly displayed a negligible bias for both the continuous (Figures 5 and S7 and Tables S17 and S18) and the binary (Figure 6 and S8, lower panels and Tables S19 and S20) estimands. However, the Observed Values estimator has a slight bias that is proportional to the proportion of affected patients. Under the multiplicative scenario, Loh's g-estimation approach displays a similar behavior when the de-mediation model is miss-specified. Furthermore, imputation methods display a bias at very high proportions of affected patients.

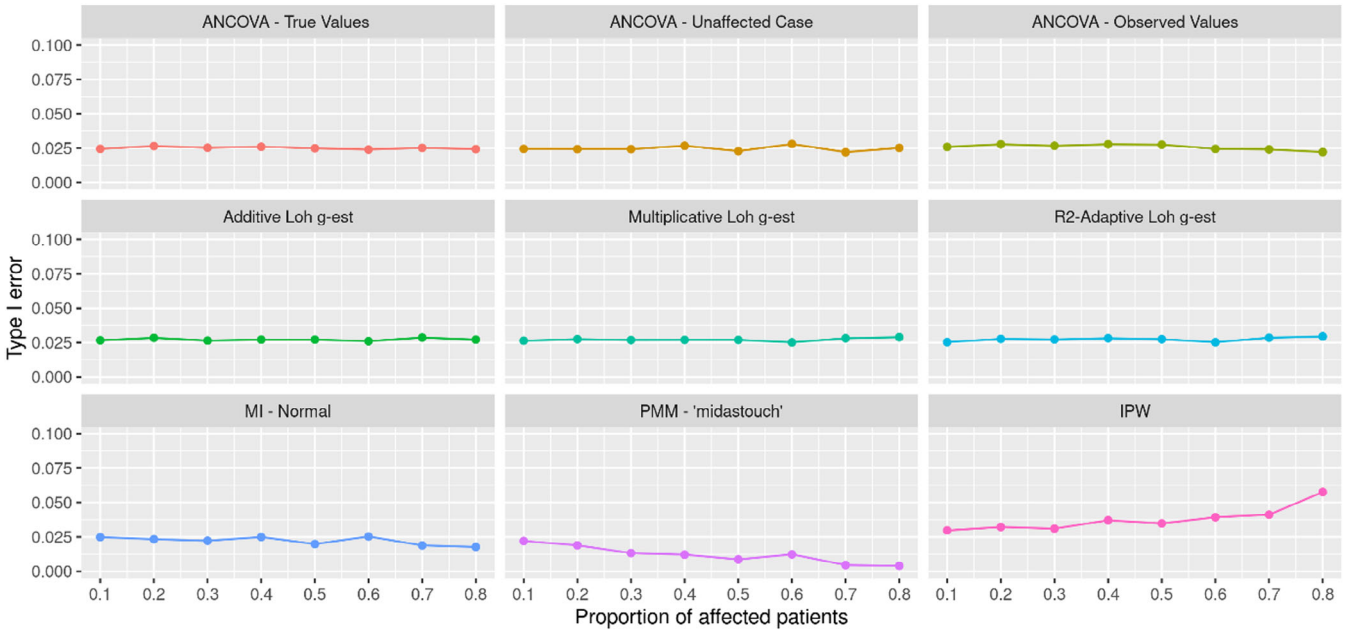

FIGURE 3 Type I error of the candidate estimators for the continuous estimand, multiplicative IE impact

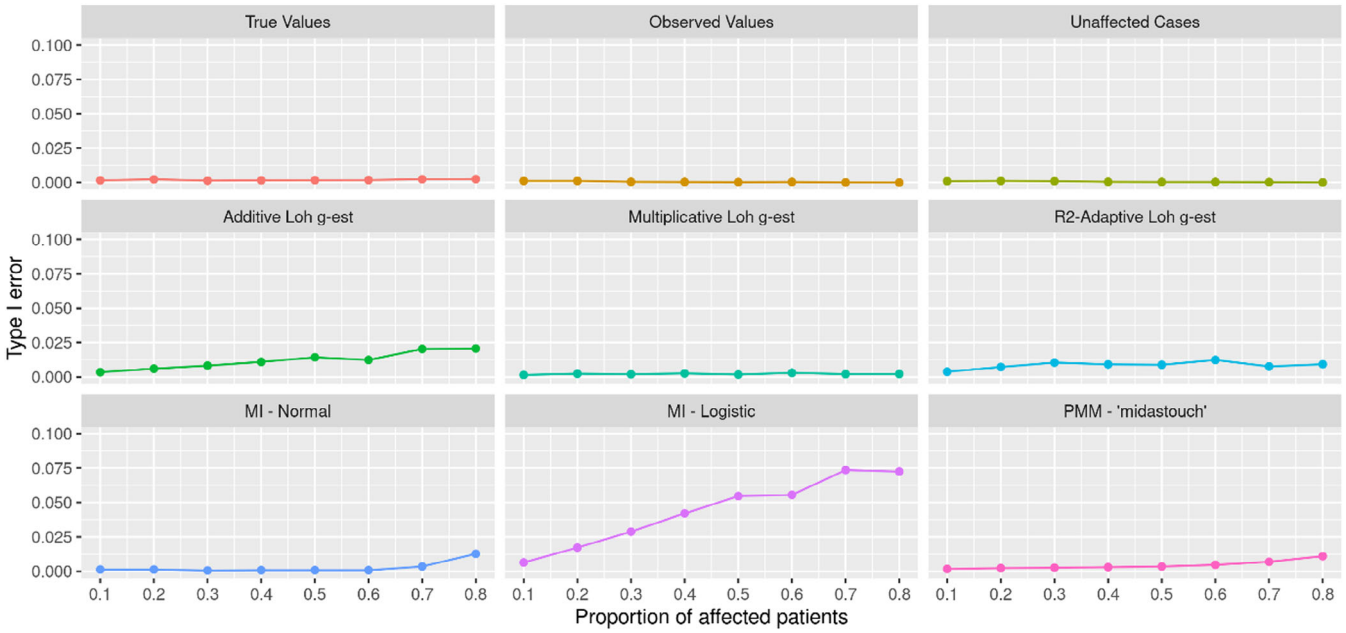

FIGURE 4 Type I error of the candidate estimators for the binary estimand, multiplicative IE impact

