Simulation study report

Title of project	Use of causal inference methods commonly used in non-interventional studies to estimate treatment effects in clinical trials.
Tender ID	SC03 to FWC EMA/2020/46/TDA/L3.02 - ROC36
Consortium	CONFIRMS (CONsortium For Innovation in Regulatory Medical Statistics)
Objectives of the project	(1) To screen the scientific literature and regulatory documents to identify (a) randomised clinical trials in which causal inference methods have been used to test and estimate treatment effects, and; (b) methodological approaches based on causal inference for testing and estimating treatment effects in randomised clinical trials. (2) (a) To identify relevant clinical trial scenarios in which the causal pathway from randomised treatment to outcome may be affected by intercurrent events; (b) To identify causal inference methods suitable for testing and estimating treatment effects within the estimand framework that appropriately account for these intercurrent events. (3) To assess the statistical properties (including bias, coverage probabilities, type 1 error rate, and power) of the selected methods in a comprehensive simulation study based on clinical trial data generated under varying scenarios.
Scientific contact person	Robin Ristl, Medical University of Vienna
Administrative contact person(s)	Karin Brauneis Medical University of Vienna Spitalgasse 23, 1090 Vienna, Austria Tel: +43 1 40400 74880 Email: medstat@meduniwien.ac.at
Date	2026-06-29
Version	1.1

Simulation study report

Authors: Robin Ristl[1], Louise Jespersen[1], Florian Klinglmüller[2], Tobias Fellinger[2], Niels Hendrickx[4], Susanne Urach[2], Paula Starke[3], Angelika Geroldinger[2], Martin Posch[1], Franz König[1], Maddalena Centanni[4], Tim Friede[3], Norbert Benda[3], Andrew Hooker[4], Maike Hohberg[3], Zhe Huang[4], Lucia Schneider[3], Maxi Schulz[3], Mats Karlsson[4].

Affiliations:
[1]Medical University of Vienna, Center for Medical Data Science, Vienna, Austria
[2]Österreichische Agentur für Gesundheit und Ernährungssicherheit GmbH (AGES), Austria
[3]University Medical Center Göttingen, Department of Medical Statistics, Göttingen, Germany
[4]Uppsala University, Department of Pharmacy, Uppsala, Sweden

Contents

1 Introduction

A simulation study in the setting of randomised clinical trials was performed to evaluate the operating characteristics of causal inference methods for estimating specific estimands that address intercurrent events or missing data. The simulation study was pre-specified in the Simulation Study Protocol Version 1.2 (Ristl et al., 2026) that was made public in the HMA-EMA Catalogues of real-world data sources and studies and can be accessed under `https://catalogues.ema.europa.eu/node/4623`.

The simulation study focused on scenarios for two-armed randomised controlled trials with an experimental treatment group and a control group, one primary endpoint, one or more baseline covariates as well as time-dependent covariates, different types of intercurrent events, different estimand strategies and different mechanisms resulting in missing data.

The simulation comprised four scenario classes: 1) scenarios with a time-to-event endpoint, treatment switching as intercurrent event and a hypothetical strategy, 2) scenarios with a continuous endpoint, administration of rescue medication as intercurrent event and treatment policy or hypothetical strategies, 3) preventive vaccine efficacy trial scenarios with a binary endpoint and possible non-adherence to the full (multi dose) vaccination regimen with an estimand aiming at the principal stratum of compliers with the vaccination regime, and 4) scenarios with a safety endpoint in a chronic disease setting, treatment discontinuation as intercurrent event and a while-on-treatment strategy. An overview on the scenarios, estimands and methods is given in Table 1.1.

Table 1.1: Scenario classes considered in the simulation study, defined through the combination of type of endpoint, summary measure, intercurrent event (ICE), estimand strategy, causal inference method. Abbreviations: RPSFTM - rank preserving structural failure time model, IPW - inverse probability weighting, TSE - Two-stage estimation, IV - instrumental variable.

Type of ICE and endpoint	**Summary measure**	**Strategy**	**Methods**
Oncology setting with switching and time to event endpoint	Hazard ratios	Hypothetical	RPSFTM IPW TSE g-computation
Diabetes trial with rescue medication and continuous endpoint	Difference in means	Treatment policy	IPW
		Hypothetical	IPW de-mediation g-computation
Vaccine trial with incomplete vaccination (lack of tolerability) and binary endpoint	Vaccine efficacy	Principal stratum of compliers with vaccination regimen	IV Propensity score
Safety trial with discontinuation of treatment and time to event endpoint	Hazard ratios	While on treatment	Competing risk IPCW

To evaluate the performance of the different analysis methods, we assessed the bias and the root mean squared error (RMSE) of point estimates, the type I error rate and power of hypothesis tests under scenarios corresponding to respective null or alternative hypotheses, as well as the coverage probability and the width of confidence intervals. More information can be found in Section 6.2 in Ristl et al. (2026).

In general, summary measures were calculated as the average of respective results across the simulation runs within each scenario. Hypothesis tests were performed as one-sided tests at a one-sided 2.5% significance level. Confidence intervals were calculated as two-sided intervals with a nominal coverage probability of 95%.

If not stated otherwise, 10,000 simulation runs were performed for each scenario.

This report is organised as follows:

Among Sections 2 to 5, each section is dedicated to one scenario class. Each of these sections contains a description of the assumed trial design, estimand, considered causal inference methods and data generating mechanisms and specifications of the simulation scenarios and deviations from the study protocol. This is followed by a presentation of simulation results, a case study for illustration of an application of the studied methods and a discussion of results.

The report is concluded in Section 6 with an overall summary of the key findings across all scenario classes, including the main observations, implications, and recommendations arising from the simulation results.

Supportive information is contained in the Appendix and in the attached files listed in Appendix E.

The full programming code of the simulation study is available at `https://github.com/CI-RCT-Sim/CI.RCT.Sim`.

The results of this simulation study should be interpreted within the context of the specific estimands and data-generating mechanisms considered. They are not intended to generalise to other settings where the causal structure, intercurrent events and handling strategies, or modelling assumptions differ.

2 Scenario class 1: Treatment switching, hypothetical strategy, time-to-event

In this set of scenarios, we considered the setting of oncological trials with an overall survival endpoint, in which switching from the control treatment to the experimental treatment may occur after disease progression. Examples for this set up include (Camidge et al., 2021; Demetri et al., 2012; Latimer et al., 2016).

In all simulated scenarios studying causal inference methods to account for treatment switching, we considered the following trial design: The study is a 1:1 randomized controlled trial with an experimental treatment group and a control group in an oncology setting. The primary endpoint of the study is overall survival. The trial follow-up length is event driven with a (maximal) recruitment phase of 2 years and a maximal overall study duration of 7 years. We assume that patients are recruited at constant rate over the recruitment period. Event times are censored by administrative censoring at the calendar time of trial termination (thus participants are administratively censored at different follow-up times). In most considered scenarios, we allowed for additional random censoring, as would occur, e.g., due to study withdrawals or patients lost to follow-up.

We assume that disease status is continuously monitored, and patients may experience a progression of disease event. At the time of disease progression, patients in the control group may switch their treatment towards the experimental treatment.

The aim of the analysis is to assess whether patients receiving the experimental treatment have greater survival times compared to patients receiving the control treatment under the hypothetical scenario that treatment switching was not possible. The respective attributes of the estimand thus were:

- Endpoint: Time from randomisation to death from any cause.
- Treatments: Experimental versus control, administered in regular intervals (not further specified in the general simulation set-up).
- Summary measure: Hazard ratio conditional on covariates X and W at baseline.
- Population: Patients diagnosed with the specific type of cancer.
- Intercurrent event: Treatment switching from the control arm to the experimental arm.
- Intercurrent event strategy: Hypothetical strategy, assuming a hypothetical scenario were switching would not occur.

We assume that for every patient the following variables are observed (including the initial treatment group and the (possibly censored) time to death):

- A, Initial treatment group
- Time to death or last follow up
- Time to disease progression or last follow up
- D(t), a time dependent indicator variable indicating whether the patient had switched at or before time t. (For treatment group patients, D(t) is always 0.)
- X, a continuous baseline covariate
- W(t), a time-dependent continuous covariate measured at baseline and subsequently at yearly intervals
- W_{prog}, the value of W at the time of disease progression (also termed secondary baseline), if applicable
- $\mathbb{1}\{W_{prog} > 0\}$: An indicator variable resembling W at the time of progression exceeding a threshold (chosen as 0), which was considered to be particularly informative for the decision to switch in certain scenarios.

In addition, we assume the presence of a further time-dependent continuous covariate L(t), which takes the role of an unobserved confounder in certain scenarios.

Analysis methods: The following causal inference methods were be applied in the simulation

- Rank preserving structural failure time model (RPSFTM) (Robins and Tsiatis, 1991)
- Two-stage estimation (TSE) (Latimer et al., 2017, 2020)
- Inverse probability of censoring weighting (IPCW) (Latimer and Abrams, 2014)
- G-computation (g-formula) (Al Tawil et al., 2024)

As a comparator method we applied a Cox model with survival times censored at the time of switching.

Assumptions
The model assumption of RPSFTM and TSE are correctly specified outcome models, resembling an accelerated failure time model. The impact of violation of this assumption will be investigated by starting from a data generating model that meets the model assumptions and then shifting parameter values to increasingly deviate from the assumption.

IPCW assumes a correct model for the propensity of switching, which is inherently time dependent and may further depend on time-varying covariates, in particular disease progression. The simulation will allow for combinations of data generating mechanism and specifications of the analysis methods such that the assumptions of the methods are matched, as well as for deviations from assumptions. For IPCW, in particular, specifications which include all required covariates will be initially used and deviations will be explored by date generating mechanisms that include unobserved confounders. IPCW further assumes that the probability for switching conditional on the covariates must be strictly less than 1. The characteristics of IPCW under near violation of this assumption will be explored via increasing the dependence of switching on a time dependent covariate threshold.

G-computation assumes correctly specified models for the dependence of time-varying covariates and outcomes on the covariate history. Similar to IPCW, the impact of this assumption will be assessed through scenarios with functional associations deviating from assumed functions and by increasing effects of unobserved covariates.

The comparator Cox model assumes that switching is independent of the counterfactual outcome, conditional on covariates included in the model. In the core simulation, both the hazard for switching and the hazard for death may depend on common covariates. By varying the effect of these covariates, deviations from the assumption will be explored.

2.1 Data generation

In the simulation, we considered a core scenario and further scenarios which deviated from the core scenario in one or more parameter values. The aim of studying a comprehensive set of scenarios was to explore the impact of changes to the mechanisms of progression, switching, death and random censoring as well as to explore settings with different effects of unobserved confounding (relevant for all methdods), near violation of the positivity assumption (for IPW) and violations of the common treatment effect assumption for RPSFTM.

Table 2.1 gives an overview of the included scenarios. The table should be read in conjunction with the following section 2.1 on data generation. It further serves as reference for the scenario names used in the results Section 2.4.

Throughout this section, we use t to denote time with respect to event times and time dependent variables.

The assigned treatment $A \in 0, 1$, with $A = 1$ denoting treatment and $A = 0$ denoting control was sampled from a Bernoulli distribution with equal probability for 0 and 1, corresponding to an allocation ratio of 1:1.

For every patient, X was sampled from a standard normal distribution.

The time-dependent covariate $W(t),\ t \geq 0$ was modelled as step function with jumps at $t = 0, 1, 2, \ldots, k$ years (and constant values between jumps). For the simulation, k was chosen larger than the maximal trial duration of 7 years. Denote the values over the respective time intervals, indexed $j = 0, 1, 2, \ldots, k$, as $W_j^* = W(t) : t \in [j, j+1)$.

Table 2.1: Description of treatment switching scenarios.

Scenario name	Description
Core	Base scenario with effects of observed variables X and W(t)
W - decrease under ctr	W has decreasing mean with time in the control group (declining health state under control)
W - decrease in both groups	W has decreasing mean with time in both groups
L - decrease under ctr, effect on death, progr., switching (unobs. conf.)	L has decreasing mean with time in the control group
L - decrease in both groups, effect on death, progr., switching (unobs. conf.)	L has decreasing mean with time in both groups
time dependent correlation - Toeplitz	Toeplitz correlation instead of exchangeable for W
progression - slow	Median time to progression of 1 year insteaf of 0.5
progresssion - no effect of W	Progression does not depend on W
progression - reduced effect of trt	Progression rates are more similar between groups (hazard ratio of 0.75 fo trt vs. ctr instead of 0.5)
progression - no effect of trt	Progression rates are identical between groups
progression - unobserved confounding with death	Death and progression share an unobserved, time-dependent confounder L(t)
switch - high probability	Approximately 90% switching after progression instead of 50%
switch - no effect of W	Switching does not depend on W
switch - W>0 near separation	Switching is strongly governed by W exceeding 0 at progression
switch - unobserved confounding with death	Death and switching share an unobserved, time-dependent confounder L(t)
death - high rate	Median time to death of 2 year instead of 4
death - no effect of W	Survival does not depend on W
death - unobserved confounding with progr. and switch	Death, progression an switching share an unobserved, time-dependent confounder L(t)
Trt effect post switch - reduced*	The treatment effect after switching is less than under initial administration (conditional hazard ratio 0.75 instead of 0.5)
Trt effect post switch - none	After switching, there is no treatment effect.
random censoring - none	No random censoring
random censoring - effect of X	Random censoring depends on X
random censoring - effect of X and W	Random censoring depends on X and W(t)
random censoring - effect of X, W and L	Random censoring depends on X, W(t) and L(t). In addition, death also depends on L to induce unobserved confounding between death and random censoring.
random censoring - high rate, effect of X, W	High censoring rate (10% per year instead of 2.5%)
random censoring - high rate, control only, effect of X, W	Censoring occurs under control only, at a high rate of approximately 10%.

The vector $W_0^*, \ldots, W_k^*$ was sampled from a multivariate normal distribution with mean $\boldsymbol{\mu}_W$ and covariance matrix $\boldsymbol{\Sigma}_W$.

The time-dependent covariate $L(t),\ t \geq 0$ was modelled as step function and sampled from a multivariate normal distribution in the same way as the observed covariate $W(t)$. The according mean vector and covariance matrix are denoted as $\boldsymbol{\mu}_L$ and covariance matrix $\boldsymbol{\Sigma}_L$.

In the core scenario, $\boldsymbol{\mu}_W$ and $\boldsymbol{\mu}_L$ were vectors of $\mathbf{0}$, corresponding to a stable health state in terms of these variables. The covariance matrices corresponded to the exchangeable covariance structure with all variances equal to 1 and correlation 0.5. In additional scenarios we considered mean vectors $(1, 0.5, 0, -1, \ldots, -1)$ in both treatment groups, which corresponds to a declining health state, as well as mean vector $(0.5, \ldots, 0.5)$ under treatment in combination with $(0.5, 0, -0.5, -1, \ldots, -1)$ under control, which corresponded to a declining state under control only.

We further considered scenarios with Toeplitz covariance structure in which the correlations for two measurements between adjacent time intervals were set to $0.9, 0.8, \ldots, 0.1$ for time differences of $1, 2, \ldots, 10$ years. The variance was set to 1.

Of note, X, W and L were generated as independent random variables.

All time to event variables (time to death, time to progression and time to random censoring) were sampled in a similar fashion: We defined a hazard function that potentially depends on an intercept (Int) and the variables A, X, W(t), L(t), D(t). Then event times were sampled based on the corresponding distribution function. In detail, the hazard functions for event $\in$ {death, censor} were defined as

$$\eta_{event}(t) = \exp(\beta_{event,Int} + X * \beta_{event,X} + W(t) * \beta_{event,W} + L(t) * \beta_{event,L} + A * \beta_{event,A} + D(t) * \beta_{event,D})$$

and the hazard function for progression does not depend on switching, as this occurs after progression. Note that the censoring here refers to the random censoring process that may depend on covariates and occurs in addition to administrative censoring. It was defined similarly otherwise:

$$\eta_{progr}(t) = \exp(\beta_{progr,Int} + X * \beta_{progr,X} + W(t) * \beta_{progr,W} + L(t) * \beta_{progr,L} + A * \beta_{progr,A})$$

Event times were sampled using the inverse cumulative distribution function method: The corresponding cumulative distribution function is $F_{event}(t) = 1 - \exp\left(-\int_0^t \eta_{event}(s)ds\right)$. A random variable $u \sim Uniform(0, 1)$ was sampled. The random time to event then was t_{event} such that $F_{event}(t_{event}) = u$.

Observed times for progression and death were censored at the event time for random censoring, if the latter was smaller, and at due to administrative censoring at the time of study termination.

Parameter values used in the core scenario for the hazard models and the model for switching, which will be presented below, are shown in Table 2.2. For all further scenarios, the parameter values deviating from the core scenario are listed in table 2.3.

Table 2.2: Paramter values in the core scenario for hazard models and the logistic model for switching.

	Intercept	X	W	W>0	L	Treatment (A)	Switched (D)
Progression	log(log(2)/0.5)	log(0.5)	log(0.5)	0	0	log(0.5)	
Death	log(log(2)/4)	log(0.5)	log(0.5)	0	0	log(0.5)	log(0.5)
Random censoring	log(-log(0.975))	0	0	0	0	0	0
Switch	log(0.5/0.5)	log(1.5)	log(1.5)	0			

In general, in settings where the covariates are centered around 0, the median time to event is approximately determined through the intercept as $log(2)/\exp(\beta_{event,Int})$. The intercepts for the core scenario were chosen

such that, by this approximation, the median time to death was 4 year and the median time to progression was 0.5 years. The intercept for random censoring was chosen such that in absence of any covariate effects, where the time to random censoring follows a common exponential distribution, the random censoring rate is 2.5% within one year.

For control group patients for whom disease progression occurred, the decision to switch was sampled from a Bernoulli distribution with probability p_{switch} depending on the covariates via the following logistic model:

$$\frac{p_{switch}}{1-p_{switch}} = \exp(\beta_{switch,Int} + X * \beta_{switch,X} + W_{prog} * \beta_{switch,W} + \mathbb{1}\{W_{prog} > 0\} * \beta_{switch,W>0}$$

In the core scenario, $\beta_{switch,Int}$ was set to 0, which results in approximately 50% of control group patients to switch after progression. In a further scenario, $beta_{switch,Int} = \log(9)$ was considered to increase the switching probability to approximately 90%. Furthermore, while in the core scenario $\beta_{switch,W>0} = 0$, in an additional scenario aimed to explore data close to the violation of the positivity assumption for the inverse probability weighting method, $\beta_{switch,W>0} = \log(0.9/0.1)/\sqrt{2/\pi} - \log(1.5) = 2.348$ was considered. (Note that $\sqrt{2/\pi}$ is the expected value of a standard normal variable conditional on being larger than 0 and so, together with the other parameter values $\beta_{switch,Int} = 0$ and $\beta_{switch,W} = \log(1.5)$, for an average patient with $W_i(t) > 0$, the probability to switch will be 90%.)

To summarise the data generating mechanisms, Figure 2.1 visualises a simplified version of the assumed structure and dependencies among the variables for this scenario.

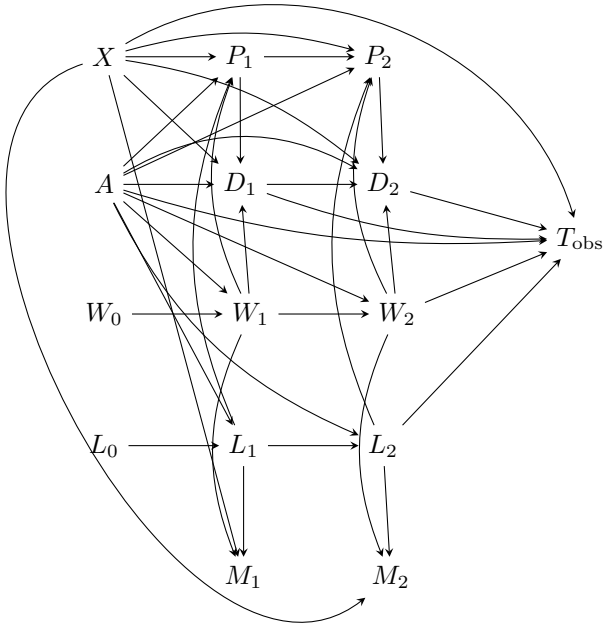


Figure 2.1: Directed acyclic graph (DAG) illustrating the assumed causal structure of the simulated longitudinal data, with baseline covariates (X), randomly assigned treatment (A), measured time varying covariates (W_j), unmeasured time varying covariates (L_j), indicator of censoring other than administrative censoring (M_j), indicator of progression (P_j) and the intercurrent event treatment switching (D_j) at visit j as well as the observed time to event (T_{obs}).

Table 2.3: Parameter values in all treatment switching simulation scenarios. The table shows the parameter values that deviate from the values in the core scenario. Also see text for further details. In general, all scenarios were investigated under small and under large sample sizes as well as under the alternative and the null hypotheses. The scenarios "Trt effect post switch - reduced" and "Trt effect post switch - none" were not included under the null hypothesis as in that case the treatemtn effect was zero in all scnearios. Furthermorer, the scenario "Trt effect post switch - reduced" was investigated only for small sample sizes as it includes a reduction of the treatment effect towards the effect size studied in the large sample setting. Note that the scenario "W - decrease under ctr" contained an effect of W on the hazard for death under the alternative, while this effect was set to 0 under the null hypothesis.

Scenario Name	Difference from core scenario
W - decrease under ctr	Trt: $\boldsymbol{\mu}_W = (0.5, \ldots, 0.5)$, Ctr: $\boldsymbol{\mu}_W = (0.5, 0, -0.5, -1, \ldots, -1)$
W - decrease in both groups	$\boldsymbol{\mu}_W = (1, 0.5, 0, -0.5, -1, \ldots, -1)$
L - decrease under ctr, effect on death, progr., switching (unobs. conf.)	Trt: $\boldsymbol{\mu}_L = (0.5, \ldots, 0.5)$, Ctr: $\boldsymbol{\mu}_L = (0.5, 0, -0.5, -1, \ldots, -1)$, $\beta_{prog,L} = \log(0.5)$, $\beta_{switch,L} = \log(1.5)$, $\beta_{death,L} = \log(0.5)$
L - decrease in both groups, effect on death, progr., switching (unobs. conf.)	$\boldsymbol{\mu}_L = (1, 0.5, 0, -0.5, -1, \ldots, -1)$, $\beta_{prog,L} = log(0.5)$, $\beta_{switch,L} = log(1.5)$, $\beta_{death,L} = log(0.5)$
time dependent correlation - Toeplitz	Toeplitz$(0.9, 0.8, \ldots, 0)$ covariance matrix for W
progression - slow	$\beta_{prog,Int} = \log(\log(2)/1)$
progresssion - no effect of W	$\beta_{prog,W} = 0$
progression - reduced effect of trt	$\beta_{prog,trt} = log(0.75)$
progression - no effect of trt	$\beta_{prog,trt} = 0$
progression - unobserved confounding with death	$\beta_{prog,L} = \log(0.5)$, $\beta_{death,L} = \log(0.5)$
switch - high probability	$\beta_{switch,Int} = \log(9)$
switch - no effect of W	$\beta_{switch,W} = 0$
switch - W>0 near separation	$\beta_{switch,Wgrw} = \log(0.9/0.1) * \sqrt{\pi/2} - \log(1.5)$
switch - unobserved confounding with death	$\beta_{switch,L} = \log(1.5)$, $\beta_{death,L} = \log(0.5)$
death - high rate	$\beta_{death,Int} = \log(log(2)/2)$
death - no effect of W	$\beta_{death,W} = 0$
death - unobserved confounding with progr. and switch	$\beta_{prog,L} = \log(0.5)$, $\beta_{switch,L} = \log(1.5)$, $\beta_{death,L} = \log(0.5)$
Trt effect post switch - reduced	$\beta_{death,switched} = \log(0.75)$
Trt effect post switch - none	$\beta_{death,switched} = 0$
random censoring - none	$\beta_{cens,Int} = -\infty$
random censoring - effect of X	$\beta_{cens,X} = \log(0.5)$
random censoring - effect of X and W	$\beta_{cens,X} = \log(0.5)$, $\beta_{cens,W} = \log(0.5)$
random censoring - effect of X, W and L	$\beta_{death,L} = \log(0.5)$, $\beta_{cens,X} = \log(0.5)$, $\beta_{cens,W} = \log(0.5)$, $\beta_{cens,L} = \log(0.5)$
random censoring - high rate, effect of X, W	$\beta_{cens,Int} = \log(-\log(0.9))$, $\beta_{cens,X} = \log(0.5)$, $\beta_{cens,W} = \log(0.5)$
random censoring - high rate, control only, effect of X, W	$\beta_{cens,Int} = \log(-\log(0.9))$, $\beta_{cens,X} = \log(0.5)$, $\beta_{cens,W} = \log(0.5)$, random censoring only under control

2.1.1 Sample size

The number of events e_{stop} required to stop the study was determined by Schoenfeld's formula (Schoenfeld, 1981) as $e_{stop} = ((z_{1-\alpha/2} + z_{1-\beta})/\log(HR_{assumed}))^2 * 4$. Here z_γ is the γ quantile of the standard normal distribution. For all simulations, the nominal two-sided significance level was $\alpha = 0.05$, the aimed for power was set to $1 - \beta = 0.8$.

We considered scenarios under the alternative where the true hazard ratio (see Section 2.1.2 was close to 0.5 or 0.75. For these scenarios, we considered $HR_{assumed} = 0.5$ or 0.75, respectively, such that the actual power would be in a realistic range of 80%. The number of events in these studies accordingly were 66 or 380. The yearly recruitment rate over the two year recruitment period was chosen to be 1.5 times the aimed for number of events. The actual number of eligible patients in a single simulation run was be sampled from a Poisson distribution with mean accordingly corresponding to 198 or 1140. In general, due to the set event rates, the study duration exceeded the recruitment phase such that this maximum number of patients was included in the sampled data set.

2.1.2 True treatment effect

The true treatment effect was determined through the data generating model in terms of a hazard ratio conditional on time-dependent covariates. The summary measure defined in the estimand was the hazard ratio conditional on covariates X and W at baseline. Since the data generating mechanism also involves unobserved covariates and time-dependent covariates, this summary measure does in general not correspond to a single parameter value in the data generating model and it will in general not meet the proportional hazards assumption. Under these circumstances, the Cox model (conditional) hazard ratio can be understood as a parameter, which minimizes a weighted difference between the hazard functions of the two groups. For illustration, consider a case without covariates to condition on. The true Cox model hazard ratio then is β such that $\int_0^\infty Y_0(t)Y_1(t)/(Y_0(t)+Y_1(t)\exp(\beta))(\lambda_1(t)-\exp(\beta)\lambda_0(t))dt = 0$, where $Y_0(t)$ and $Y_1(t)$ are the at-risk functions - probability to still be at risk at time t - in the control and the treatment group, $\lambda_0(t)$ and $\lambda_1(t)$ are the marginal hazard functions and p is the allocation probability for the treatment group. Of note, this value depends on the censoring distribution which in turn is in part determined through trial design.

In the simulation we considered as true treatment effect the expected value of the Cox model log hazard ratio when calculated from an (ideally infinitely) large population under the hypothetical scenario that no treatment switching is possible.

For the main simulations, the true treatment effect was calculated with censoring. This was done to allow for an isolated assessment of the properties of the causal inference methods targeting the hypothetical estimand with respect to treatment switching. For the simulations with additional inverse probability of censoring weights, the true treatment effect was calculated from data with complete follow-up (no random censoring), because here the aim was to see if the methods could properly account for potential bias due to censoring.

The determined the true value for the log hazard ratio, for each considered scenario, we sampled 40 independent data sets with the required number of events set to 10,000 (and the recruitment rate set accordingly to 15,000 per year). The true value was then defined as the average of the 40 large-sample log hazard ratio estimates. The obtained true values are included in the summary tables 2.5 to 2.10.

2.1.3 Summary metrics

For all scenarios, 10,000 simulation runs were performed. To assess bias, we calculated the relative bias of the hazard ratio as $\exp(\widehat{logHR} - logHR)$ where $\widehat{logHR}$ is the log hazard ratio estimated by an analysis method and $logHR$ is the value for the true log hazard. The relative bias was preferred over the absolute bias at the hazard ratio scale as it allow an interpretation independent of the actual value of the hazard ratio and better comparison across different scenarios. To include a measure of uncertainty for the reported bias, we calculated the standard error of the bias $\widehat{logHR} - logHR$ at the log hazard ratio scale as $SE_{bias} = \sqrt{SE^2_{\widehat{logHR}} + SE^2_{logHR}}$, where SE denotes the respective simulation standard error. Note that $logHR$, even though obtained through a large number of 400,000 events in total has some residual variability that is here taken into account. The standard

error of the relative bias is obtained by the delta method as $\exp(\widehat{logHR} - logHR) * SE_{bias}$. Furthermore, we assessed the root mean squared error (RMSE) of the log hazard ratio, the rejection rate (type I error rate and power) of one-sided hypothesis tests under scenarios corresponding to the null or alternative hypotheses, at a one-sided significance level of 2.5%, as well as the coverage probability and the width of two sided nominal 95% confidence intervals at the log hazard ratio scale.

2.2 Analysis methods

2.2.1 Rank preserving structural failure time model (RPSFTM)

RPSFTM assumes an accelerated failure time model such that for the counterfactual time in the study, assuming a patient was "off-treatment" (meaning under control) throughout, is defined for all patients as

$$U_i(\psi) = T_{\text{off},i} + T_{\text{on},i} \exp(\psi),$$

where $T_{\text{off},i}$ is the observed time "off" experimental treatment (and thus "on" control treatment) for participant i, $T_{\text{on},i}$ is the observed time on treatment, and ψ is the acceleration parameter of interest Robins and Tsiatis (1991); White et al. (1999). ψ is estimated via a g-estimation approach, such that the resulting times $\hat{U}_i(\psi)$ show no association (p-value=1) with the randomised treatment group, conditional on the covariates X and W, in an appropriate hypothesis test. In the simulation, the function rpsftm in the R package trtswitch was used and as test for independence a Cox model adjusted for X and W at baseline was used.

A key assumption of the RPSFTM is that the effect of treatment in terms of the acceleration factor ψ is constant, regardless of when the treatment is started and of the time under treatment. In the data generating models, violation of this assumption was be explored by including scenarios labeled "Trt effect post switch - reduced" and "Trt effect post switch - none", where the treatment effect after switching is weaker than the initial treatment effect or absent at all.

The conditional hazard ratio HR_{RPSFTM} was then estimated by a Cox model using the $\hat{U}_i(\psi)$ for the control group and observed event times for the treatment group. The Cox model included treatment group and X and W at baseline as covariates.

The calculation of $\hat{U}_i(\psi)$ may introduce informative censoring and a recensoring approach has been proposed (White et al., 1999). The RPSFTM approach was used with and without recensoring.

Confidence intervals for the hazard ratio were calculated using the boostrap approach implemented in the rpsftm function from the trtswitch package with 100 bootstrap iterations. In brief, the hazard ratio is estiamted from each bootstrap sample, the empirical standard deviation is calculated and used as estimate for the standard deviation in Wald tests and Wald confidence intervals.

2.2.2 Simple Two-stage Estimation (TSE)

In the simple TSE (Latimer et al. (2017)), the time point of disease progression is considered as secondary baseline. In a first step, the effect of switching to active treatment versus staying on the control is estimated using data from control group patients who experienced disease progression: Survival times of these patients, starting from the secondary baseline, are modelled by a parametric accelerated failure time (AFT) model that includes baseline covariates, time dependent covariates with their respective values at the time of progression (secondary baseline) and the switching indicator. The acceleration parameter for switching to active treatment is estimated from this model, and counterfactual survival times for those who switched are calculated analogous to the RPSFTM. A Cox model is then fitted using counterfactual survival times for switchers.

This approach is applicable in particular when treatment switching occurs within a small time interval from progression, as is the case in the simulation scenarios. The presence of unobserved confounders affecting the decision to switch and the survival time may have an impact on the calculation of the counterfactual survival times and this was be explored in the scenarios

- L - decrease under ctr, effect on death, progr., switching (unobs. conf.)

- L - decrease in both groups, effect on death, progr., switching (unobs. conf.)
- switch - unobserved confounding with death
- death - unobserved confounding with progr. and switch

that included effects of the unobserved confounder L on the times to progression, switching and death.

A Cox model is then fitted using counterfactual survival times for switchers.

The TSE was implemented using the tsesimp function from the trtswitch package in R, using a Weibull AFT model and X and W at the time of progression as baseline covariates. The outcome model (Cox model) include treatment group and X and W at baseline as covariates in order to estimate the conditional hazard ratio. The TSE was be fitted, both, without recensoring and with recensoring for administrative censoring. Inference was based on the bootstrap approach implemented in the tsesimp function, which is similar as described for RPSFTM, using 100 bootstrap samples.

2.2.3 Inverse probability of censoring weighting (IPCW)

The aim of the IPCW is to construct a pseudo population, corresponding to a population in which no patient switched. In the model fitting process, survival times for patients who switched are censored at the time of switching, and the remaining participants are weighted with the inverse of the probability of staying on treatment without switching Robins and Finkelstein (2000); Al Tawil et al. (2024). A key assumption for this approach is that there are no unobserved confounders. Violations of this assumption were explored in scenarios that include unobserved confounders:

- L - decrease under ctr, effect on death, progr., switching (unobs. conf.)
- L - decrease in both groups, effect on death, progr., switching (unobs. conf.)
- switch - unobserved confounding with death
- death - unobserved confounding with progr. and switch

In the simulation scenarios, treatment switching may only occur at the time of progression and only from the control arm to the treatment arm. In the analysis, a logistic regression model for the probability to switch was estimated using data from the control group patients who experienced a disease progression. The model included as predictors the baseline covariate X and the time-dependent covariate W with the value observed at the time of progression, i.e. W_{prog}. The function glm in R was be used for this step. For patients who did not switch, the model based probability to switch, $\hat{\pi}_{sw,i}$ was be calculated, as well as according inverse probability weights. To estimate the treatment effect, a Cox model for the primary endpoint was be fit that included X and W at baseline as covariates and furthermore included time dependent weights: All patients in the treatment group received a weight of 1. For patients in the control group, weights were 1 until the time of progression. After progression, patients who did not switch received the weight $w_i = 1/(1 - \hat{\pi}_{sw,i})$. These weights remained constant across further time, as switching was no longer possible post progression in the considered scenarios. For patients who switched, survival times were considered as censored at the time of progression. The function coxph from the R package survival was applied for this step. Time dependent weights were implemented by structuring the data in the time-dependent format including start and stop times for intervals prior to progression and post progression and according weights.

Inference was be based on robust standard errors for the weighted Cox model.

2.2.4 G-computation (parametric g-formula)

For g-computation, a set of regression models is fit that explains the distribution of the values of time-dependent covariates, switching status and event status in discretised time through the previously observed values for covariates, progression status and switching status Robins (1986); McGrath et al. (2020). The fitted models are used to sample a large number of covariate and event time trajectories with starting values given through randomised treatment and baseline covariates, albeit after modifying the models such that switching is not

allowed. The sampled data is used to estimate the hazard ratio. Bootstrap is used to calculate the standard error of this estimate and confidence intervals.

The g-computation approach was implemented by reformatting the data in a long format across discrete time intervals and the using standard regression model functions lm and glm in R. This implementation followed the implementation of the parametric g-formula in the function gformula_survival in the R package gfoRmula. The latter was not used directly, though, as its computations were highly time consuming precluded its application in the large scale simulation study. Of note, the method is inherently computation intense due the need to sample a large amount of data from the fitted models and then to repeat this process several times in order to obtain bootstrap based inference. Also the custom implementation had computation times that by far exceed those of the other studied methods. The g-computation was therefore only included for the small sample scenarios.

In detail, the data for each patient was first structured into discrete time intervals with length 1/12 year. The data was then brought to a long format across these discrete intervals. For each interval the long data set included the value for X the (last observed) value for the time dependent variable W as well as W lagged by one time interval. Furthermore, the long data set included an indicator variable ev for death, with a value of 1 in the patient's last time interval if death had been observed and a value of 0 otherwise. A further indicator variable was included for switching, with a value of 1 for the interval in which switching occurred and all subsequent intervals, and 0 before switching occurred.

Based on this data, the following linear and logistic regression models were fit to model the trajectory of W and the occurrence of death separately for each treatment group. Here, D is the long data set.

```
mod_death<-glm(ev~X+W+time+switch,family=binomial,data=D,weights=w,subset=trt==trt_group)
mod_W<-lm(W~X+W_lag+time,data=D,weights=w,subset=trt==trt_group)
```

Note that the models include weights statements. Weights were used in additional simulations where the aim was to counterbalance random censoring. See Section 2.2.6 for details.

Based on these models and the observed values of X and W at baseline, trajectories of W and death were sampled. In detail, sampled date was built up stepwise over the discrete time intervals, starting at baseline. In each step, the next value for W was sampled from normal distribution with mean implied by X, W and lagged W through the fitted model for W, the and variance corresponding to the residual variance of the model. Next, the potential occurrence of death was sampled from Bernoulli distributions with probability implied by the logistic model for death, taking into account the previously sampled values for X at baseline and W for the current interval. Importantly, while switching state was included when fitting the model for death, the switching covariate was not used when sampling data under the hypothetical assumption that no switching was possible. This is a key step of the g-formula approach.

Once a death event was sampled, the iterative process was stopped. Otherwise the sampling continued until the individual possible maximum follow-up time point for the patient. The maximum follow-up depended on the calendar time of recruitment (which was included when simulating the original data) and the calendar time of study termination. This implementation of the g-formula thus takes into account administrative censoring, which was considered suitable to address the fact the the assumed true value for the hazard ratio may depend on the study design through administrative censoring.

For every patient, $m = 10$ such trajectories were sampled.

Finally, the sampled long format data was back transformed to one line per trajectory (i.e. m lines per patient with information of the observed baseline covariates X and W, the initial treatment allocation and the g-formula-sampled time to death. This data was used to fit a Cox model and estimate the log-hazard ratio for treatment versus control adjusted for X and W at baseline.

The whole process was subject to bootstrapping: Starting with the initial long data set for n patients, n patients were sampled with replacement and the subsequent model fitting, sampling of trajectories, back transformation and estimation of the log hazard ratio were repeated for B such bootstrap sample. The bootstrap standard error of the log hazard ratio, SE_{BS} was calculated as empirical standard deviation across the hazard ratios obtained through bootstrapping.

In the simulation, bootstrap was performed with $B = 20$ samples. The number was chosen relatively small in order to allow for feasibly computation times and including the g-formula approach in the simulation of 52 scenarios with 10,000 simulation runs each. Opposed to the calculation of quantile based bootstrap confidence interval, we considered the bootstrap standard deviation to be sufficiently stably estimable also from a small number of samples.

Finally, a one sided p-value and a two sided 95% nominal confidence interval (CI) were based on log the log hazard ratio (loghr) estimate and t-distribution with $B - 1$ degrees of freedom as illustrated in the below code:

```
    p <- pt(loghr / SE_BS,df=B-1)
   CI <- exp(loghr + c(-1, 1) * SE_BS * qt(0.975,df=B-1))
```

2.2.5 Cox proportional hazards model as comparator method

A Cox model was be fit for overall survival, in which event times were censored at the time of switching. The model included as predictor the initial treatment allocation and the observed covariates X and W at baseline.

2.2.6 Methods to account for non-administrative censoring

The considered data generating mechanisms included scenarios with random censoring (in addition to administrative censoring), that was either independent of covariates, dependent on the baseline covariate X or dependent on both baseline and time dependent covariates, see Table 2.3.

The analysis models for all considered methods included baseline covariates as predictor variables. Under these models, censoring is assumed to be at random conditional on the utilized baseline information such that we expect no substantial bias in scenarios with completely random censoring or censoring dependent on baseline covariates.

In the scenario with censoring dependent on time-dependent covariates, adjusting for baseline covariates may, however, not be a sufficient adjustment. We therefore included an additional analysis approach that utilizes time-dependent inverse probability of censoring weights as follows:

A Cox model was be fit for the time to censoring (other than administrative), including treatment group and X as baseline covariates and $W(t)$ as time-dependent covariate. Death and administrative censoring were treated as censoring events in this model. Based on this model, cumulative distribution function for time to censoring given the covariate history were estimated for each patient, using the function `survfit.coxph` in the R package survival. Next, for each patient and each observed event time t, the probability $\pi_i(t)$ that the time to random censoring is larger than t was calculated. Finally, $1/\pi_i(t)$ was applied as time-dependent inverse probability weight in the outcome Cox models for the RPSFTM, TSE and IPW methods described above. Note that for IPW, the weights accounting for random censoring are used in addition to the weights that account for treatment switching and the final weights are the product of both weights. RPSFTM and TSE were applied with recensoring in the simulations where the additional inverse probability of censoring weights were used.