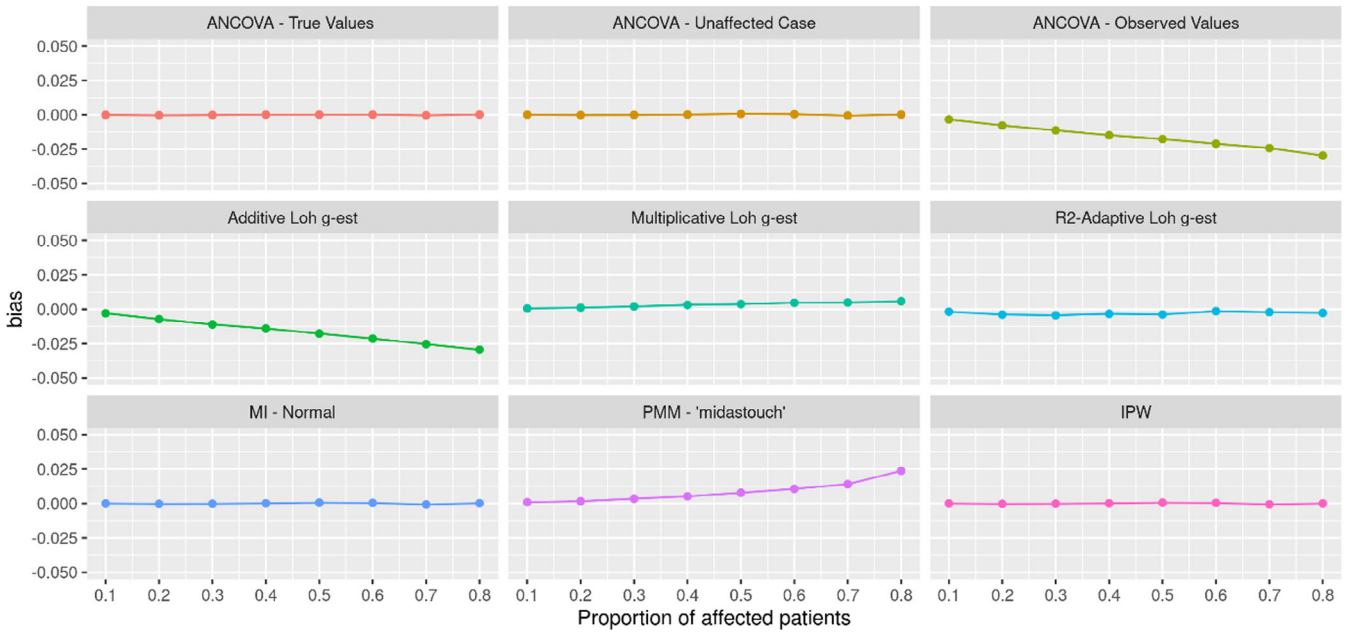

FIGURE 5 Bias of the candidate estimators for the continuous estimand under the alternative hypothesis, multiplicative IE impact

For both the continuous (Figures 7 and S9 and Tables 2 and 3) and the binary (Figures 8 and S10 and Tables 4 and 5) estimands, all methods that do not use post-IE data show sever power losses in all scenarios with at least 30% affected patients as compared to the reference estimator. The Observed values estimator retains power for the continuous analysis under the additive scenario, but not under the multiplicative scenario. The de-mediation approaches retain power in all scenarios, regardless of the misspecification of the de-mediation model. The observed values estimator displays a similar behavior regarding precision (Figures S11–S14 and Tables S21–S24) and confidence interval length (Figures S15 and S16, and Tables S25 and S26). Confidence interval length consistently increases with the increase of the percentage of affected patients for all the methods that do not use post-IE data. Despite larger confidence intervals, coverage is