For the g-computation, the inverse probability weights for random censoring were computed based on the discrete long format. The probability to stay free of random censoring locally within each interval was calculated from a logistic regression model with treatment (trt), X, $W(t)$ and time t as well as all treatment by covariate interactions as predictors. The probability to stay free of random censoring up to a given time interval was then calculated as product of the local probabilities up to the given time interval. The weights were then the inverse of this cumulative probabilities:

```
    mod_cens<-glm(!random_cens_event~trt*(X+W+time)+switch,data=D,family=binomial)
    PR<-predict(mod_cens,type="response")
    Ag<-aggregate(PR~D$id,FUN=cumprod)
    cumpr<-unlist(Ag[[2]])
    w<-1/cumpr
```

Weights larger than 5 were truncated to a value of 5 to avoid overly influential points. The weights were included in the respective models for time dependent W, the occurrence of progression and the occurrence of death. (A similar approach is available in the gfoRmula R package, which was initially considered for this method.)

2.3 Deviations from the study protocol

The implementation of the simulation study for the treatment switching scenarios contained the following deviations from the study protocol.

The yearly recruitment rate was increased in all scenarios from 1 time the aimed for number of events towards 1.5 times the aimed for number of events. This deviation was introduced because the initially planned recruitment rate too frequently resulted in studies that did not reach the number of events within the set maximum duration of seven years.

In the scenario "W - decrease under ctr", the mean vectors for W(t) were changed from $(0, \ldots, 0)$ under treatment and $(1, 0.5, 0, -1, \ldots, -1)$ under control towards $(0.5, \ldots, 0.5)$ under treatment and $(0.5, 0, -0.5, -1, \ldots, -1)$ under control. This change was introduced to ensure that both groups have the same expected value for W at baseline, as should be the case under randomized treatment allocation.

The true hazard ratio was calculated using 40 samples with 10,000 events each. The initial plan was to use one sample with 100,000. This effective increase of the number of events was performed to improve the precision of the simulated true value.

The confidence intervals for RPSFTM were calculated using bootstrap, while the protocol said to use p-value inversion. However, the latter was found to produce inaccurate results and therefore was replaced by bootstrapping.

The g-formula method was performed using a used custom implementation instead of the gfoRmula R package. This change was made because the methods in the gfoRmula package required long and thus limiting computation times. However, also the custom implementation was computation intense and in particular in conjunction with the required bootstrap inference was considerably more time consuming than the other studied methods. It was only feasibly to include this method with a small number of bootstrap samples of 20 and in the small sample scenarios. However, the number of trajectories sampled for every patient was increased from one trajectory towards 10 trajectories. This change was introduced because, with small sample sizes, too few trajectories resulted in considerably variability when the method was repeatedly applied to the same data set.

Scenarios random "censoring - high rate, effect of X, W" and "random censoring - high rate, control only, effect of X, W" were not prespecified in the protocol but added to explore the impact of larger amount of censoring and in particular of differential censoring patterns between the study groups.

2.4 Results

This section contains the results of simulations for the treatment switching scenarios. The first subsection 2.4.1 show summarising statistics of the simulated data. In the following sections, results on bias, RSMSE, type I error rate, power, and confidence interval coverage and width are presented. Throughout these Sections, the operating characteristics will be presented figures, where each figure corresponds to one of six scenario blocks. These blocks are: Small sample settings under the alternative hypothesis (labelled H1) and under the null hypothesis (labelled H0), large sample settings under the alternative hypothesis and under the null hypothesis, as well as additional simulations for IPCW methods addressing random censoring under the alternative and the null hypothesis.

In all result plots presented for the treatment switching scenarios, the assessed methods were labelled with abbreviations according to Table 2.4.

In addition to the plots presented below, numeric values of all results can be found in the supplementary files trt_switch_small_n_H1_results.xlsx, trt_switch_small_n_H0_results.xlsx, trt_switch_large_n_H1_results.xlsx, trt_switch_large_n_H0_results.xlsx, trt_switch_extra_cens_H1_results.xlsx and trt_switch_extra_cens_H0_results.xlsx.

2.4.1 Descriptive summaries of simulated trial data

Tables 2.5 to 2.10 provide summary statistics over the 10,000 simulation runs for each scenario. The tables show that in the vast majority of simulated trials the intended number of events was achieved within the maximum trial duration of seven years. The single exception is the scenario "random censoring - high rate, effect of X,

Table 2.4: Abbreviation of considered analysis methods for the estimation of a hypothetical estimand in the treatment switching scenarios used in result plots.

Abbreviation	Method
cens	Cox model with censored event times at switching
ipw	Inverse probability weights for switching applied to a Cox model with censored event times at switching
rpsftm	Rank-preserving structural failure time model
rpsftm_rc	Rank-preserving structural failure time model with recensoring
tse	Simplified two stage estimation
tse_rc	Simplified two stage estimation with recensoring
gformula	Parametric g-formula
ipw_IPCW	Application of inverse probability weights for switching and additional inverse probability weights for censoring in a Cox model with censored event times at switching
rpsftm_rc_IPCW	Application of addtional inverse probability weights for censoring to RPSFTM with recensoring
tse_rc_IPCW	Application of addtional inverse probability weights for censoring to TSE with recensoring
gformula_IPCW	Application of addtional inverse probability weights for censoring to the g-formula

W". In this scenario, due to the high amount of censoring, in approximately 22% of simulation runs the number of events was lower than the intended number of 66, albeit the average number of events was still close to the target value.

Table 2.5: Descriptive statistics for the data simulated under scenarios with small sample size under the alternative hypothesis. Statistics are mean values for the number of patients under treatment and control (trt, ctr), the number of events in either group (ev trt, ev ctr), the number of control group patients who switched (switch), the number of patients for whom the time do death was censored by a random censoring event (rcens), the number of simulated data sets out of 10,000 in which the aimed for number of events was not achieved within the maximum study duration (ev<aim), the average maximum follow up time in years (FU) and the true hazard ratio conditional on the covariates X and W at baseline (HR). For each scenario, 10,000 data sets were simulated.

Scenario	trt	ctr	ev trt	ev ctr	switch	rcens	ev<aim	FU	HR
Core	99.0	98.9	27.7	38.3	38.3	3.8	4	4.0	0.50
W - decrease under ctr	99.1	99.1	23.2	42.8	41.3	3.9	3	4.3	0.36
W - decrease in both groups	99.1	99.0	28.3	37.7	42.9	3.9	0	4.4	0.51
L - decrease under ctr, effect on death, progr., switching (unobs. conf.)	98.9	99.2	24.3	41.7	38.9	3.7	0	4.0	0.41
L - decrease in both groups, effect on death, progr., switching (unobs. conf.)	99.0	98.8	28.5	37.5	40.3	3.7	0	4.2	0.54
time dependent correlation - Toeplitz	99.0	99.1	27.8	38.2	38.5	3.9	11	4.1	0.50
progression - slow	99.0	99.1	27.2	38.8	30.2	3.8	0	3.9	0.50
progresssion - no effect of W	99.1	99.0	27.7	38.3	39.9	3.9	1	4.0	0.50
progression - reduced effect of trt	99.0	99.0	27.8	38.2	38.5	3.8	0	4.0	0.50
progression - no effect of trt	99.2	98.9	27.7	38.3	38.3	3.9	3	4.0	0.51
progression - unobserved confounding with death	99.1	99.1	28.1	37.9	37.2	3.6	0	3.6	0.54
switch - high probability	99.1	99.0	30.3	35.6	72.1	4.2	23	4.4	0.50
switch - no effect of W	99.0	98.9	27.9	38.1	39.7	3.9	5	4.0	0.51
switch - W>0 near separation	99.0	99.1	28.6	37.4	51.7	4.0	6	4.1	0.50
switch - unobserved confounding with death	99.0	98.9	28.0	38.0	37.3	3.6	0	3.6	0.54
death - high rate	99.1	99.1	27.5	38.5	29.8	2.5	0	2.5	0.50
death - no effect of W	99.0	99.0	27.4	38.6	38.8	4.2	15	4.5	0.50
death - unobserved confound ing with progr. and switch	99.0	98.9	28.0	38.0	35.9	3.6	0	3.6	0.53
Trt effect post switch - reduced	99.1	99.0	26.4	39.6	38.0	3.7	1	3.8	0.51
Trt effect post switch - none	99.0	99.1	25.4	40.6	37.6	3.6	0	3.6	0.50
random censoring - none	98.9	99.0	27.7	38.3	38.7	0.0	0	3.8	0.50
random censoring - effect of X	99.1	99.0	27.7	38.3	38.5	4.1	4	4.1	0.50
random censoring - effect of X and W	99.1	99.0	27.8	38.2	38.8	4.4	14	4.2	0.51
random censoring - effect of X, W and L	99.1	99.2	27.9	38.1	37.4	4.3	5	3.8	0.53
random censoring - high rate, effect of X, W	99.1	99.1	27.2	37.7	38.8	10.0	2178	5.7	0.51
random censoring - high rate, control only, effect of X, W	99.1	98.9	32.4	33.5	37.7	5.6	67	4.5	0.51

Table 2.6: Descriptive statistics for the data simulated under scenarios with small sample size under the null hypothesis. Statistics are mean values for the number of patients under treatment and control (trt, ctr), the number of events in either group (ev trt, ev ctr), the number of control group patients who switched (switch), the number of patients for whom the time do death was censored by a random censoring event (rcens), the number of simulated data sets out of 10,000 in which the aimed for number of events was not achieved within the maximum study duration (ev<aim), the average maximum follow up time in years (FU) and the true hazard ratio conditional on the covariates X and W at baseline (HR). For each scenario, 10,000 data sets were simulated.

Scenario	trt	ctr	ev trt	ev ctr	switch	rcens	ev<aim	FU	HR
Core	99.2	99.0	33.0	33.0	34.9	2.9	0	2.9	1.00
W - decrease under ctr	99.0	99.2	32.9	33.1	36.4	3.2	0	3.2	1.00
W - decrease in both groups	99.0	98.8	33.1	32.9	39.9	3.4	0	3.6	1.00
L - decrease under ctr, effect on death, progr., switching (unobs. conf.)	99.1	99.0	33.0	33.0	33.2	2.9	0	2.9	1.00
L - decrease in both groups, effect on death, progr., switching (unobs. conf.)	99.2	98.9	33.0	33.0	37.1	3.2	0	3.4	1.00
time dependent correlation - Toeplitz	99.0	99.1	33.0	33.0	35.0	3.0	0	3.0	1.00
progression - slow	99.1	98.9	33.0	33.0	26.1	2.9	0	2.9	1.00
progresssion - no effect of W	99.0	99.2	32.9	33.1	36.8	2.9	0	2.9	1.00
progression - reduced effect of trt	98.9	98.9	33.0	33.0	35.0	2.9	0	2.9	1.00
progression - no effect of trt	98.9	98.9	33.0	33.0	34.9	3.0	0	2.9	1.00
progression - unobserved confounding with death	98.9	99.0	33.0	33.0	33.4	2.8	0	2.7	1.00
switch - high probability	98.9	99.2	32.9	33.1	65.5	3.0	0	2.9	1.00
switch - no effect of W	99.0	99.0	32.9	33.1	36.2	3.0	0	2.9	1.00
switch - W>0 near separation	99.0	99.2	33.0	33.0	46.9	3.0	0	2.9	1.00
switch - unobserved confounding with death	99.1	98.9	33.1	32.9	33.6	2.8	0	2.7	1.00
death - high rate	96.7	97.0	32.9	33.1	25.2	2.1	0	2.0	1.00
death - no effect of W	99.1	99.0	33.1	32.9	36.0	3.2	0	3.2	1.00
death - unobserved confound ing with progr. and switch	99.1	98.9	33.0	33.0	32.2	2.8	0	2.7	1.00
random censoring - none	98.9	98.9	33.0	33.0	35.0	0.0	0	2.9	1.00
random censoring - effect of X	99.0	98.9	33.0	33.0	35.1	3.1	0	3.0	1.00
random censoring - effect of X and W	99.0	99.1	33.1	32.9	35.4	3.3	0	3.0	1.00
random censoring - effect of X, W and L	99.1	99.0	33.0	33.0	33.8	3.4	0	2.8	1.00
random censoring - high rate, effect of X, W	99.0	99.0	33.0	33.0	35.9	8.1	5	3.6	1.00
random censoring - high rate, control only, effect of X, W	99.0	98.9	36.2	29.8	34.5	4.7	0	3.1	1.00

Table 2.7: Descriptive statistics for the data simulated under scenarios with large sample size under the alternative hypothesis. Statistics are mean values for the number of patients under tretament and control (trt, ctr), the number of events in either group (ev trt, ev ctr), the number of control group patients who switched (switch), the number of patients for whom the time do death was censored by a random censoring event (rcens), the number of simulated data sets out of 10,000 in which the aimed for number of events was not achieved within the maximum study duration (ev<aim), the average maximum follow up time in years (FU) and the true hazard ratio conditional on the covariates X and W at baseline (HR). For each scenario, 10,000 data sets were simulated.

Scenario	trt	ctr	ev trt	ev ctr	switch	rcens	ev<aim	FU	HR
Core	569.9	570.4	177.1	202.9	211.5	7.9	0	3.3	0.75
W - decrease under ctr	569.7	570.4	156.1	223.9	229.4	8.4	0	3.7	0.58
W - decrease in both groups	570.1	570.4	178.4	201.6	239.3	8.6	0	4.0	0.76
L - decrease under ctr, effect on death, progr., switching (unobs. conf.)	569.9	569.9	160.8	219.2	213.8	8.0	0	3.4	0.62
L - decrease in both groups, effect on death, progr., switching (unobs. conf.)	569.6	570.2	178.5	201.5	223.3	8.2	0	3.7	0.77
time dependent correlation - Toeplitz	569.9	570.0	177.5	202.5	211.8	7.9	0	3.4	0.75
progression - slow	570.0	569.9	175.9	204.1	161.6	7.7	0	3.3	0.75
progresssion - no effect of W	569.5	570.1	177.1	202.9	221.5	7.8	0	3.3	0.75
progression - reduced effect of trt	569.8	569.9	177.3	202.7	211.7	7.8	0	3.3	0.75
progression - no effect of trt	569.9	570.2	177.3	202.7	211.5	7.8	0	3.3	0.75
progression - unobserved confounding with death	569.9	569.7	178.3	201.7	203.1	7.3	0	3.0	0.77
switch - high probability	569.6	570.0	183.1	196.9	396.7	8.0	0	3.4	0.76
switch - no effect of W	570.0	569.9	177.7	202.3	219.2	7.8	0	3.3	0.75
switch - W>0 near separation	569.8	570.3	179.1	200.9	283.8	7.9	0	3.3	0.75
switch - unobserved confounding with death	570.0	570.4	177.8	202.2	205.1	7.3	0	3.0	0.77
death - high rate	569.8	570.1	176.8	203.2	157.6	5.4	0	2.2	0.75
death - no effect of W	569.8	570.2	176.4	203.6	215.8	8.4	0	3.6	0.75
death - unobserved confound ing with progr. and switch	570.0	569.9	177.9	202.1	196.2	7.3	0	3.0	0.77
Trt effect post switch - none	570.1	570.1	171.8	208.2	209.4	7.6	0	3.2	0.75
random censoring - none	570.0	570.1	177.3	202.7	212.8	0.0	0	3.2	0.76
random censoring - effect of X	570.3	570.1	177.1	202.9	213.0	8.3	0	3.3	0.75
random censoring - effect of X and W	570.2	570.2	177.0	203.0	214.0	8.8	0	3.4	0.75
random censoring - effect of X, W and L	569.9	570.1	177.6	202.4	204.9	8.9	0	3.2	0.77
random censoring - high rate, effect of X, W	570.7	570.1	177.2	202.8	216.1	22.3	0	4.2	0.75
random censoring - high rate, control only, effect of X, W	569.8	570.2	198.5	181.5	208.8	12.3	0	3.6	0.76

Table 2.8: Descriptive statistics for the data simulated under scenarios with large sample size under the null hypothesis. Statistics are mean values for the number of patients under tretament and control (trt, ctr), the number of events in either group (ev trt, ev ctr), the number of control group patients who switched (switch), the number of patients for whom the time do death was censored by a random censoring event (rcens), the number of simulated data sets out of 10,000 in which the aimed for number of events was not achieved within the maximum study duration (ev<aim), the average maximum follow up time in years (FU) and the true hazard ratio conditional on the covariates X and W at baseline (HR). For each scenario, 10,000 data sets were simulated.

Scenario	trt	ctr	ev trt	ev ctr	switch	rcens	ev<aim	FU	HR
Core	570.1	570.3	189.8	190.2	202.3	7.1	0	2.9	1.00
W - decrease under ctr	569.9	569.8	190.1	189.9	210.5	7.5	0	3.2	1.00
W - decrease in both groups	570.0	569.9	190.1	189.9	230.7	8.1	0	3.6	1.00
L - decrease under ctr, effect on death, progr., switching (unobs. conf.)	570.0	570.0	190.2	189.8	192.1	6.9	0	2.9	1.00
L - decrease in both groups, effect on death, progr., switching (unobs. conf.)	570.1	569.8	190.0	190.0	214.6	7.9	0	3.4	1.00
time dependent correlation - Toeplitz	570.1	570.0	190.0	190.0	202.5	7.2	0	3.0	1.00
progression - slow	569.9	569.9	190.0	190.0	151.5	7.1	0	2.9	1.00
progresssion - no effect of W	569.5	570.3	189.9	190.1	212.6	7.0	0	2.9	1.00
progression - reduced effect of trt	570.0	570.4	190.0	190.0	202.3	7.0	0	2.9	1.00
progression - no effect of trt	569.7	570.0	189.9	190.1	202.3	7.0	0	2.9	1.00
progression - unobserved confounding with death	570.1	570.2	190.1	189.9	193.4	6.6	0	2.7	1.00
switch - high probability	570.1	569.7	190.1	189.9	378.1	7.1	0	2.9	1.00
switch - no effect of W	570.0	570.0	189.9	190.1	209.9	7.0	0	2.9	1.00
switch - W>0 near separation	569.6	569.8	189.9	190.1	270.6	7.0	0	2.9	1.00
switch - unobserved confounding with death	570.0	570.4	189.9	190.1	194.9	6.7	0	2.7	1.00
death - high rate	567.6	567.4	190.1	189.9	145.4	4.9	0	2.0	1.00
death - no effect of W	569.9	569.8	190.0	190.0	207.8	7.6	0	3.2	1.00
death - unobserved confound ing with progr. and switch	569.9	570.0	190.0	190.0	186.3	6.7	0	2.7	1.00
random censoring - none	569.6	570.1	189.9	190.1	203.0	0.0	0	2.9	1.00
random censoring - effect of X	570.4	569.8	190.1	189.9	203.1	7.4	0	2.9	1.00
random censoring - effect of X and W	570.1	569.8	190.1	189.9	204.5	8.1	0	3.0	1.00
random censoring - effect of X, W and L	570.2	570.3	189.9	190.1	195.2	8.0	0	2.8	1.00
random censoring - high rate, effect of X, W	569.8	570.1	190.0	190.0	207.5	19.2	0	3.5	1.00
random censoring - high rate, control only, effect of X, W	570.2	570.0	208.2	171.8	199.9	11.3	0	3.1	1.00

Table 2.9: Descriptive statistics for the data simulated under scenarios with different random censoring patterns under the alternative hypothesis. Statistics are mean values for the number of patients under treatment and control (trt, ctr), the number of events in either group (ev trt, ev ctr), the number of control group patients who switched (switch), the number of patients for whom the time do death was censored by a random censoring event (rcens), the number of simulated data sets out of 10,000 in which the aimed for number of events was not achieved within the maximum study duration (ev<aim), the average maximum follow up time in years (FU) and the true hazard ratio conditional on the covariates X and W at baseline (HR). For each scenario, 10,000 data sets were simulated.

Scenario	trt	ctr	ev trt	ev ctr	switch	rcens	ev<aim	FU	HR
Core	99.0	99.0	27.8	38.2	38.4	3.9	1	4.0	0.51
random censoring - none	98.8	98.8	27.7	38.3	38.7	0.0	0	3.8	0.50
random censoring - effect of X	99.0	99.1	27.7	38.3	38.6	4.1	1	4.1	0.51
random censoring - effect of X and W	99.0	99.0	27.6	38.4	38.8	4.5	11	4.2	0.51
random censoring - effect of X, W and L	98.9	99.0	27.9	38.1	37.3	4.4	6	3.8	0.53
random censoring - high rate, effect of X, W	98.9	98.9	27.2	37.7	38.8	10.1	2241	5.7	0.50
random censoring - high rate, control only, effect of X, W	99.0	99.0	32.4	33.6	37.7	5.6	63	4.5	0.51

Table 2.10: Descriptive statistics for the data simulated under scenarios with different random censoring patterns under the null hypothesis. Statistics are mean values for the number of patients under treatment and control (trt, ctr), the number of events in either group (ev trt, ev ctr), the number of control group patients who switched (switch), the number of patients for whom the time do death was censored by a random censoring event (rcens), the number of simulated data sets out of 10,000 in which the aimed for number of events was not achieved within the maximum study duration (ev<aim), the average maximum follow up time in years (FU) and the true hazard ratio conditional on the covariates X and W at baseline (HR). For each scenario, 10,000 data sets were simulated.

Scenario	trt	ctr	ev trt	ev ctr	switch	rcens	ev<aim	FU	HR
Core	98.8	98.9	33.0	33.0	35.0	3.0	0	2.9	1.00
random censoring - none	99.1	99.1	33.1	32.9	35.2	0.0	0	2.9	1.00
random censoring - effect of X	99.1	99.1	33.0	33.0	35.1	3.1	0	3.0	1.00
random censoring - effect of X and W	99.0	98.8	33.0	33.0	35.3	3.4	0	3.0	1.00
random censoring - effect of X, W and L	99.0	99.1	33.0	33.0	33.7	3.3	0	2.8	1.00
random censoring - high rate, effect of X, W	99.2	99.2	33.0	33.0	36.1	8.0	7	3.6	1.00
random censoring - high rate, control only, effect of X, W	99.2	99.0	36.2	29.8	34.6	4.8	0	3.1	1.00

2.4.2 Bias of estimated log hazard ratio

Figures 2.2 to 2.7 show the empirical relative bias for the estimated hazard ratios of the studied methods across scenarios.

Overall, most methods had only small relative bias, with deviations of only few percentage points, in scenarios without explicit violations of their assumptions. In the small sample settings, bias was more pronounced, while in the large sample settings bias was negligible in many scenario/method combinations and in general agreement between methods was larger. There were, however, several combinations of methods/scenarios where bias was considerable.

Notable bias due to small sample sizes (and not violation of assumptions) was observed when recensoring was applied to RPSFTM and TSE.

The IPW approach provided unbiased estimates under the core scenario and under scenarios with different progression mechanisms. In scenarios with unobserved confounding of death with switching or progression (or both), the IPW estimator was notably biased towards too strong effect sizes. These settings correspond to a violation of the key assumption of exchangeability (no unobserved confounding) for IPW. A high switching rate and presence of near separation in W were included to model near-violation of the positivity assumption. These scenarios resulted in notable bias in the small sample setting under H0, though these effect diminished under large sample size.

A small though significant bias was observed for IPW in the scenario where the time dependent covariate W had a declining time trend under control but stable mean under treatment (scenario "W - decrease under ctr"). Of note, this scenario did not include unobserved confounders. Instead, the bias could be due to patients who reach disease progression at different follow-up times having systematically different covariate profiles at the time of switching and different subsequent hazard functions. Simply put, when patients switch treatment, they should be implicitly replaced, from that time on, by similar patients who did not switch. However, since the IPW model does not take into account the time of switching, the "replacement" is a mixture of patients who experienced disease progression at a different time-points. While these patients may have a matching probability to switch, their future hazard for death may be different, depending on the current values of X and W(t) and the future trajectory of W(t). A possibly remedy could be to calculate weights not only depending on switching at the time of progression but to include the time-dependent probability to progress in the model for the weights. This mechanism warrants further exploration, as it may impact the validity of IPW even in settings where potential confounders have been observed.

RPSFTM was performed with and without recensoring. In the small sample setting, results diverged between these two approaches. With small sample size under the alternative, RPSFTM without recensoring was unbiased in many scenarios and recensoring resulted in overestimation of the effect, whereas under H0 the approach with recensoring was typically closer to the true value and RPSFTM without recensoring was slightly biased towards favouring the control group.

With large sample sizes, there was only little difference between the two approaches, albeit estimates using RPSFTM with recensoring often had lower bias.

RPSFTM performed particularly poor under H1 in the beforementioned scenario "W - decrease under ctr" and in the related scenario "L - decrease under ctr, effect on death, progr., switching (unobs. conf.)" where the unobserved variable L has differential patterns in treatment and control and induces confounding. In both scenarios, though, recensoring removed the bias.

In scenarios where the treatment effect after switching is lower than the effect for the initial treatment group, the key assumption for RPSFTM of a constant treatment effect was violated and the treatment effect was strongly overestimated. In these cases, the bias was aggravated by recensoring.

Finally, RPSFTM was more sensitive to a high switching rate under H0 then other methods. Under small sample size, a high switching rate (and hence a low data base for non-switchers) resulted in notable bias for all methods. Under large sample size in this scenario, only RPSFTM without recensoring and, to a lesser degree, with recensoring remained biased.

Two stage estimation was mostly affected by unobserved confounding, violating the key assumption of no unobserved confounding at secondary baseline, and in these scenarios often showed a bias similar to the IPW method. For TSE, recensoring in most scenarios had only a small impact on the estimate. Exceptions were the scenarios where W or L had declining trends under control. Here both TSE approaches were biased albeit in opposite directions.

G-formula was only studied under the small sample size to the high computational time burden of this method. The estimates obtained via g-formula showed a consistent bias favouring active treatment, both under H0 and under H1, over most scenarios. It had a particularly strong bias in the scenario "W - decrease under ctr", albeit here bias was in the direction of favouring the control treatment. A reason for this bias may be that the specified g-formula models were not sufficiently complex to correctly capture the time trend of W. Bias was also notably pronounced under a high switching rate and under scenarios with unobserved confounding.

The comparator method of censoring event times at the time of switching had only small bias in many scenarios. However, opposed to the causal inference methods, in many scenarios this bias remained in the large sample setting. This points to an inherent bias of this approach. Also, compared to the causal inference methods, it was more strongly affected by high switching rates. Finally, the censoring approach was particularly biased in scenarios with unobserved confounding, which is understandable as it relies on a correct model specification to adjust for artificial censoring of switchers.

The additional application of inverse probability of censoring weights to IPW, RPSFTM, TSE and g-formula in scenarios with different random censoring mechanisms in general had very little impact on the bias, see Figures 2.6 and 2.7. A notable exception is high rate in control group. Here the combination of IPW for swithcing and IPCW removed the small amount of bias that was present with oure IPW. under H0, albeit resulted in overestimation of the effect under H1.

In summary, the studied causal inference methods performed well under scenarios where their identifying assumptions were satisfied within the specified data-generating mechanism, albeit considerably bias was observed under violation of model assumptions. A small sample size in general may lead to biased results. Of note, the scenario with decreasing trend of W under control was challenging for all methods. This is of importance, as this scenario corresponds to a realistic case in which patients under treatment and under control do not only differ in their primary endpoint but also in mediating variables.

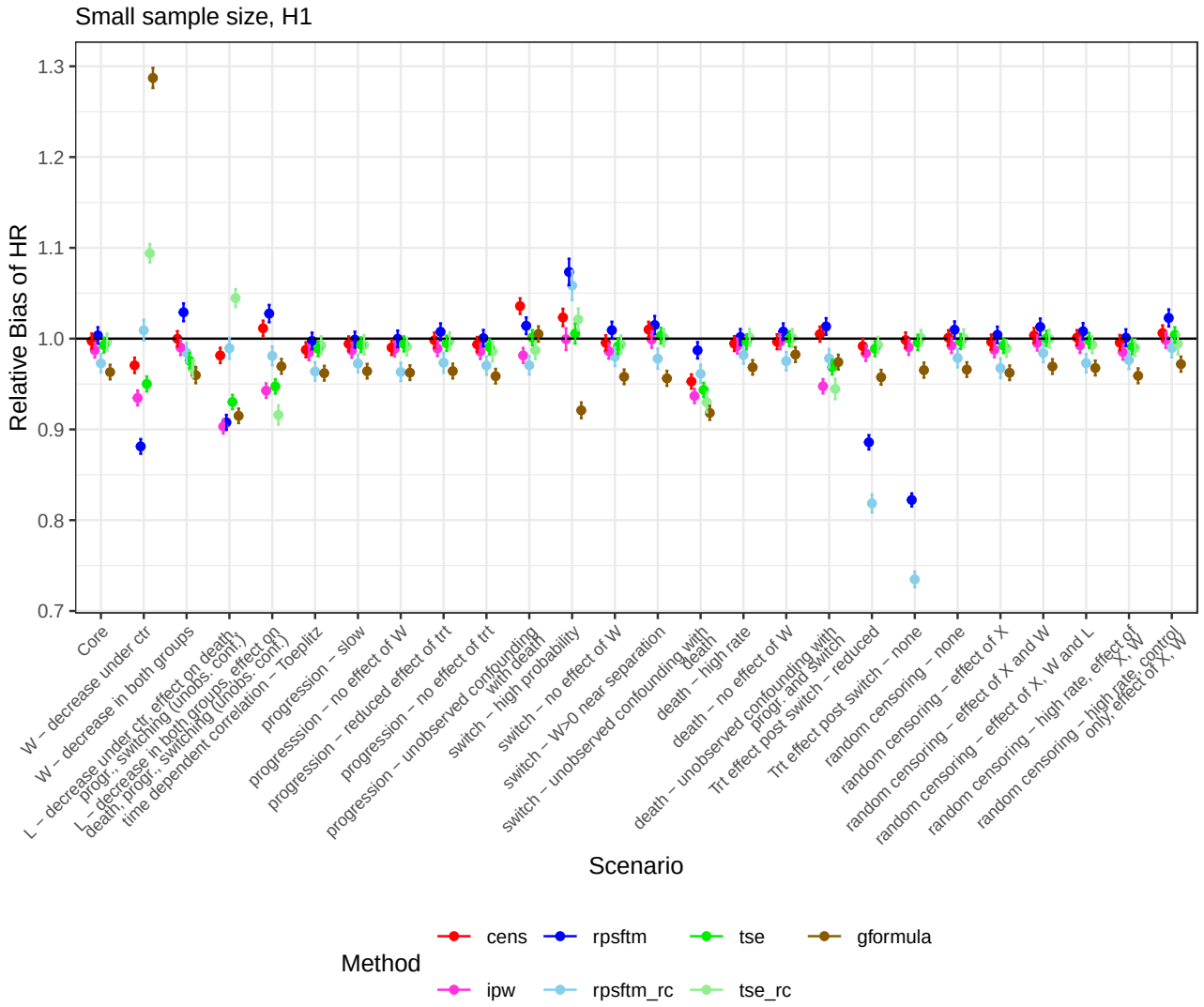


Figure 2.2: Relative bias of the estimated hazard ratio in treatment switching scenarios with small sample size under the alternative.

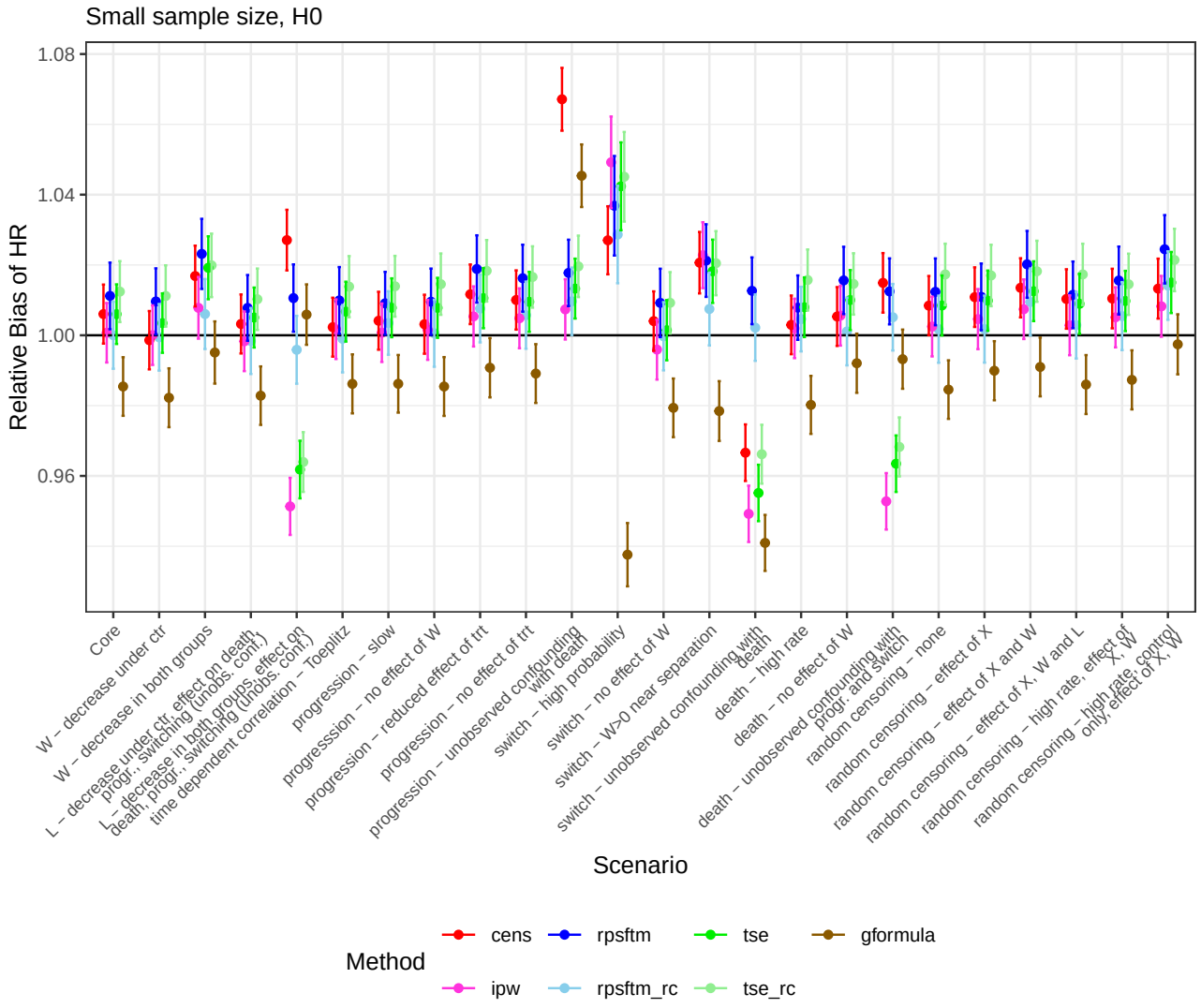


Figure 2.3: Relative bias of the estimated hazard ratio in treatment switching scenarios with small sample size under the null hypothesis.

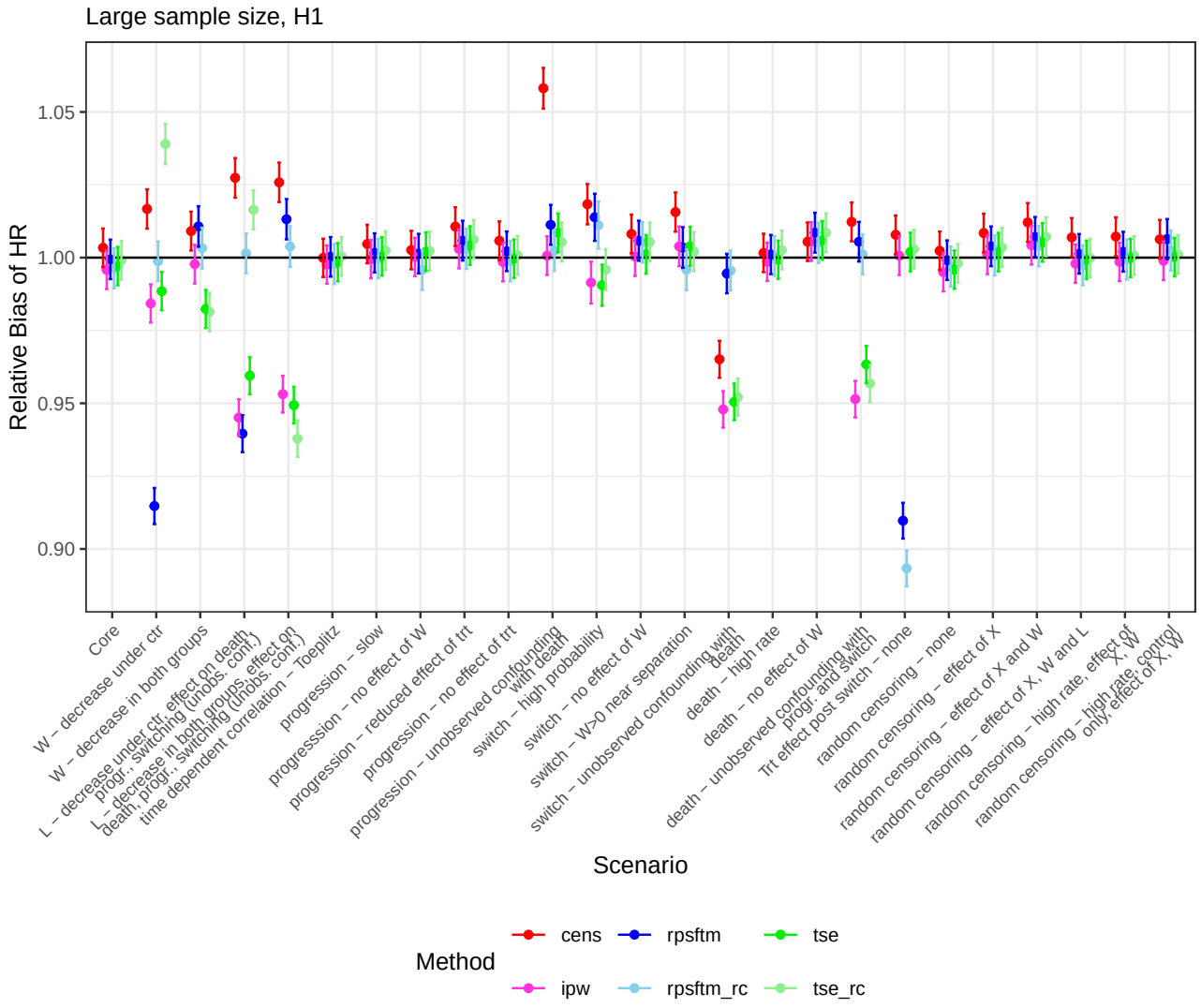


Figure 2.4: Relative bias of the estimated hazard ratio in treatment switching scenarios with large sample size under the alternative.