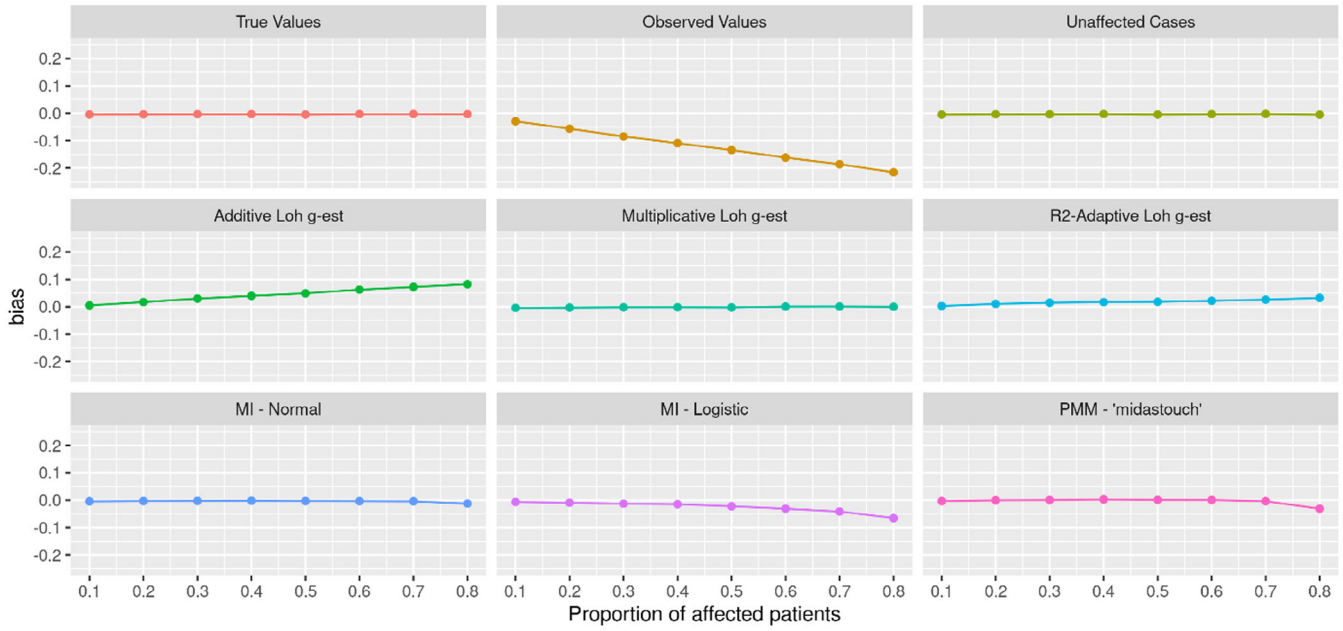

FIGURE 6 Bias of the candidate estimators for the binary estimand under the alternative hypothesis, multiplicative IE impact

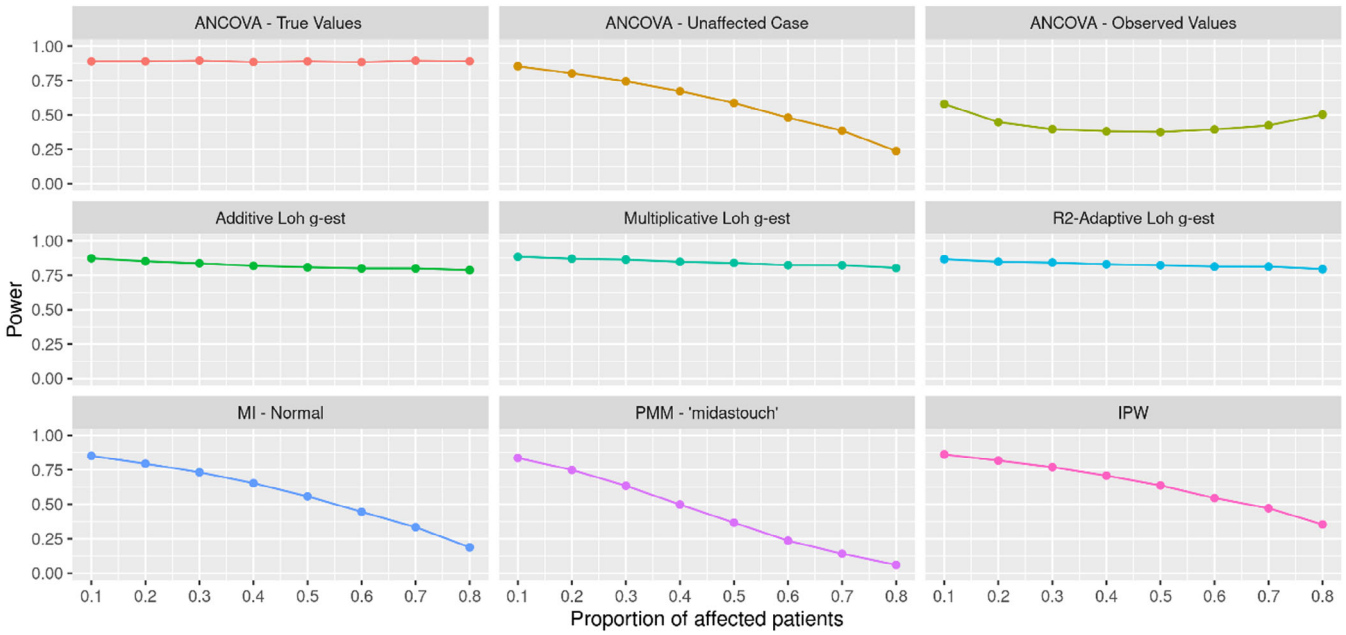

FIGURE 7 Power of the candidate estimators for the continuous estimand under the alternative hypothesis, multiplicative IE impact

markedly decreased at high proportions of affected patients for the Multiple Imputation methods and for the weighting-based approaches (Tables S27 and S28).

It needs to be noted that the treatment effect specified in the data-generating mechanism is smaller in the continuous scenario as compared to the binary scenario to ensure a nominal power of 90% in each scenario. This could contribute to the more rapid power loss of the multiple imputation approaches in the continuous scenario as compared to the binary scenario (see performance of 'MI normal' in Figures 7 and 8). However, not all methods that use the continuous endpoint also as a basis for the binary analysis show the same pattern. Notably, the de-mediation approaches also use the continuous endpoint for de-mediation before dichotomizing the de-mediated outcome, but performance is equally good in the binary and the continuous scenarios (see Figures 7, 8, S9, and S10).

TABLE 2 Power for the continuous analysis under the alternative hypothesis, additive scenario