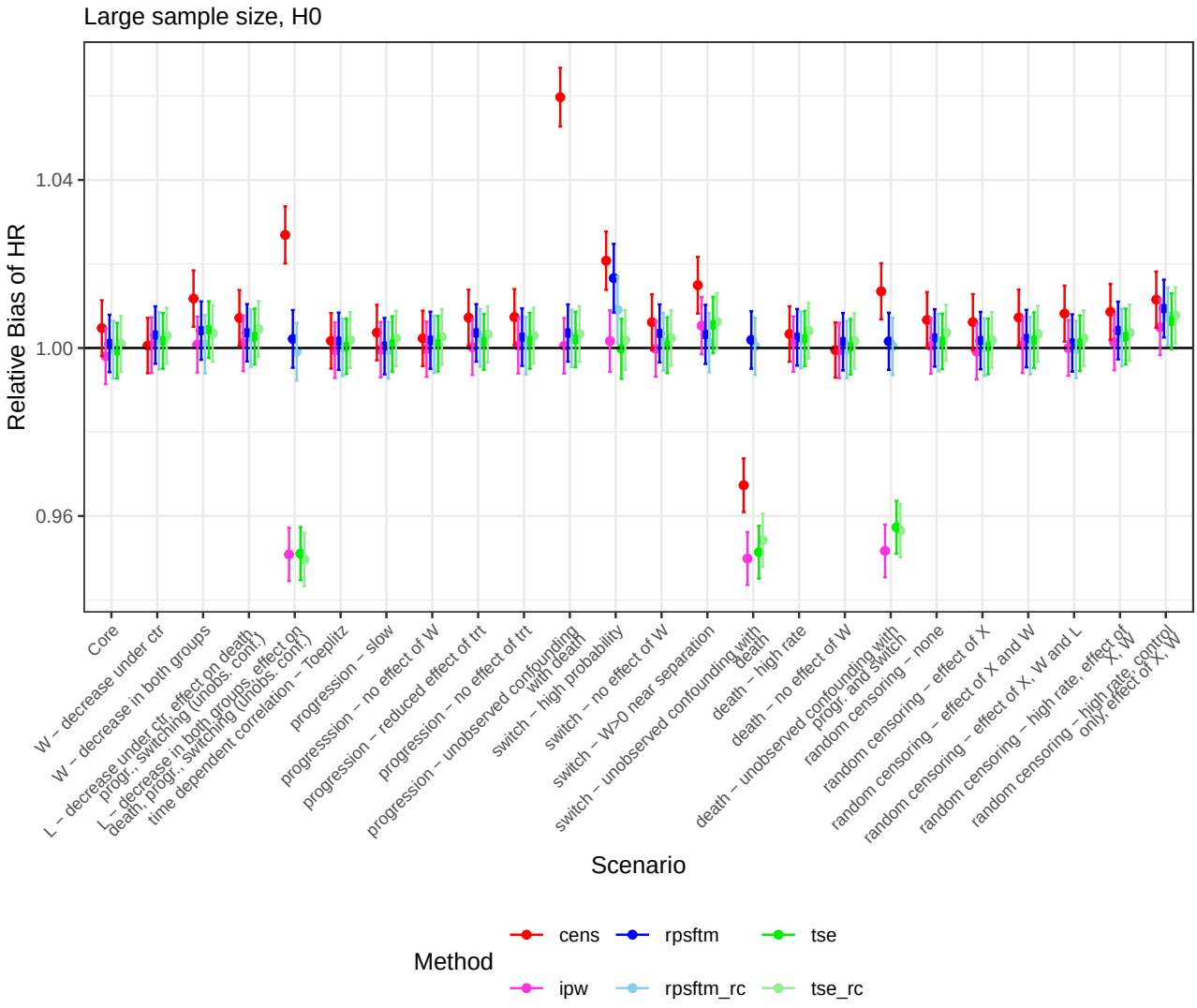


Figure 2.5: Relative bias of the estimated hazard ratio in treatment switching scenarios with small sample size under the null hypothesis.

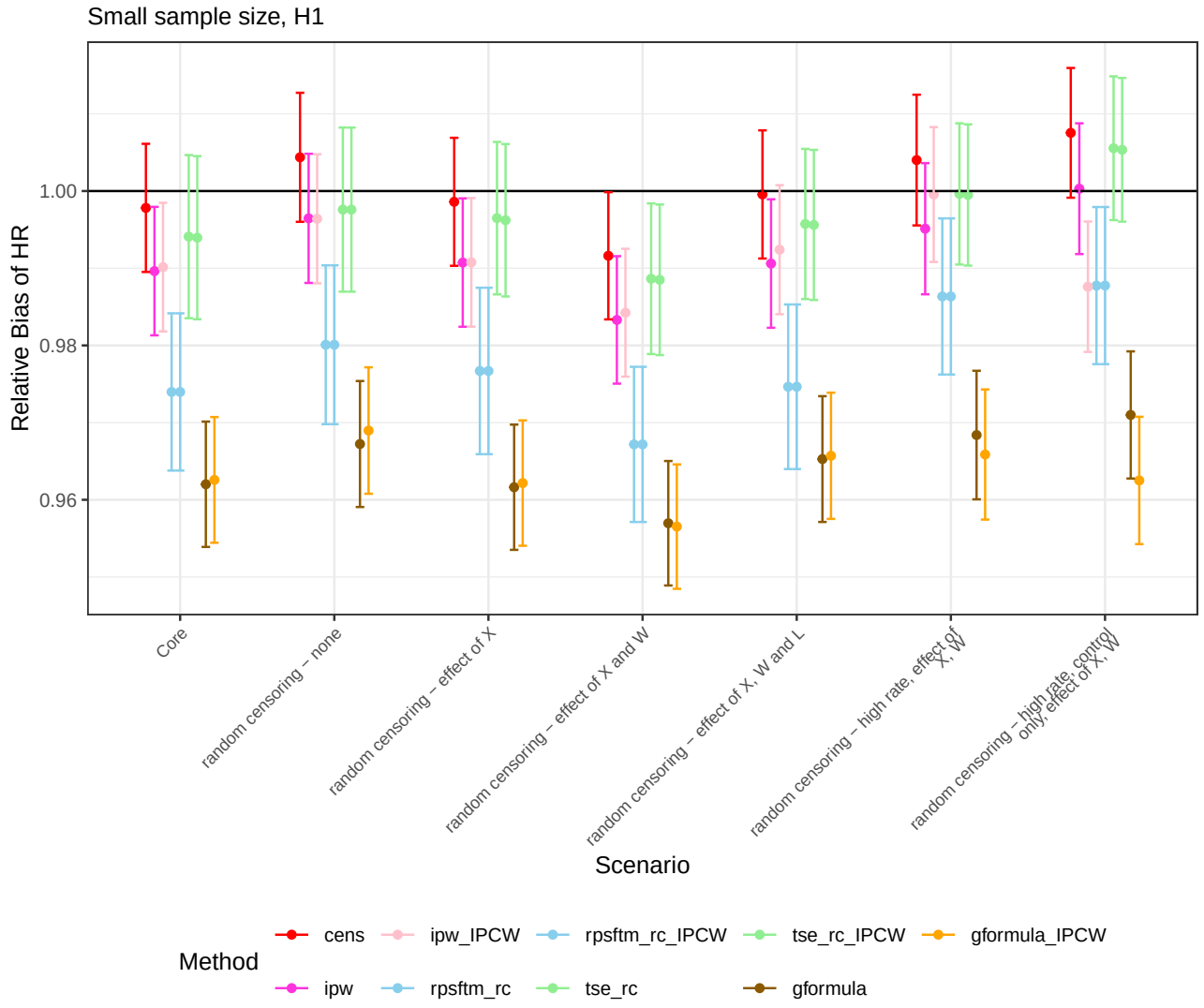


Figure 2.6: Relative bias of the estimated hazard ratio for methods with and without additional inverse probability weighting for random censoring in treatment switching scenarios with small sample size under the alternative.

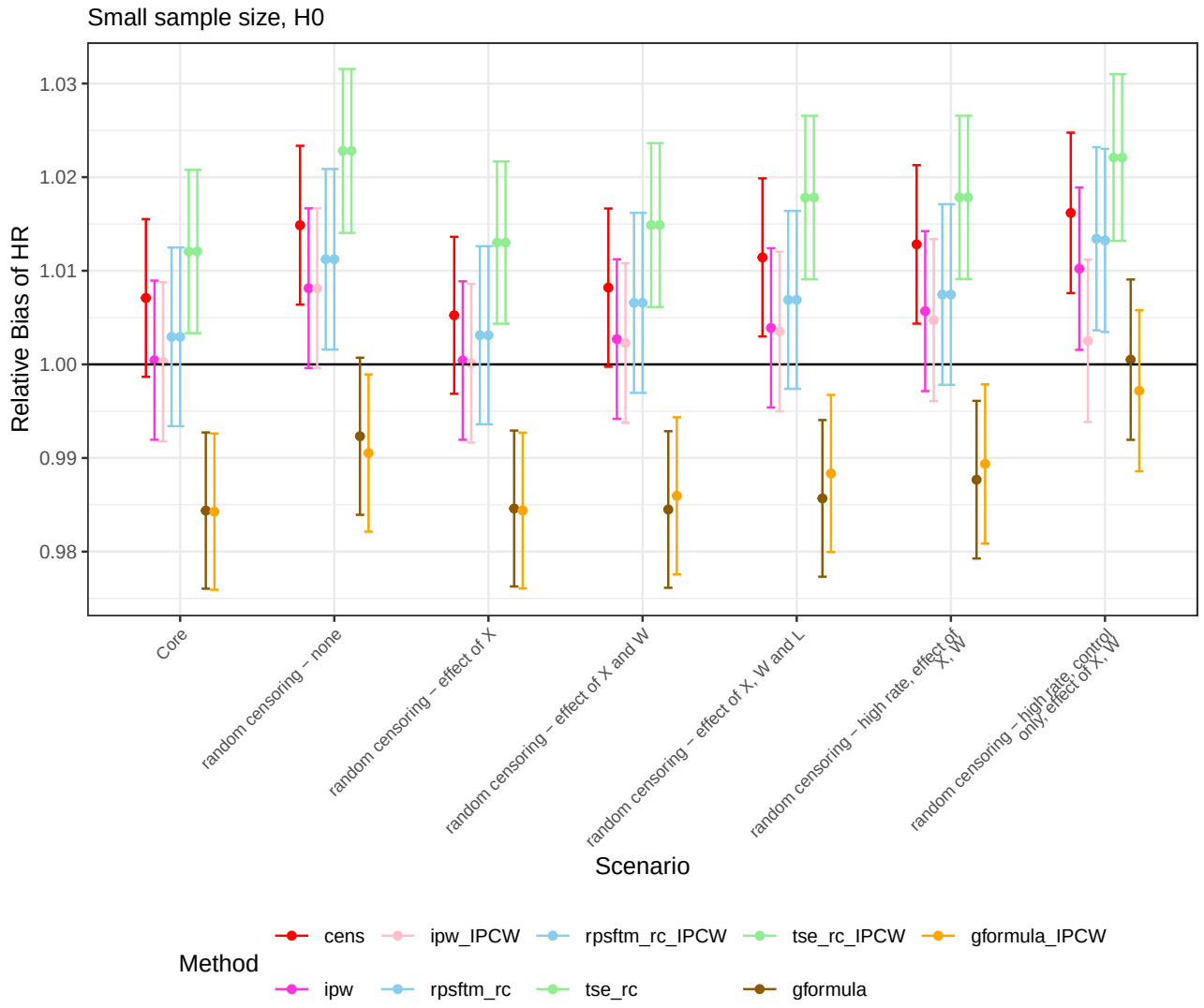


Figure 2.7: Relative bias of the estimated hazard ratio for methods with and without additional inverse probability weighting for random censoring in treatment switching scenarios with small sample size under the null hypothesis.

2.4.3 Type I error rate

Empirical type I error rates are displayed in Figures 2.8 to 2.10. The investigation of type I error rates showed that, independent of other scenario characteristics, type I error rate control was affected by small sample sizes for several methods. In particular, IPW was in general too liberal and TSE was too conservative with small sample sizes, whereas both methods had type I error close to the nominal level in large sample scenarios where their assumptions were met. It should be noted that type I error rate inflation may result from biased estimates of the log hazard ratio or from biased estimates of its variance. For IPW, however, when assumptions of no unobserved confouding and positivity were met, there was no bias favouring treatment under H0. Hence the small sample inflation likely results from insufficient small sample properties of the robust variance estimator.

RPSFTM was slightly too liberal also in large sample settings. Type I error rates for the g-formula approach (which was only studied under small sample sizes) were relatively close to 0.025. Also for the censoring comparator approach there was not notably inflation of type I error rate due to small sample size. For both RPSFTM and TSE, recensoring resulted in clearly conservative behaviour, although this effect was less pronounced with larger sample size. In settings with unobserved confound, tough, IPW, TSE and the censoring approach showed deviations from the nominal level and these could result in inflations up to approximately 5-6%. This type of inflation was more pronounced under large sample sizes.

In the scenario with high switching probability all methods showed deviations from the nominal level to some extent. IPW, RPSFTM and g-formula showed an inflation of the type I error towards above 6% (IPW) or around 4% (RPSFTM and g-formula) in this scenario with small sample size. In the large sample settings, type I error rate was closer to the nominal 2.5% level for most methods in most scenarios. With small sample sizes, IPW, RPSFTM and TSE showed an inflation towards approximately 3%. The type I error inflation due to high switching probability is not due to a strict violation if the positivity assumption (as the assumption was met in the data generating mechanism), but due to potentially unstable models with a too small data basis in small samples combined with near-violation of the assumption.

Different random censoring mechanisms had only a small impact on type I error rate. The scenario with high censoring in the control group only resulted in conservative behaviour of all methods when the sample size was large. The additional application of IPCW, studied for small sample sizes, turned the IPW method slightly more liberal in some scenarios. The g-formula approach was more conservative after inclusion of IPCW; for RPSFTM and TSE (with reconsoring) additional IPCW had little impact, these methods stayed conservative.

Overall, type I error rate control of the studied methods was found to be sensitive to too small sample sizes and, to some degree, to unobserved confounding.

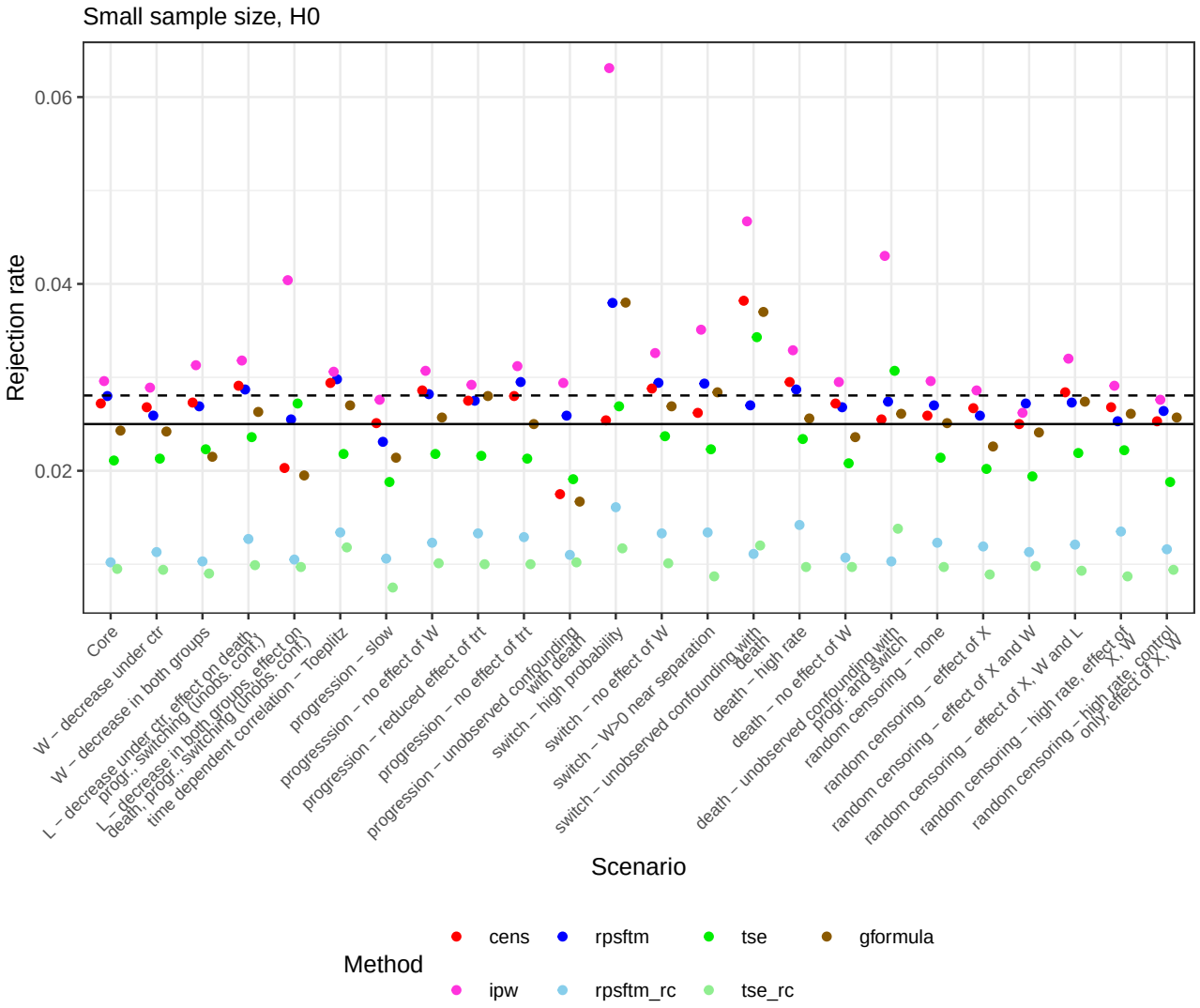


Figure 2.8: Type I error rate of the studied methods in treatment switching scenarios with small sample size. The solid reference line corresponds to the nominal one-sided significance level of 0.025. The dashed reference line indicates simulation uncertainty. It marks the value below which the empirical type I error rates should be with 97.5% probability (by asymptotic approximation) if the true type I error rate was 0.025.

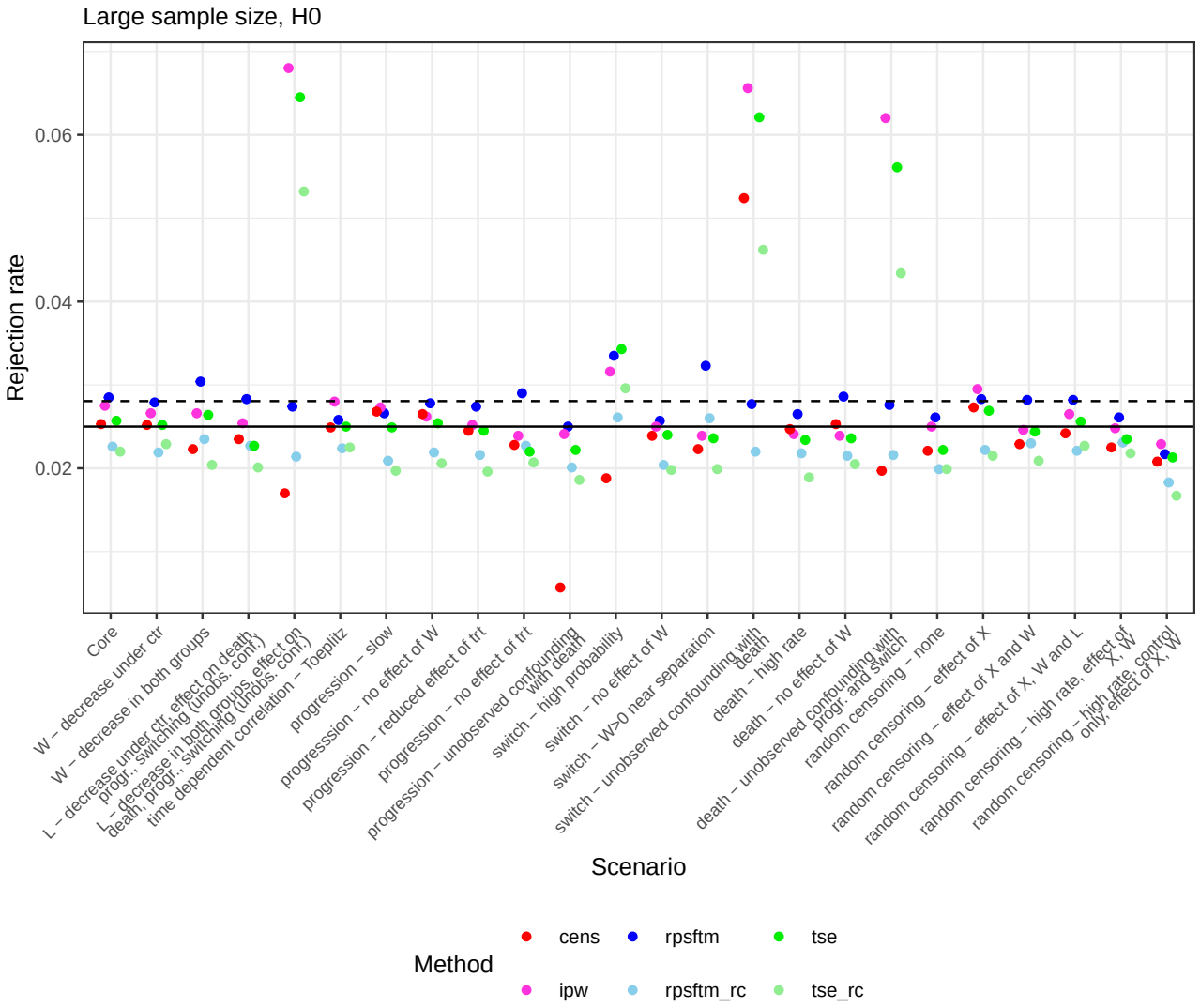


Figure 2.9: Type I error rate of the studied methods in treatment switching scenarios with large sample size. The solid reference line corresponds to the nominal one-sided significance level of 0.025. The dashed reference line indicates simulation uncertainty. It marks the value below which the empirical type I error rates should be with 97.5% probability (by asymptotic approximation) if the true type I error rate was 0.025.

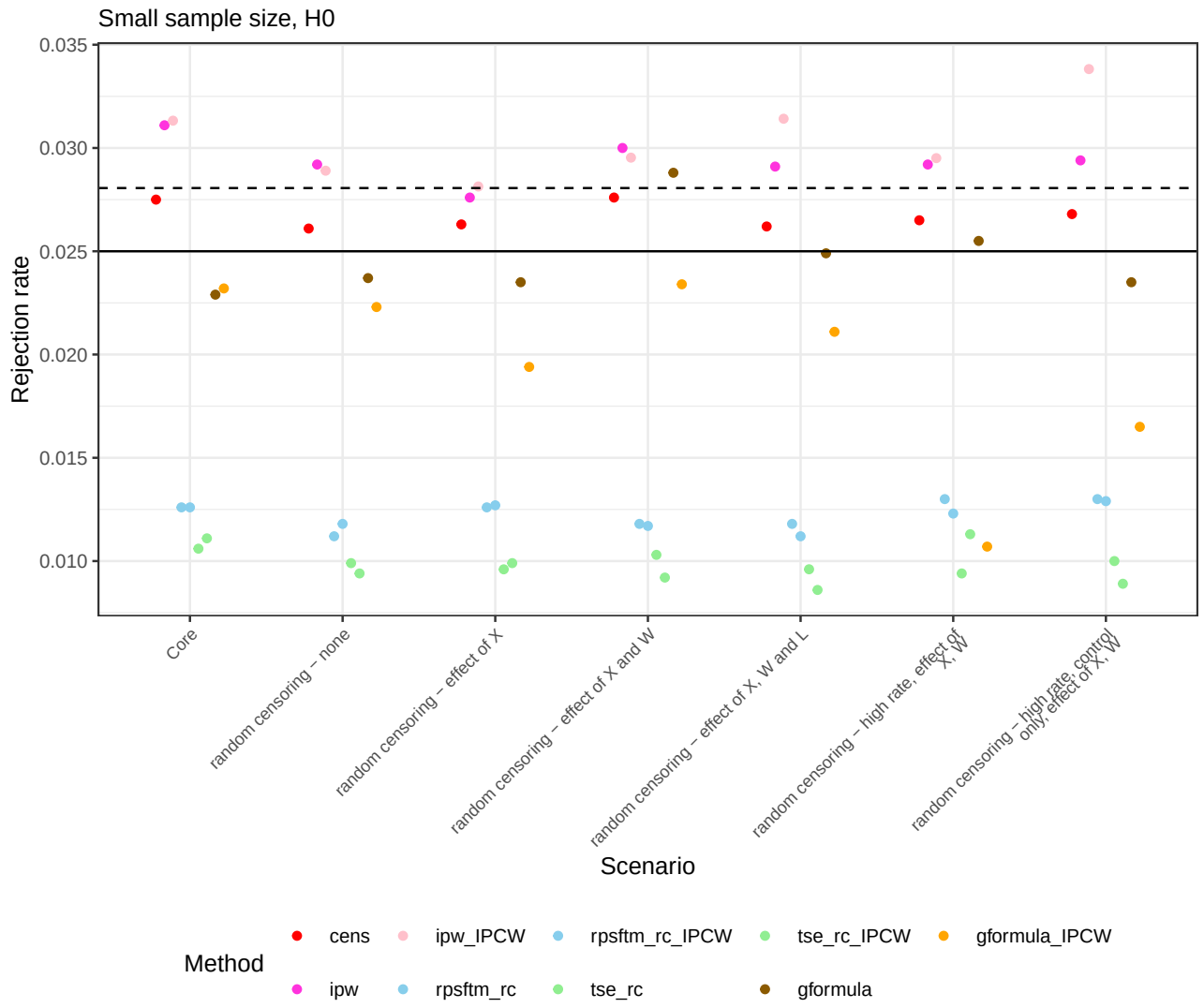


Figure 2.10: Type I error rate of methods with and without additional inverse probability weighting for random censoring in treatment switching scenarios with small sample size. The solid reference line corresponds to the nominal one-sided significance level of 0.025. The dashed reference line indicates simulation uncertainty. It marks the value below which the empirical type I error rates should be with 97.5% probability (by asymptotic approximation) if the true type I error rate was 0.025.

2.4.4 Power

The empirical power is shown in Figures 2.11 to 2.13. The power of the studied methods differed substantially and there was a clear pattern that was consistent across scenarios: In general IPW and the censoring approach had the highest power, around 70% and performed similar. TSE was slightly less powerful, followed by the g-formula. RPSFTM had substantially less power, with values around 50%, than the beforementioned methods. An exception was the scenario in which there was no effect of the experimental treatment when applied after switching. However, the point estimate of RSFTM was considerably biased towards too strong effect sizes in this scenario (see results on bias). Recensoring severaly reduced the power of TSE and RPSFTM. In the small sample setting, the respective power dropped to values around 25%. In the large sample setting, the decrease in power due to recensoring was still notable though far less pronounced with decreases in the order of a few percentage points.

Of note, in most scenarios the true treatment effect was very close to the effect assumed for sample size planning for 80% power. However, for most scenarios the power of all studied methods was below 80%, with typical values for the most powerful methods around 70%. For methods that censor event times at switching, i.e. IPW and the comparator censoring approach, a likely explanation for this smaller than intended power is that the effective number of event is reduced due to artificial censoring. Similarly, the drop in power after recensoring in RPSFTM and TSE may be due to masked events. For IPW, the additional variability due to weights which are also estimated values (and accounted for in the robust variance estimator), may further reduce power. Similarly, calculations via g-formula include steps that introduce additional variability, accounted for by the bootstrap variance estimator, compared to relatively simple methods such as purely censoring times after switching. For RPSFTM the main source of additional variability is the estimated acceleration parameter which is applied to calculate counterfactual times and thereby potentially passes on variability to a large part of the data.

When comparing power across scenarios, not that in the scenarios where either W or L decrease in the control group only, the true hazard ratios are much smaller than in other scenarios due to the detrimental effect of smaller values of the time dependent covariates.

Of further note, in the scenario with high switching rate, the true effect is similar to the majority of the other scenarios, however the power of all methods much smaller in this scenario, reaching at most around 50%. As discussed above, methods with censoring after switching will utilize a much smaller event number here, and other methods such as TSE may suffer from a too small data base for non-switched control patients to reliable calculate differential treatment effects.

In summary, the studied methods were found to differ substantially their power to identify a treatment effect under the hyothetical estimand and these differences are not driven by scenario characteristics but seem to be inherent and suggest that in particular RPSFTM and the application of recensoring lead to inefficient hypothesis tests. Of further relevance, the power was consistently below the nominaly value suggested by traditional sample size planning. Hence, power and sample size planning for causal inference methods under treatment switching may need refinement to adjust for their larger variability compared to more simple outcome models.

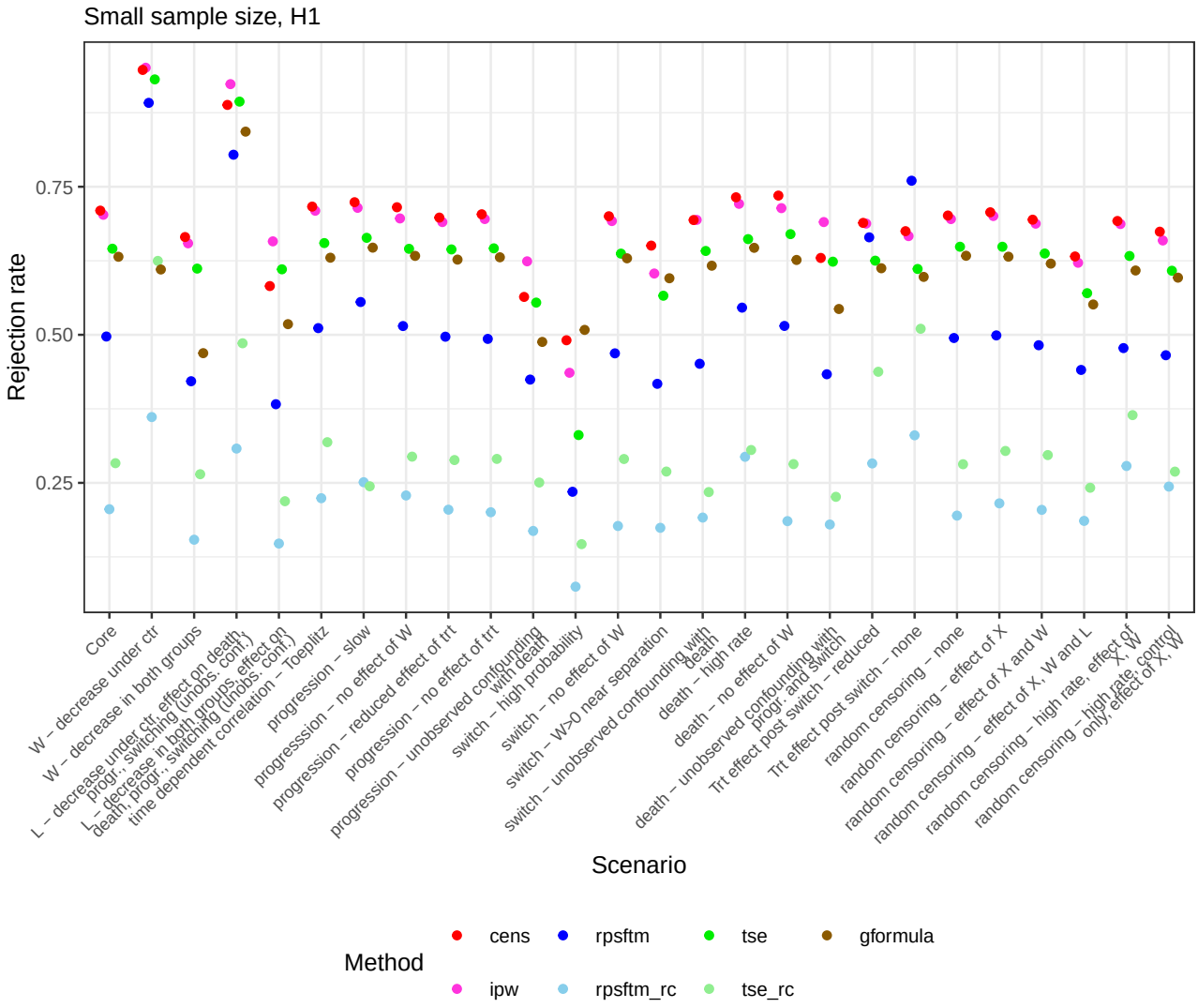


Figure 2.11: Power of the studied methods in treatment switching scenarios with small sample size.

2.4.5 Coverage and width of confidence intervals for the hazard ratio

Figures 2.14 to 2.19 show the empirical coverage probability of nominal 95% confidence intervals for the log hazard ratio. In small sample settings, IPW and the censoring comparator method showed a trend towards slight undercoverage, whereas TSE, g-formula as well as RPSFTM and TSE with recensoring typically had larger than nominal coverage probabilitits. For IPW, undercoverage was particularly pronounced in scenarios with near violations of the positivity assumption (high switching, near separation in W) and in scenarios with unobserved confounding. However, in addition to violations of assumptions, small sample size had an impact in coverage for the IPW method. A possible explanation is that the robust variance estimator, used to account for weighting, has non-negligibla bias under small sample sizes.

In scenarios where time dependent covariates had differential time trends between groups, intervals obtained through IPW, censoring approach, g-formula and RPSFTM were prone to undercoverage, likely due to their models being misspecified under these scenarios. Note that this effect is only observed under the alternative, as in null hypothesis scenarios, the time dependent covariates had no effect in the hazard for death. RPSFTM without recensoring showed a mixed picture. For this method, the coverage was too small by a few percentage points in scenarios where time dependent covariates had differential time trends between groups, For RPSFTM

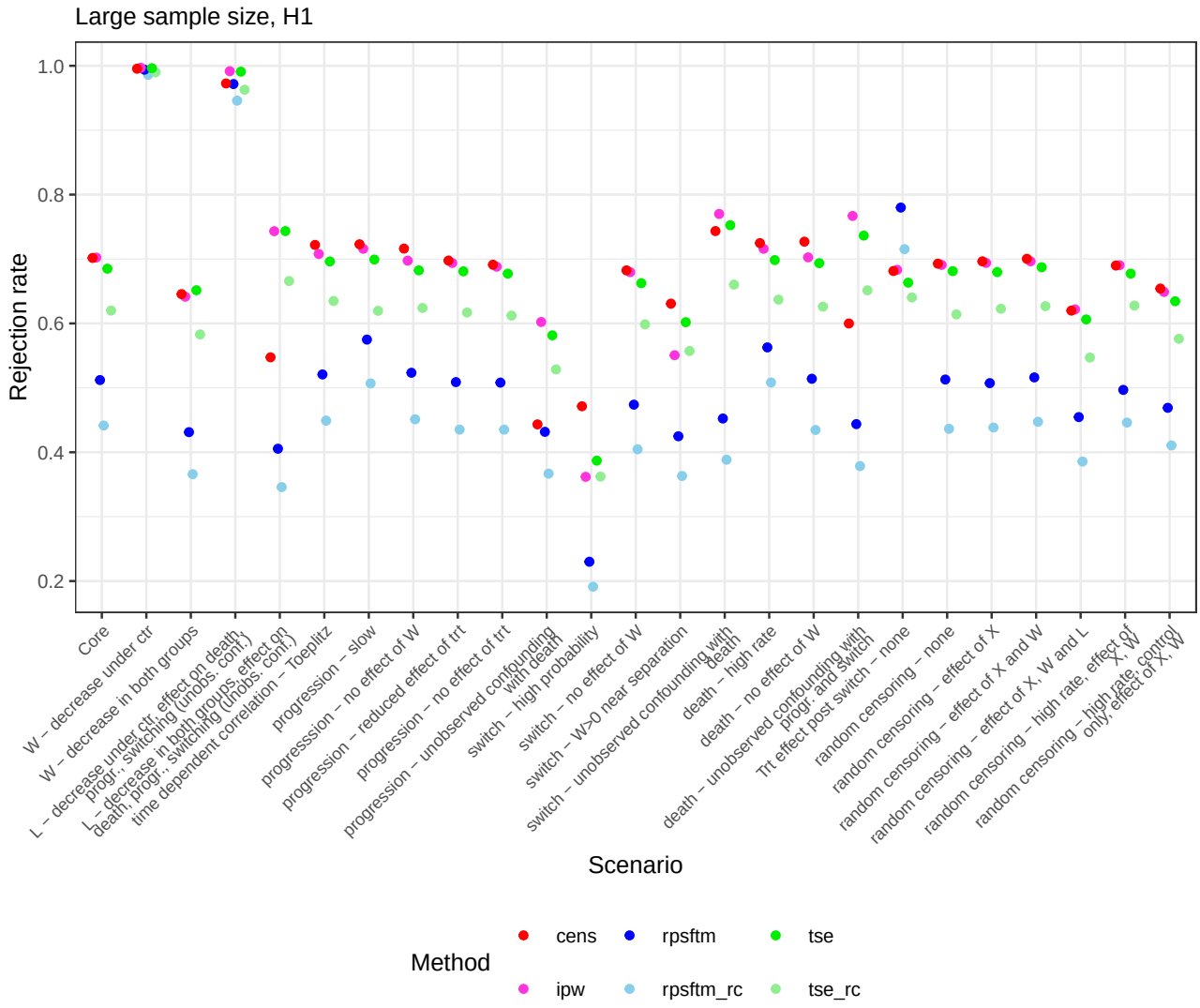


Figure 2.12: Power of the studied methods in treatment switching scenarios with large sample size.

coverage was also too small in the scenario with high switching probability and in scenarios where the treatment effect was different for switchers than for initial treatment group patients, and thus the assumption of the method was violated. In other scenarios RPSFTM tended to results in higher than nominal coverage.

In the large sample, coverage was moving closer to the nominal level for most method/scenario combinations that showed higher than nominal coverage in the small sample setting. This held in particular for TSE and for recensoring applied to TSE and RPSFTM and suggests that the corresponding higher coverage was due to the pronounced bias of these methods in small sample sizes. However, the discussed situations with undercoverage mostly persisted or even became more pronounced. With large sample sizes, it became visible that also TSE is affected by unobserved confounding, leading to undercoverage.

Similar to other operating characteristics, the additional application of IPCW to account for random censoring had mostly negligible impact on coverage. An exception was g-formula where in some scenarios, additional IPCW resulted in an increase of the, already larger than nominal, coverage probability.

In summary, similar to finding on the type I error rate, the confidence interval coverage probabilities of all methods were found to be sensitive to meeting the model assumptions of the respective methods and to too

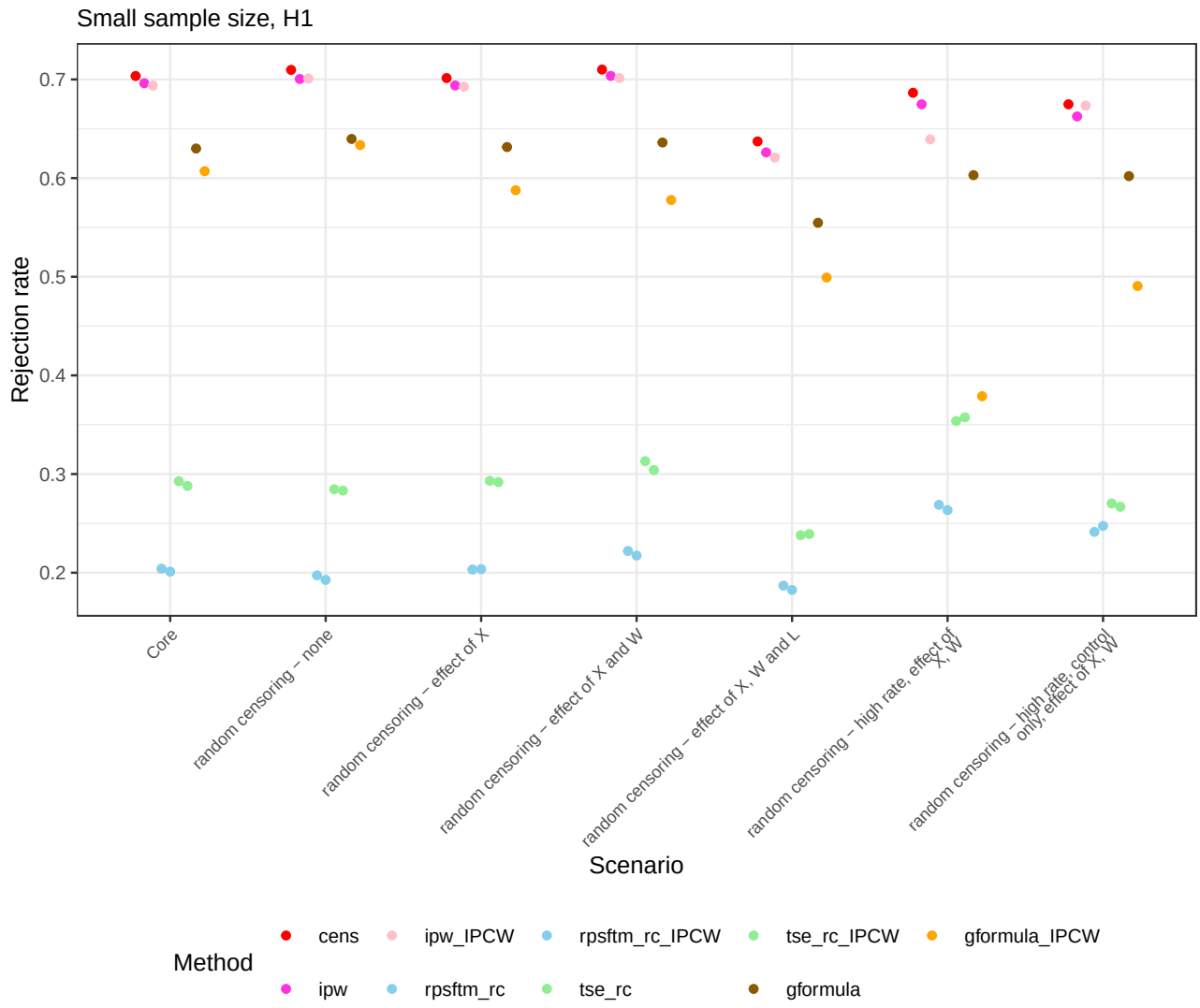


Figure 2.13: Power of methods with and without additional inverse probability weighting for random censoring in treatment switching scenarios with small sample size.

small sample size.