Method	10%	20%	30%	40%	50%	60%	70%	80%
ANCOVA—true values	0.89 (0.003)	0.885 (0.003)	0.89 (0.003)	0.886 (0.003)	0.884 (0.003)	0.884 (0.003)	0.89 (0.003)	0.89 (0.003)
ANCOVA—unaffected case	0.858 (0.003)	0.798 (0.004)	0.747 (0.004)	0.675 (0.005)	0.597 (0.005)	0.487 (0.005)	0.382 (0.005)	0.238 (0.004)
ANCOVA—observed values	0.882 (0.003)	0.872 (0.003)	0.874 (0.003)	0.868 (0.003)	0.863 (0.003)	0.863 (0.003)	0.874 (0.003)	0.878 (0.003)
ANCOVA—covariate	0.883 (0.003)	0.879 (0.003)	0.884 (0.003)	0.878 (0.003)	0.877 (0.003)	0.877 (0.003)	0.881 (0.003)	0.884 (0.003)
Sequential g-estimation	0.892 (0.003)	0.885 (0.003)	0.89 (0.003)	0.888 (0.003)	0.885 (0.003)	0.885 (0.003)	0.889 (0.003)	0.893 (0.003)
Additive Loh g-est, de-med change	0.889 (0.003)	0.887 (0.003)	0.89 (0.003)	0.885 (0.003)	0.883 (0.003)	0.883 (0.003)	0.887 (0.003)	0.889 (0.003)
Additive Loh g-est, de-med Y1	0.89 (0.003)	0.885 (0.003)	0.889 (0.003)	0.885 (0.003)	0.884 (0.003)	0.884 (0.003)	0.889 (0.003)	0.889 (0.003)
Multiplicative Loh g-est	0.889 (0.003)	0.884 (0.003)	0.89 (0.003)	0.884 (0.003)	0.882 (0.003)	0.882 (0.003)	0.887 (0.003)	0.887 (0.003)
Adaptive Loh g-est—r-squared	0.889 (0.003)	0.884 (0.003)	0.889 (0.003)	0.886 (0.003)	0.883 (0.003)	0.883 (0.003)	0.888 (0.003)	0.888 (0.003)
MI—Normal	0.855 (0.004)	0.792 (0.004)	0.733 (0.004)	0.655 (0.005)	0.567 (0.005)	0.444 (0.005)	0.333 (0.005)	0.185 (0.004)
PMM—"pmm"	0.852 (0.004)	0.78 (0.004)	0.713 (0.005)	0.627 (0.005)	0.538 (0.005)	0.43 (0.005)	0.344 (0.005)	0.225 (0.004)
PMM—"midastouch"	0.842 (0.004)	0.749 (0.004)	0.633 (0.005)	0.505 (0.005)	0.366 (0.005)	0.233 (0.004)	0.143 (0.003)	0.062 (0.002)
Inverse probability weighting	0.87 (0.003)	0.817 (0.004)	0.771 (0.004)	0.713 (0.005)	0.643 (0.005)	0.547 (0.005)	0.463 (0.005)	0.361 (0.005)

The approach we have implemented for a data-driven selection of the de-mediation model based on the R^2 of the respective de-mediation models shows mixed performance in discriminating between an additive and a multiplicative underlying data-generating model (Table S29) with miss-classification up to 40% under the alternative hypothesis. While in the additive scenarios, the proportion of runs in which the adaptive approach wrongly selects the multiplicative de-mediation model does not change qualitatively depending on the proportion of affected patients, in the multiplicative scenarios the proportion of correctly selecting the multiplicative de-mediation model increases with the proportion of affected patients. However, classification improves with higher proportions of affected patients (where using the right de-mediation model has a larger impact on the results). Especially, in the multiplicative scenarios, this leads to comparable performance to the correct multiplicative de-mediation model.

9 | DISCUSSION

Based on the CANNA-TICS trial, this simulation study has investigated the performance of candidate estimators for an estimand applying a hypothetical strategy for COVID-19-related IEs with the relative change in the YGTSS-TTS scale as continuous and a dichotomized responder criterion as binary endpoints under a small sample size. The IEs was modeled to affect the underlying value on the YGTSS-TTS both in an additive and in a multiplicative way, and the proportion of affected patients was varied from 10% to 80% to cover a wide range of application and investigate the performance of the estimators also in extreme situations.

As expected, there was no bias under the null hypothesis. Under the alternative hypothesis modest bias could be observed for some candidate estimators and the bias depends both on the data-generating model for the impact of the IE (additive or multiplicative) and the proportion of affected patients. For the binary analysis, the only methods showing an inflated type I error rate are multiple imputation with a logistic model and weighting-based methods. For multiple imputation on the continuous scale via a normal model followed by dichotomization, an inflated type I error rates could only be observed in the scenario where 80% of the patients are affected by the IE. Overall, type I error was inflated only for high proportions of affected patients in our simulation. For the continuous analysis, the weighting-based approaches displayed increased type I error rates for high proportions of affected patients.

Of more concern, all methods that discarded post-IE data displayed a sharp decrease of power proportional to the number of affected patients.

TABLE 3 Power for the continuous analysis under the alternative hypothesis, multiplicative scenario

Method	10%	20%	30%	40%	50%	60%	70%	80%
ANCOVA—true values	0.889 (0.003)	0.889 (0.003)	0.894 (0.003)	0.884 (0.003)	0.889 (0.003)	0.884 (0.003)	0.894 (0.003)	0.89 (0.003)
ANCOVA—unaffected case	0.854 (0.004)	0.802 (0.004)	0.744 (0.004)	0.672 (0.005)	0.587 (0.005)	0.48 (0.005)	0.386 (0.005)	0.238 (0.004)
ANCOVA—observed values	0.579 (0.005)	0.448 (0.005)	0.397 (0.005)	0.382 (0.005)	0.376 (0.005)	0.395 (0.005)	0.424 (0.005)	0.503 (0.005)
ANCOVA—covariate	0.867 (0.003)	0.845 (0.004)	0.827 (0.004)	0.812 (0.004)	0.799 (0.004)	0.791 (0.004)	0.791 (0.004)	0.779 (0.004)
Sequential g-estimation	0.877 (0.003)	0.851 (0.004)	0.84 (0.004)	0.823 (0.004)	0.812 (0.004)	0.804 (0.004)	0.805 (0.004)	0.794 (0.004)
Additive Loh g-est, de-med change	0.873 (0.003)	0.851 (0.004)	0.836 (0.004)	0.818 (0.004)	0.807 (0.004)	0.799 (0.004)	0.799 (0.004)	0.787 (0.004)
Additive Loh g-est, de-med Y1	0.854 (0.004)	0.823 (0.004)	0.802 (0.004)	0.781 (0.004)	0.768 (0.004)	0.761 (0.004)	0.761 (0.004)	0.755 (0.004)
Multiplicative Loh g-est	0.884 (0.003)	0.87 (0.003)	0.863 (0.003)	0.847 (0.004)	0.839 (0.004)	0.823 (0.004)	0.823 (0.004)	0.803 (0.004)
Adaptive Loh g-est—r-squared	0.867 (0.003)	0.847 (0.004)	0.842 (0.004)	0.829 (0.004)	0.822 (0.004)	0.813 (0.004)	0.813 (0.004)	0.794 (0.004)
MI—normal	0.852 (0.004)	0.795 (0.004)	0.733 (0.004)	0.653 (0.005)	0.557 (0.005)	0.445 (0.005)	0.334 (0.005)	0.188 (0.004)
PMM—"pmm"	0.846 (0.004)	0.785 (0.004)	0.71 (0.005)	0.625 (0.005)	0.527 (0.005)	0.43 (0.005)	0.348 (0.005)	0.223 (0.004)
PMM—"midastouch"	0.837 (0.004)	0.75 (0.004)	0.635 (0.005)	0.498 (0.005)	0.367 (0.005)	0.236 (0.004)	0.142 (0.003)	0.06 (0.002)
Inverse probability weighting	0.862 (0.003)	0.818 (0.004)	0.77 (0.004)	0.708 (0.005)	0.637 (0.005)	0.545 (0.005)	0.472 (0.005)	0.354 (0.005)

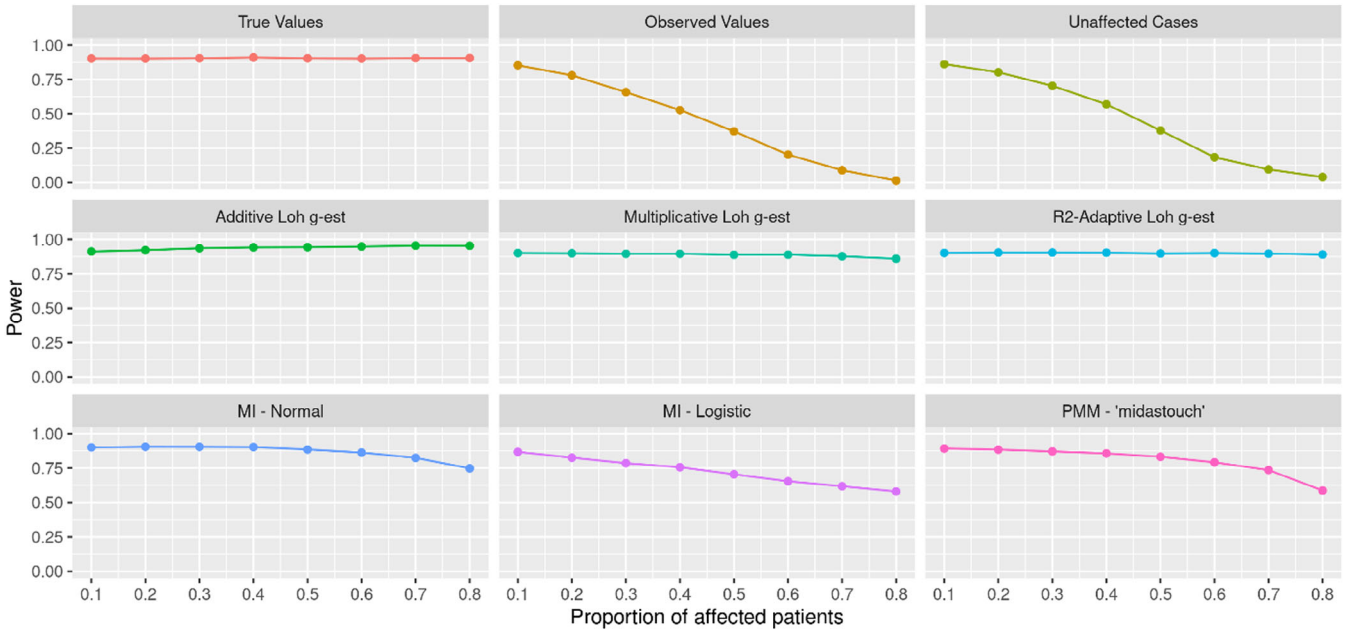

FIGURE 8 Power of the candidate estimators for the binary estimand under the alternative hypothesis, multiplicative IE impact

While there is no universally best estimator for both the additive and the multiplicative data-generating mechanism, neither for the continuous nor for the binary analysis, we generally observe that de-mediation approaches have the best performance. This is in line with what we have recently shown[41] for handling the use of symptomatic medication with a hypothetical strategy in clinical trials for disease modifying Alzheimer's Disease treatments. While the data-generating mechanism and the causal structure are different in this case, in both settings the observed values—albeit affected by an IE to be treated with the hypothetical strategy—can be assumed to carry information on the targeted (unobservable) value. Here, we formulate the hypothesis that in settings where this condition is met the de-mediation approaches have an advantage compared to estimators that set data as missing. Interestingly, this also applies where the IE is not a mediator (as in this case, where the treatment is not a cause of the IE).

In the scenario where the IE influences the YGTSS-TTS in an additive way, the de-mediation approach proposed by Loh et al.[21] on the additive scale (de-mediating the effect of the IE on the continuous outcome variable *relative change*) followed by dichotomization performs best for the binary analysis. Similarly, in the scenario of a multiplicative impact of the IE on the YGTSS-TTS, the modified de-mediation approach using the log-transformed YGTSS-TTS for de-mediation (de-mediation on a multiplicative scale) followed by dichotomisation performs best. While there is no universally best estimator, the performance of the additive de-mediation under a multiplicative impact of the IE and vice versa are close to the optimal estimator. Also in the continuous analysis, the de-mediation approach based on the correct model for the impact of the IE performs best. However, in contrast to the binary analysis, the difference between the additive and the multiplicative de-mediation is notable. Especially, under a multiplicative impact of the IE, the additive de-mediation is biased.