Figures 2.20 to 2.25 show width of the nominal 95% confidence intervals for the log hazard ratio. The pattern here closely reflects what was found for the power of the different methods: IPW, censoring and TSE in general had the smallest intervals, followed by g-formula and, with notably larger intervals, RPSFTM. Recensoring increased the width of confidence intervals for RPSFTM and TSE substantially.

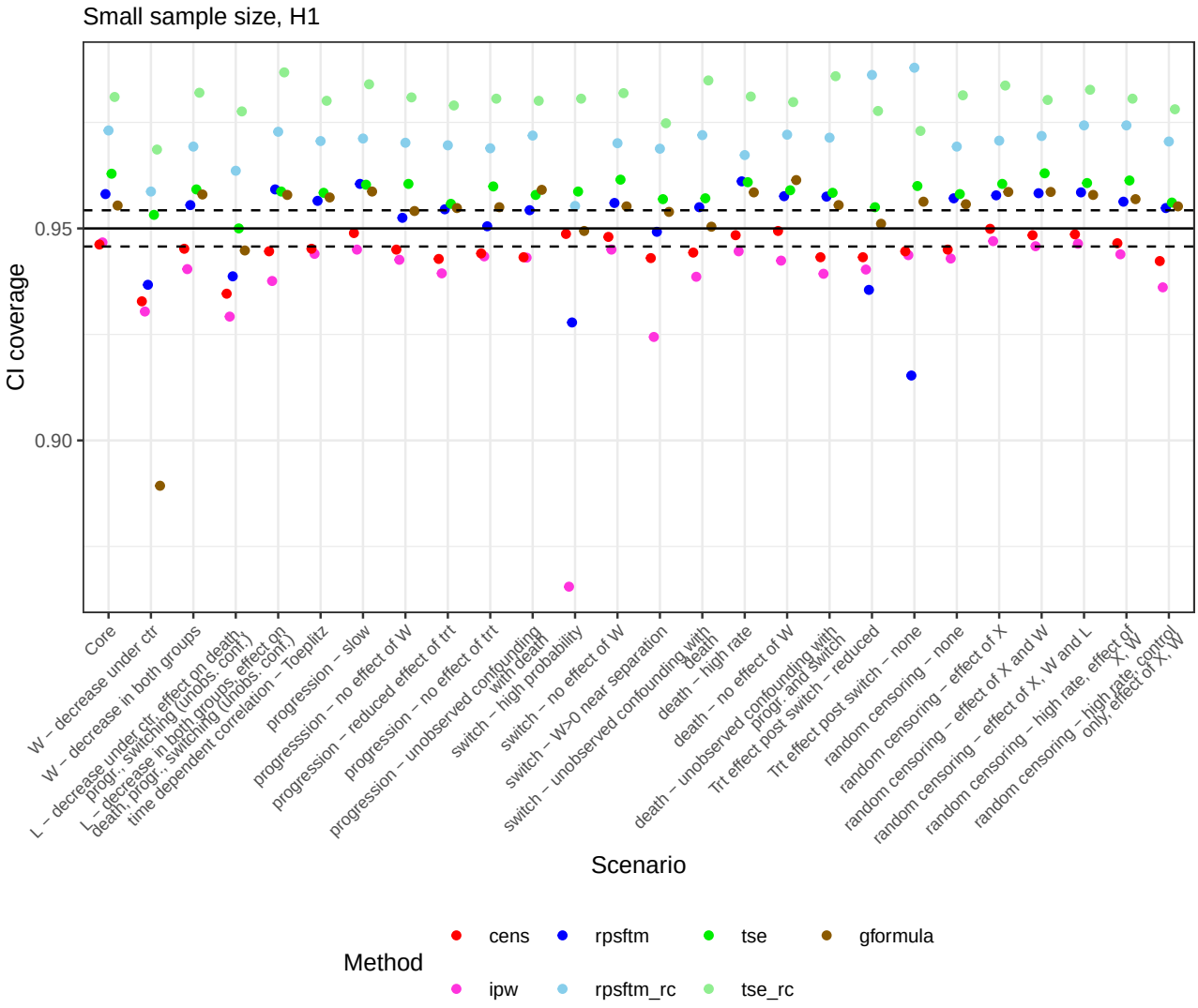


Figure 2.14: Confidence interval coverage of the studied methods in treatment switching scenarios with small sample size under the alternative. The solid reference line corresponds to the nominal 95% level. The dashed reference lines indicates simulation uncertainty. They mark the range, in which the empirical coverage probability should be with 95% probability (by asymptotic approximation) if the true coverage was 95%.

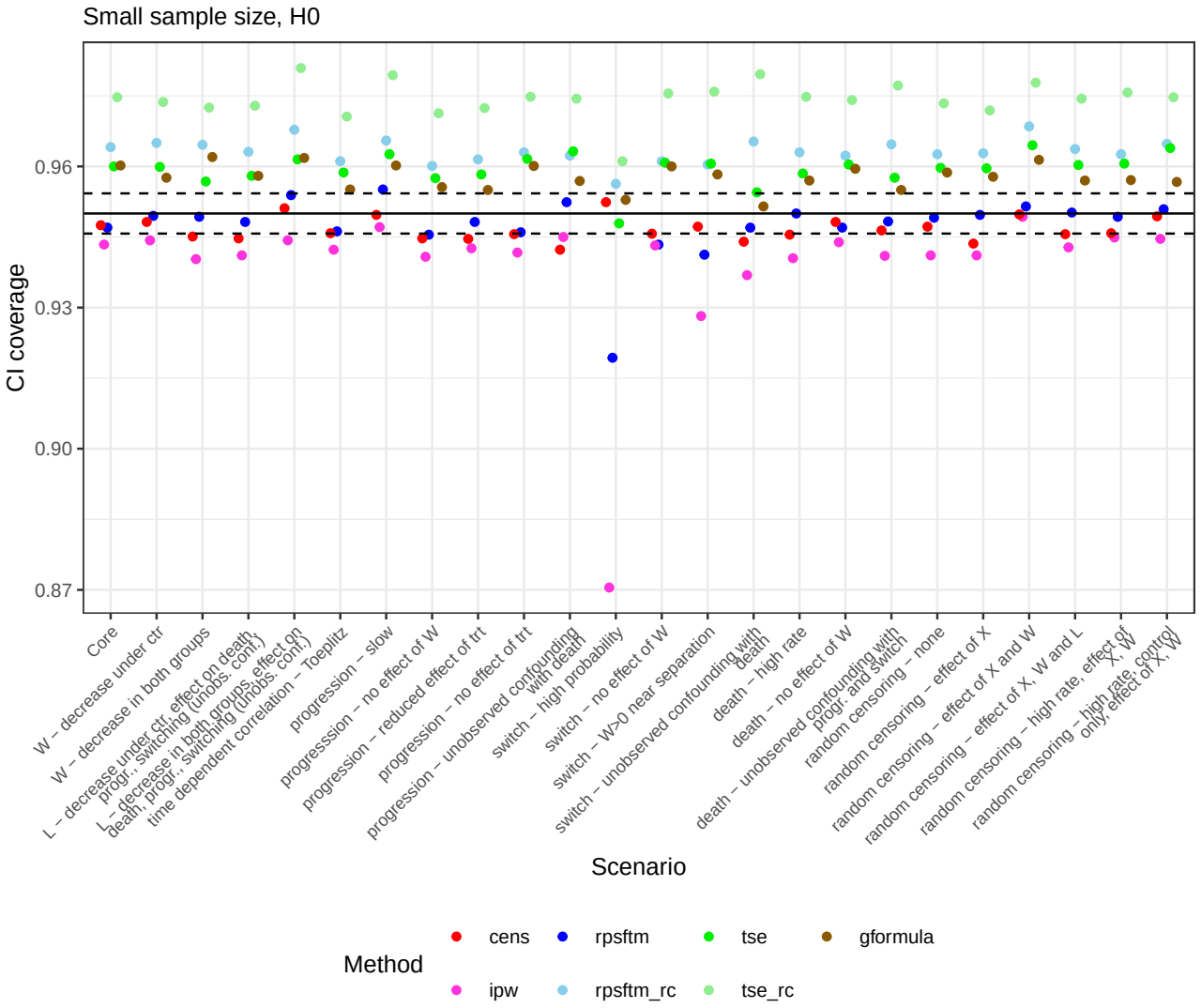


Figure 2.15: Confidence interval coverage of the studied methods in treatment switching scenarios with small sample size under the null hypothesis. The solid reference line corresponds to the nominal 95% level. The dashed reference lines indicates simulation uncertainty. They mark the range, in which the empirical coverage probability should be with 95% probability (by asymptotic approximation) if the true coverage was 95%.

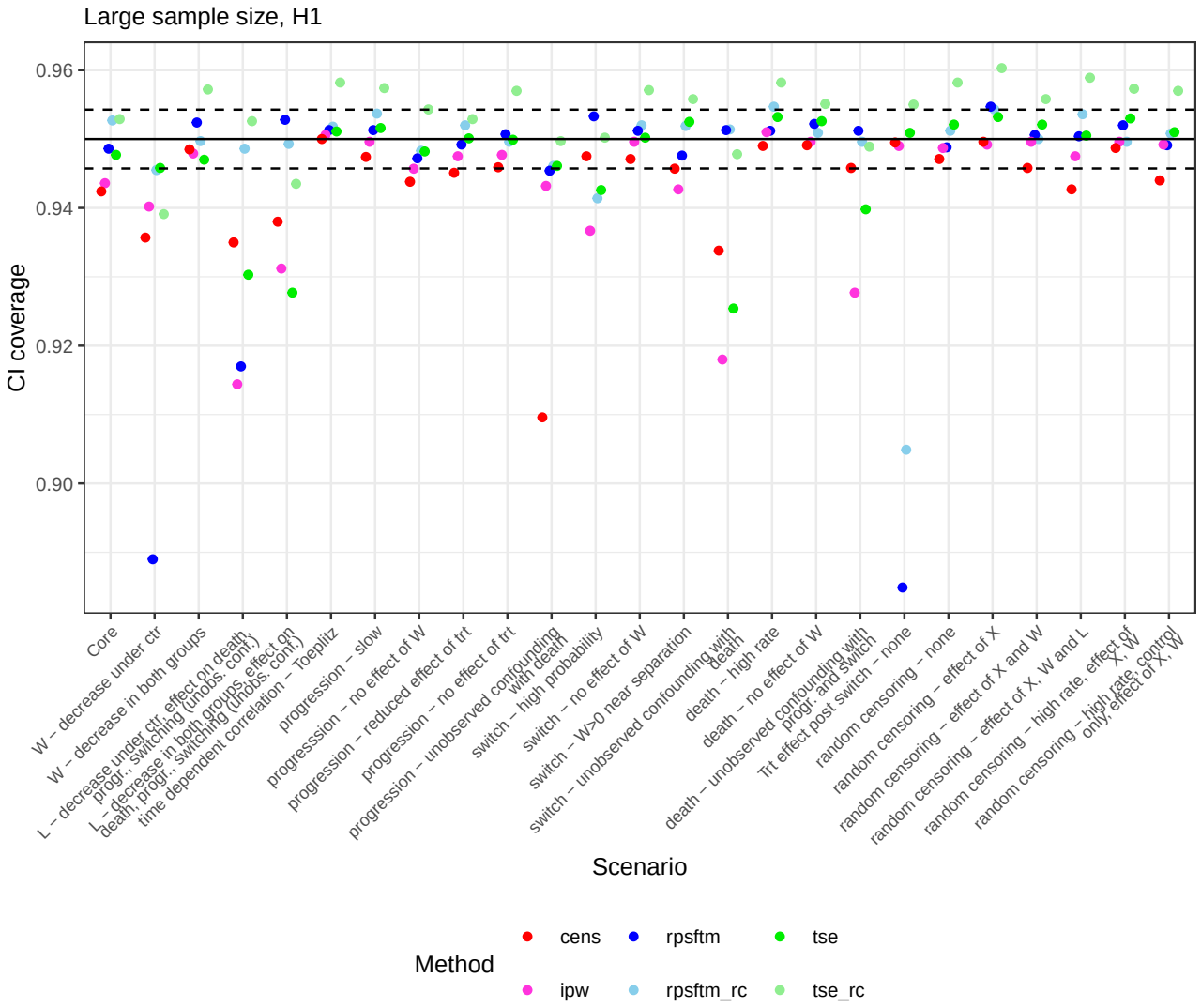


Figure 2.16: Confidence interval coverage of the studied methods in treatment switching scenarios with large sample size under the alternative. The solid reference line corresponds to the nominal 95% level. The dashed reference lines indicates simulation uncertainty. They mark the range, in which the empirical coverage probability should be with 95% probability (by asymptotic approximation) if the true coverage was 95%.

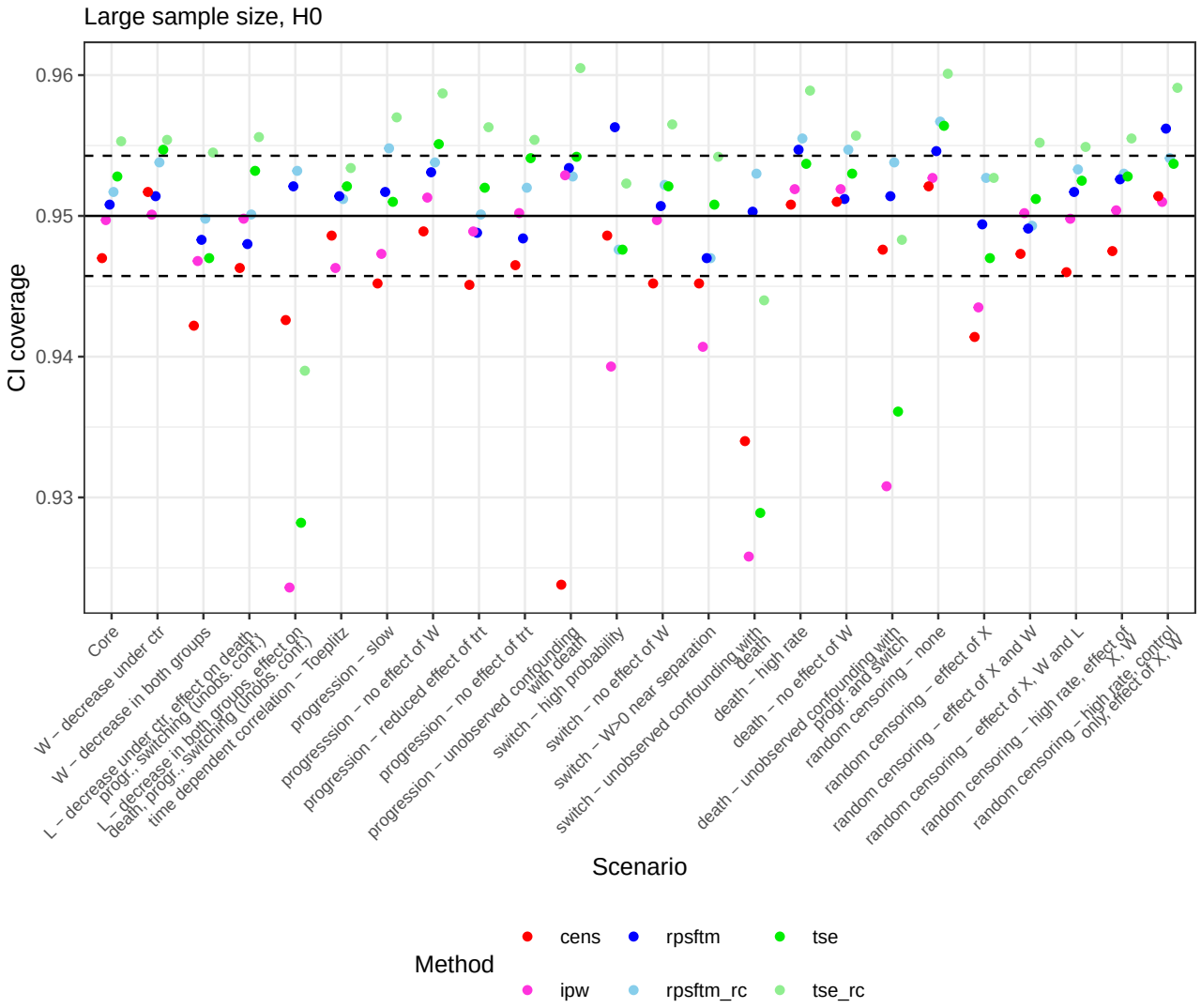


Figure 2.17: Confidence interval coverage of the studied methods in treatment switching scenarios with large sample size under the null hypothesis. The solid reference line corresponds to the nominal 95% level. The dashed reference lines indicates simulation uncertainty. They mark the range, in which the empirical coverage probability should be with 95% probability (by asymptotic approximation) if the true coverage was 95%.

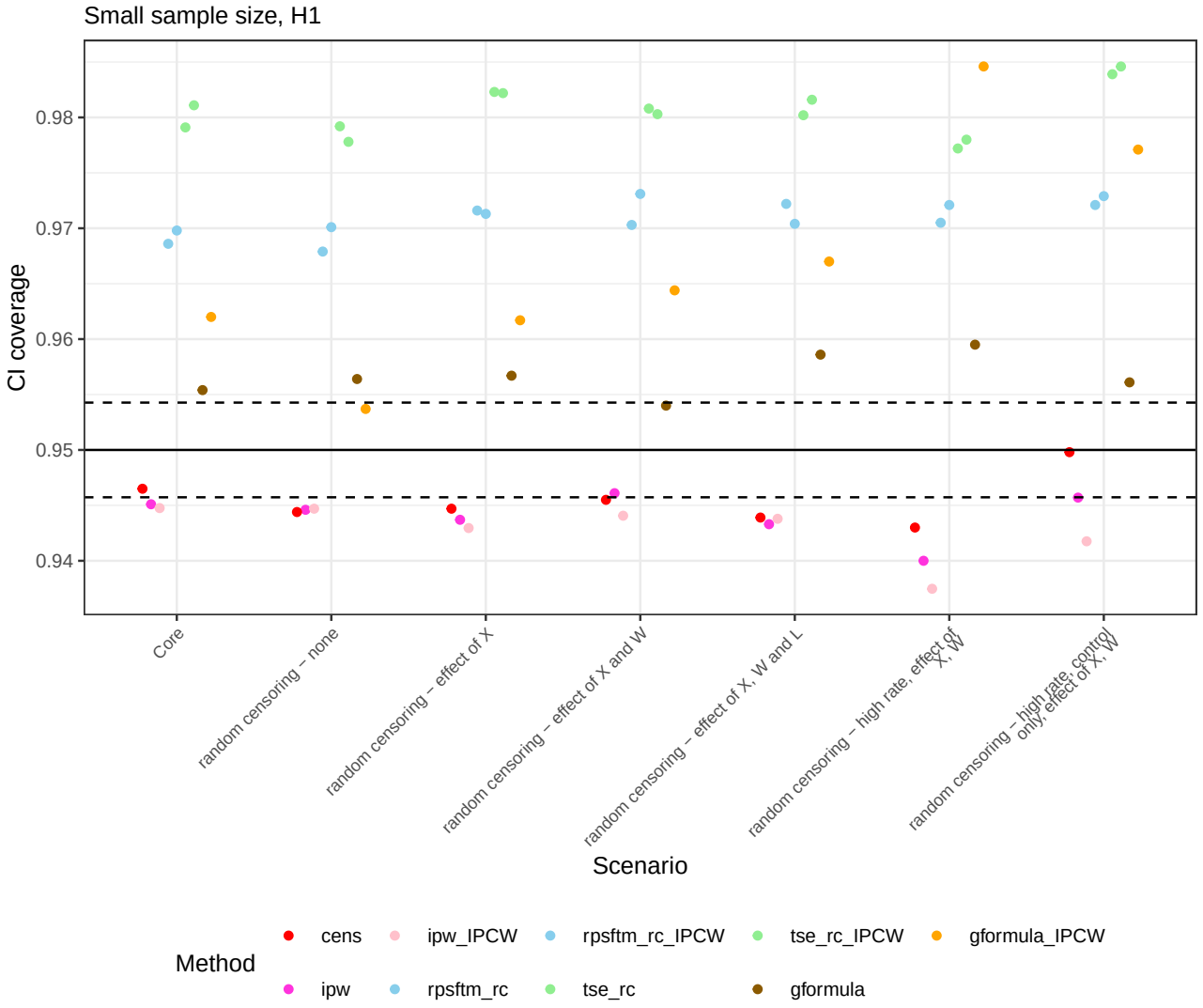


Figure 2.18: Confidence interval coverage of the studied methods with and without additional inverse probability weighting for random censoring in treatment switching scenarios with small sample size under the alternative. The solid reference line corresponds to the nominal 95% level. The dashed reference lines indicates simulation uncertainty. They mark the range, in which the empirical coverage probability should be with 95% probability (by asymptotic approximation) if the true coverage was 95%.