The performance of the adaptive de-mediation approach is in-between the additive and multiplicative approaches. Generally, the proposed data-driven selection of de-mediation model does not consistently identify the underlying data-generating mechanism. However, the discrimination is poorest in situations where the difference between the de-mediation approaches is neglectable (null hypothesis, additive impact of Covid-19). In contrast, in situations where the difference between the de-mediation approaches is largest (multiplicative impact of Covid-19, high proportion of affected patients), the discrimination works well, and the adaptive approach shows comparable performance to using the correct de-mediation. Future work is needed to investigate different approaches to identify the optimal de-mediation model.

Within the multiple imputation approaches for the binary analysis, imputing the continuous endpoint before dichotomizing had a better performance than imputing the response directly, in line with the literature.[42]

For the extrapolation of our findings it is important to note the dependence of our results on the causal structure that we have imposed on the data. In particular, we have not included a causal link between the treatment assignment

TABLE 4 Power for the binary analysis under the alternative hypothesis, additive scenario

Method	10%	20%	30%	40%	50%	60%	70%	80%
True values	0.904 (0.003)	0.904 (0.003)	0.903 (0.003)	0.907 (0.003)	0.9 (0.003)	0.908 (0.003)	0.904 (0.003)	0.902 (0.003)
Observed values	0.884 (0.003)	0.854 (0.004)	0.817 (0.004)	0.78 (0.004)	0.746 (0.004)	0.687 (0.005)	0.624 (0.005)	0.538 (0.005)
Unaffected cases	0.865 (0.003)	0.805 (0.004)	0.705 (0.005)	0.564 (0.005)	0.391 (0.005)	0.187 (0.004)	0.099 (0.003)	0.039 (0.002)
Additive Loh g-est	0.903 (0.003)	0.898 (0.003)	0.895 (0.003)	0.895 (0.003)	0.889 (0.003)	0.886 (0.003)	0.872 (0.003)	0.846 (0.004)
Multiplicative Loh g-est	0.902 (0.003)	0.893 (0.003)	0.885 (0.003)	0.884 (0.003)	0.876 (0.003)	0.869 (0.003)	0.85 (0.004)	0.826 (0.004)
Adaptive Loh g-est—r_squared	0.903 (0.003)	0.895 (0.003)	0.891 (0.003)	0.893 (0.003)	0.883 (0.003)	0.878 (0.003)	0.859 (0.003)	0.839 (0.004)
MI—normal	0.907 (0.003)	0.906 (0.003)	0.907 (0.003)	0.902 (0.003)	0.894 (0.003)	0.865 (0.003)	0.831 (0.004)	0.746 (0.004)
MI—logistic	0.873 (0.003)	0.827 (0.004)	0.788 (0.004)	0.745 (0.004)	0.709 (0.005)	0.658 (0.005)	0.632 (0.005)	0.58 (0.005)
MI—CART	0.868 (0.003)	0.77 (0.004)	0.63 (0.005)	0.495 (0.005)	0.361 (0.005)	0.226 (0.004)	0.114 (0.003)	0.033 (0.002)
PMM—"midastouch"	0.898 (0.003)	0.884 (0.003)	0.875 (0.003)	0.857 (0.003)	0.836 (0.004)	0.793 (0.004)	0.737 (0.004)	0.588 (0.005)

TABLE 5 Power for the binary analysis under the alternative hypothesis, multiplicative scenario

Method	10%	20%	30%	40%	50%	60%	70%	80%
True values	0.902 (0.003)	0.901 (0.003)	0.904 (0.003)	0.909 (0.003)	0.903 (0.003)	0.902 (0.003)	0.905 (0.003)	0.905 (0.003)
Observed values	0.852 (0.004)	0.78 (0.004)	0.657 (0.005)	0.527 (0.005)	0.372 (0.005)	0.203 (0.004)	0.089 (0.003)	0.015 (0.001)
Unaffected cases	0.861 (0.003)	0.802 (0.004)	0.703 (0.005)	0.569 (0.005)	0.379 (0.005)	0.184 (0.004)	0.095 (0.003)	0.04 (0.002)
Additive Loh g-est	0.912 (0.003)	0.922 (0.003)	0.936 (0.002)	0.942 (0.002)	0.943 (0.002)	0.948 (0.002)	0.955 (0.002)	0.954 (0.002)
Multiplicative Loh g-est	0.901 (0.003)	0.899 (0.003)	0.895 (0.003)	0.895 (0.003)	0.888 (0.003)	0.889 (0.003)	0.877 (0.003)	0.86 (0.003)
Adaptive Loh g-est—r_squared	0.902 (0.003)	0.904 (0.003)	0.904 (0.003)	0.903 (0.003)	0.897 (0.003)	0.9 (0.003)	0.896 (0.003)	0.89 (0.003)
MI—normal	0.901 (0.003)	0.906 (0.003)	0.905 (0.003)	0.904 (0.003)	0.886 (0.003)	0.863 (0.003)	0.825 (0.004)	0.749 (0.004)
MI—logistic	0.868 (0.003)	0.826 (0.004)	0.786 (0.004)	0.757 (0.004)	0.705 (0.005)	0.655 (0.005)	0.618 (0.005)	0.581 (0.005)
MI—CART	0.863 (0.003)	0.765 (0.004)	0.626 (0.005)	0.485 (0.005)	0.352 (0.005)	0.233 (0.004)	0.112 (0.003)	0.036 (0.002)
PMM—"midastouch"	0.893 (0.003)	0.885 (0.003)	0.872 (0.003)	0.857 (0.003)	0.833 (0.004)	0.792 (0.004)	0.736 (0.004)	0.587 (0.005)

and the risk to be affected by the pandemic during the trial. While this was judged to be a reasonable assumption for the motivating IEs of this article based on the CANNTA-TICs trial, this is not the case, for example, (i) for trials with response-adaptive allocation or (ii) if the assessment method can be chosen by the patient. In case a causal link exists, the IE becomes a mediator, a case more similar to what has been studied in Lasch et al[41]. Additionally, this work only considers the occurrence of one IE within the trial. For different IEs occurring in the same trial, the estimators would need to be expanded and potentially different directions of the effects of the IE would need to be considered. Finally, we did not simulate missing data, including missingness that might be caused by the intercurrent event. While this limits the direct applicability of our results for the methods that rely on post-IE data, the good performance of the de-mediation approaches, which use the post-IE data, stress the importance not to interrupt data collection after the occurrence of an IE for which the hypothetical strategy is in place. On the other hand, the findings of our simulation study can also be applied to IEs that are not related to the pandemic. While this simulation study is motivated by COVID-19-related IEs, the performance of the candidate estimators can be extrapolated to other scenarios with the same causal structure. For more complex scenarios (e.g., in case the intercurrent event can occur at more than one timepoint), more research on the adaptation of the estimators proposed is needed.

In the conclusion, de-mediation via g-estimation is a promising family of estimators for an estimand that handles COVID-19-related IEs with a hypothetical strategy. Besides showing the best performance in our simulation study, these approaches allow us to estimate the effect of the IE on the outcome and cross-compare between different studies affected by similar (COVID-19 related) IEs. In case results from the previous studies are available and the effect of the IEs has been estimated, these can inform the choice of de-mediation model for subsequent trials and serve as a diagnostic of the analysis model on an interpretable clinical scale. Potentially, prior information (e.g., observational data in the target condition during the pandemic) could even be incorporated directly into the estimation of the effect of the IE. Additionally, discrepancies between studies regarding a well-fitting de-mediation model or the magnitude or direction of the effect of the IEs could trigger follow-up investigations and help to understand the impact of COVID-19 on trials to learn about the future analysis of trials affected by the pandemic.

AUTHOR CONTRIBUTIONS

Conceptualization, Methodology, Software, Formal analysis, Writing—Original Draft: Florian Lasch. *Conceptualization, Methodology, Software, Formal analysis, Writing—Original Draft*: Lorenzo Guizzaro.

ACKNOWLEDGMENTS

We thank the anonymous reviewers and the editor for their valuable comments that helped to improve the manuscript.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest for this article.

DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

ORCID

Lorenzo Guizzaro https://orcid.org/0000-0002-0390-270X

REFERENCES

1. FDA. Statistical Considerations for Clinical Trials During the COVID-19 Public Health Emergency. 2020;
2. EMA. Points to consider on implications of Coronavirus disease (COVID-19) on methodological aspects of ongoing clinical trials.
3. Hamasaki T, Bretz F, Cooner F, LaVange LM, Posch M. Statistical challenges in the conduct and Management of Ongoing Clinical Trials during the COVID-19 pandemic. *Stat Biopharm Res*. 2020;12:397-398. doi:10.1080/19466315.2020.1828692
4. Meyer RD, Daniel Meyer R, Ratitch B, et al. Statistical issues and recommendations for clinical trials conducted during the COVID-19 pandemic. *Stat Biopharm Res*. 2020;12:399-411. doi:10.1080/19466315.2020.1779122
5. ICH. Addendum of Estimands and sensitivity analysis in. *Clin Trials*. 2019. https://database.ich.org/sites/default/files/E9-R1_Step4_Guideline_2019_1203.pdf
6. Akacha M, Kothny W. Estimands: a more strategic approach to study design and analysis. *Clin Pharmacol Ther*. 2017;102:894-896. doi:10.1002/cpt.872
7. Jin M, Liu G. Estimand framework: delineating what to be estimated with clinical questions of interest in clinical trials. *Contemp Clin Trials*. 2020;96:106093. doi:10.1016/j.cct.2020.106093

8. Phillips A, Clark T. Estimands in practice: bridging the gap between study objectives and statistical analysis. *Pharm Stat*. 2020;20:68-76. doi:10.1002/pst.2056
9. Hemmings R. Under a black cloud glimpsing a silver lining: comment on statistical issues and recommendations for clinical trials conducted during the COVID-19 pandemic. *Stat Biopharm Res*. 2020;12:414-418. doi:10.1080/19466315.2020.1785931
10. Degtyarev E, Rufibach K, Shentu Y, et al. Assessing the impact of COVID-19 on the clinical trial objective and analysis of oncology clinical trials—application of the Estimand framework. *Stat Biopharm Res*. 2020;12:427-437.
11. Leckman JF, Riddle MA, Hardin MT, et al. The Yale global tic severity scale: initial testing of a clinician-rated scale of tic severity. *J Am Acad Child Adolesc Psychiatry*. 1989;28:566-573.
12. Jakubovski E, Pisarenko A, Fremer C, et al. The CANNA-TICS study protocol: a randomized multi-center double-blind placebo controlled trial to demonstrate the efficacy and safety of nabiximols in the treatment of adults with chronic tic disorders. *Front Psych*. 2020;11:1330.
13. Lasch F, Guizzaro L, Aguirre Dávila L, Müller-Vahl K, Koch A. Potential impact of COVID-19 on ongoing clinical trials: a simulation study with the neurological Yale global tic severity scale based on the CANNA-TICS study. *Pharm Stat*. 2021;20:675-691. doi:10.1002/pst.2100
14. Mataix-Cols D, Ringberg H, Fernández de la Cruz L. Perceived worsening of tics in adult patients with Tourette syndrome after the COVID-19 outbreak. *Mov Disord Clin Pract*. 2020;7:725-726. doi:10.1002/mdc3.13004
15. Grady B, Myers KM, Nelson E-L, et al. American telemedicine association Telemental health S, guidelines working G. evidence-based practice for telemental health. *Telemed J E Health*. 2011;17:131-148. doi:10.1089/tmj.2010.0158
16. Schrag A, Martino D, Apter A, et al. European multicentre tics in children studies (EMTICS): protocol for two cohort studies to assess risk factors for tic onset and exacerbation in children and adolescents. *Eur Child Adolesc Psychiatry*. 2019;28:91-109. doi:10.1007/s00787-018-1190-4
17. Grady BJ, Melcer T. A retrospective evaluation of TeleMental healthcare services for remote military populations. *Telemed J E Health*. 2005;11:551-558. doi:10.1089/tmj.2005.11.551
18. Loh PK, Ramesh P, Maher S, Saligari J, Flicker L, Goldswain P. Can patients with dementia be assessed at a distance? The use of telehealth and standardised assessments. *Intern Med J*. 2004;34:239-242. doi:10.1111/j.1444-0903.2004.00531.x
19. Van Lancker K, Tarima S, Bartlett J, Bauer M, Bharani-Dharan B, Bretz F, Flournoy N, Michiels H, Olarte Parra C, Rosenberger JL, Cro S. Estimands and their estimators for clinical trials impacted by the COVID-19 pandemic: a report from the NISS Ingram Olkin forum series on unplanned clinical trial disruptions. In arxivorg 2022.
20. Richiardi L, Bellocco R, Zugna D. Mediation analysis in epidemiology: methods, interpretation and bias. *Int J Epidemiol*. 2013;42:1511-1519.
21. Loh WW, Moerkerke B, Loeys T, Poppe L, Crombez G, Vansteelandt S. Estimation of controlled direct effects in longitudinal mediation analyses with latent variables in randomized studies. *Multivar Behav Res*. 2020;55:763-785.
22. Vansteelandt S. Estimating direct effects in cohort and case–control studies. *Epidemiology*. 2009;20:851-860.
23. Acharya A, Blackwell M, Sen M. Explaining causal findings without bias: detecting and assessing direct effects. *Am Polit Sci Rev*. 2016;110:512-529.
24. Acharya A, Blackwell M, Sen M, Kuriwaki S, Blackwell MM. *Package 'DirectEffects'*. 2018. https://cran.r-hub.io/web/packages/DirectEffects/DirectEffects.pdf
25. Hernan MA, Robins J. *Causal Inference: What if*. Chapman & Hill/CRC; 2020.
26. Mallinckrodt C, Molenberghs G, Lipkovich I, Ratitch B. *Estimands, Estimators and Sensitivity Analysis in Clinical Trials*. CRC Press; 2019.
27. Little RJA, Rubin DB. *Statistical Analysis with Missing Data*. John Wiley & Sons; 2002.
28. Morris TP, White IR, Royston P. Tuning multiple imputation by predictive mean matching and local residual draws. *BMC Med Res Methodol*. 2014;14:1-13.
29. Buuren S, Groothuis-Oudshoorn K. Mice: multivariate imputation by chained equations in R. *J Stat Softw*. 2010;45:1-68.
30. Schafer JL. Multiple imputation: a primer. *Stat Methods Med Res*. 1999;8:3-15. doi:10.1177/096228029900800102
31. Rubin DB. *Multiple Imputation for Survey Nonresponse*. Wiley; 1987.
32. Eekhout I, van de Wiel MA, Heymans MW. Methods for significance testing of categorical covariates in logistic regression models after multiple imputation: power and applicability analysis. *BMC Med Res Methodol*. 2017;17:129. doi:10.1186/s12874-017-0404-7
33. Gaffert P, Meinfelder F, Bosch V. Towards an MI-proper predictive mean matching. https://www.uni-bambergde/fileadmin/uni/fakultaeten/sowi_lehrstuehle/statistik/Personen/Dateien_Florian/properPMMpdf2016
34. Rubin DB. *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons; 2009.
35. Burgette LF, Reiter JP. Multiple imputation for missing data via sequential regression trees. *Am J Epidemiol*. 2010;172:1070-1076. doi:10.1093/aje/kwq260
36. Doove LL, Van Buuren S, Dusseldorp E. Recursive partitioning for missing data imputation in the presence of interaction effects. *Comput Stat Data Anal*. 2014;72:92-104. doi:10.1016/j.csda.2013.10.025
37. Vansteelandt S, Carpenter J, Kenward MG. Analysis of incomplete data using inverse probability weighting and doubly robust estimators. *Methodol*. 2010;6:37-48.
38. Davison AC, Hinkley DV. *Bootstrap Methods and Their Application*. Cambridge University Press; 1997.
39. Canty A, Ripley B. *boot: Bootstrap R (S-Plus) Functions. R Package Version 1.2–43*. Cambridge University Press; 2019.
40. Morris TP, White IR, Crowther MJ. Using simulation studies to evaluate statistical methods. *Stat Med*. 2019;38:2074-2102.

41. Lasch F, Guizzaro L, Pétavy F, Gallo C. A simulation study on the estimation of the effect in the hypothetical scenario of no use of symptomatic treatment in trials for disease-modifying agents for Alzheimer's disease. *Stat. Biopharm. Res*. 2022. doi:10.1080/19466315.2022.2055633
42. Floden L, Bell ML. Imputation strategies when a continuous outcome is to be dichotomized for responder analysis: a simulation study. *BMC Med Res Methodol*. 2019;19:1-11.

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: Lasch F, Guizzaro L. Estimators for handling COVID-19-related intercurrent events with a hypothetical strategy. *Pharmaceutical Statistics*. 2022;1-23. doi:10.1002/pst.2244