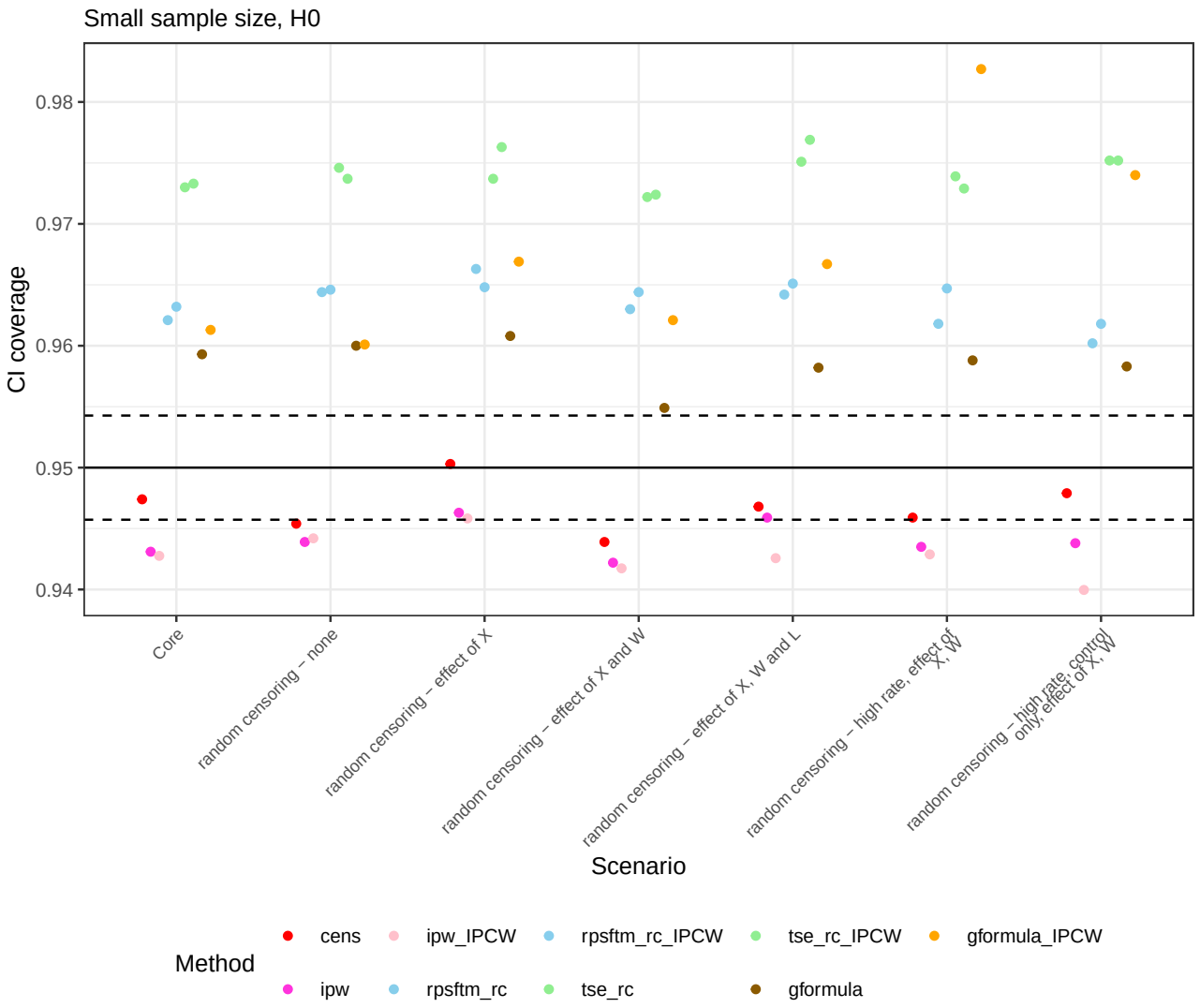


Figure 2.19: Confidence interval coverage of the studied methods with and without additional inverse probability weighting for random censoring in treatment switching scenarios with small sample size under the null hypothesis. The solid reference line corresponds to the nominal 95% level. The dashed reference lines indicates simulation uncertainty. They mark the range, in which the empirical coverage probability should be with 95% probability (by asymptotic approximation) if the true coverage was 95%.

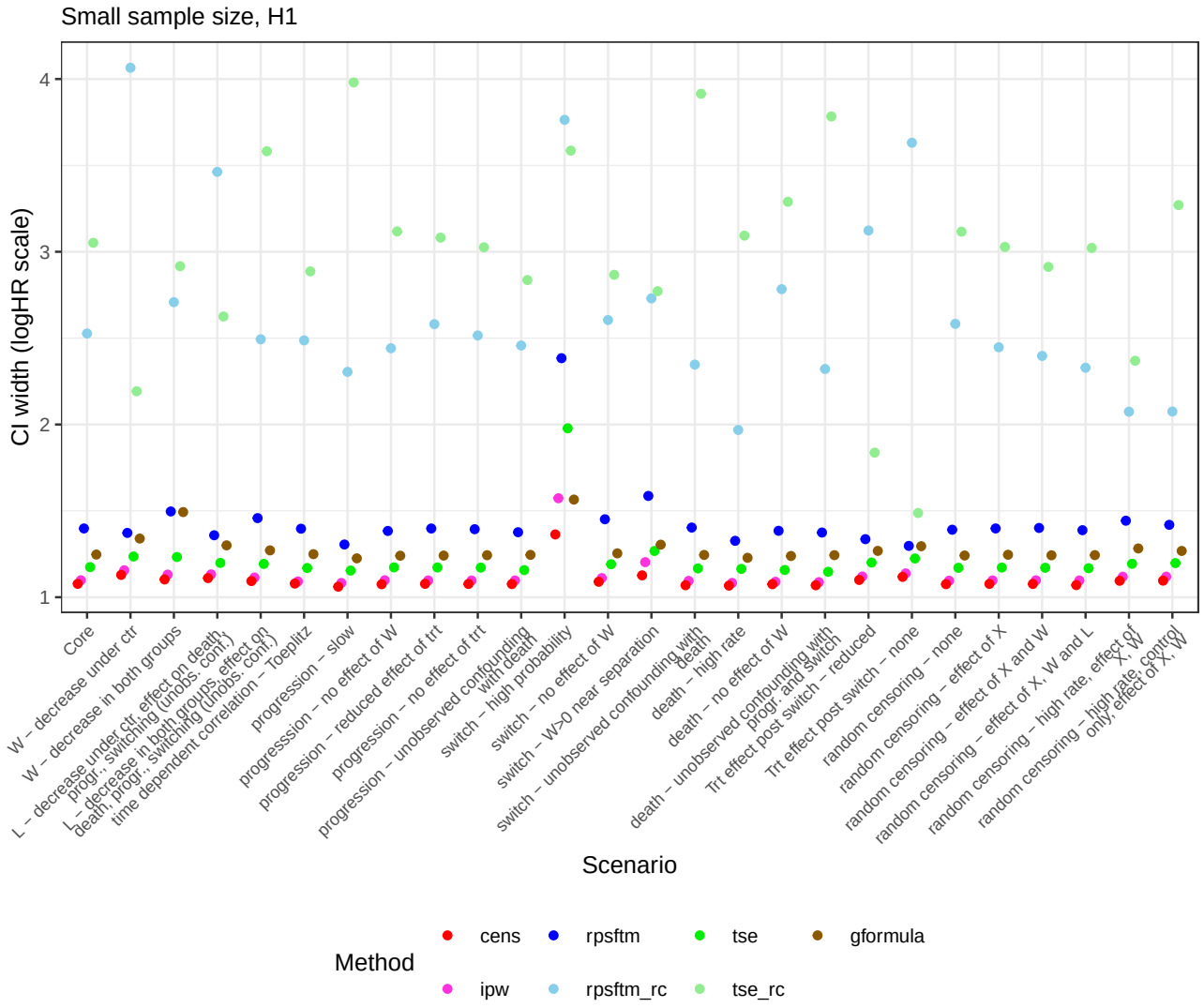


Figure 2.20: Width of confidence intervals for the log hazard ratio in treatment switching scenarios with small sample size under the alternative.

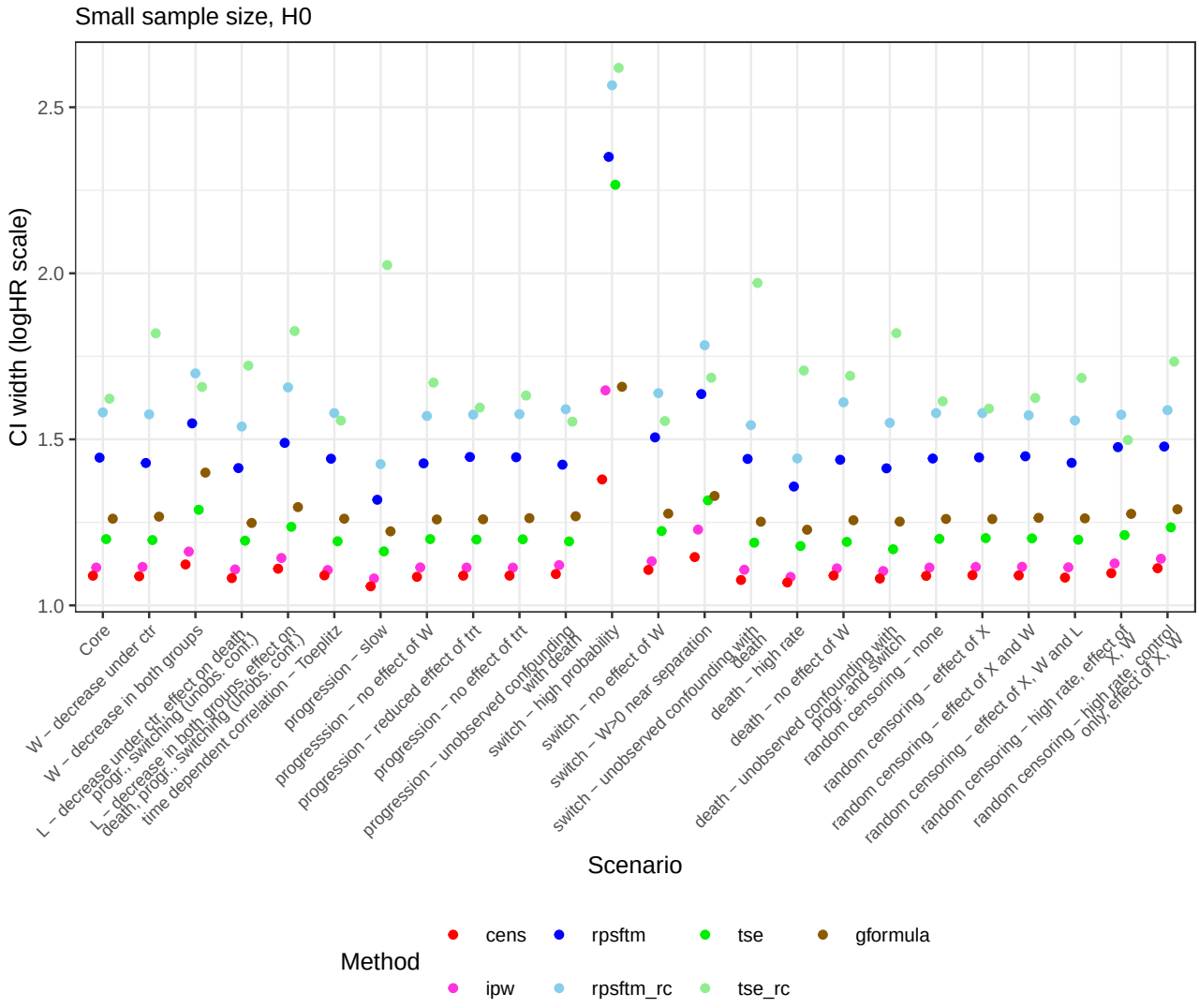


Figure 2.21: Width of confidence intervals for the log hazard ratio in treatment switching scenarios with large sample size under the null hypothesis.

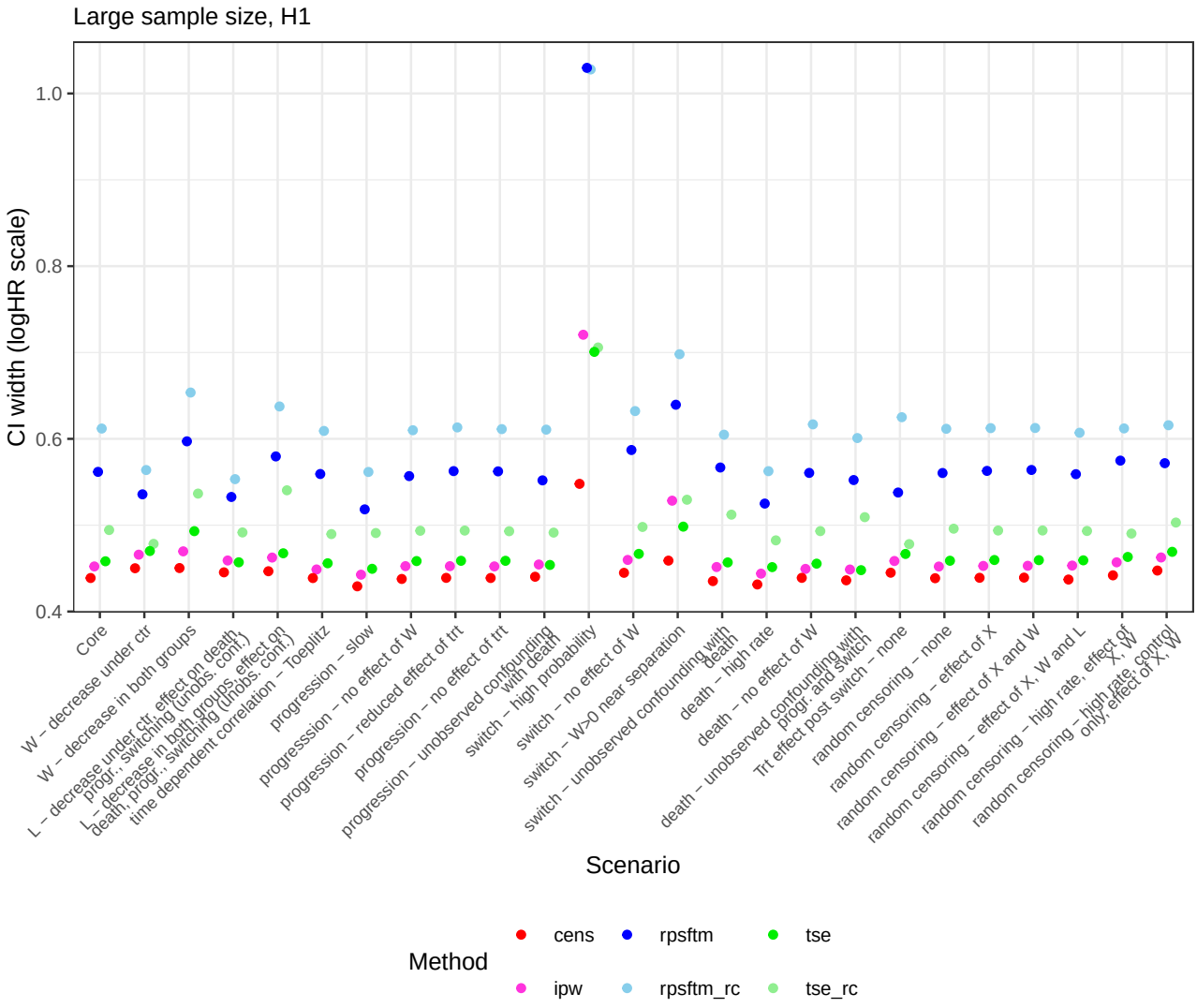


Figure 2.22: Width of confidence intervals for the log hazard ratio in treatment switching scenarios with small sample size under the alternative.

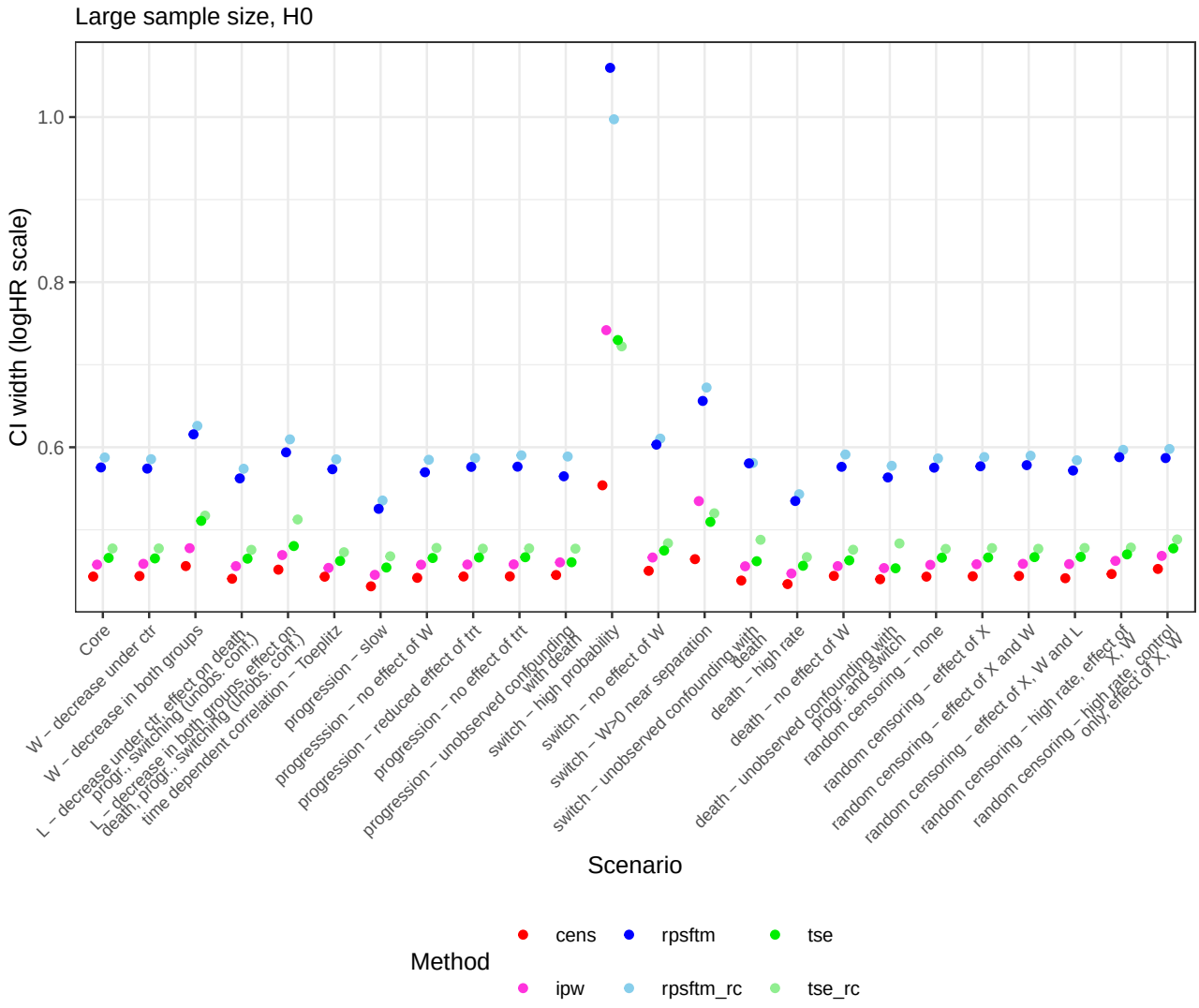


Figure 2.23: Width of confidence intervals for the log hazard ratio in treatment switching scenarios with large sample size under the null hypothesis.

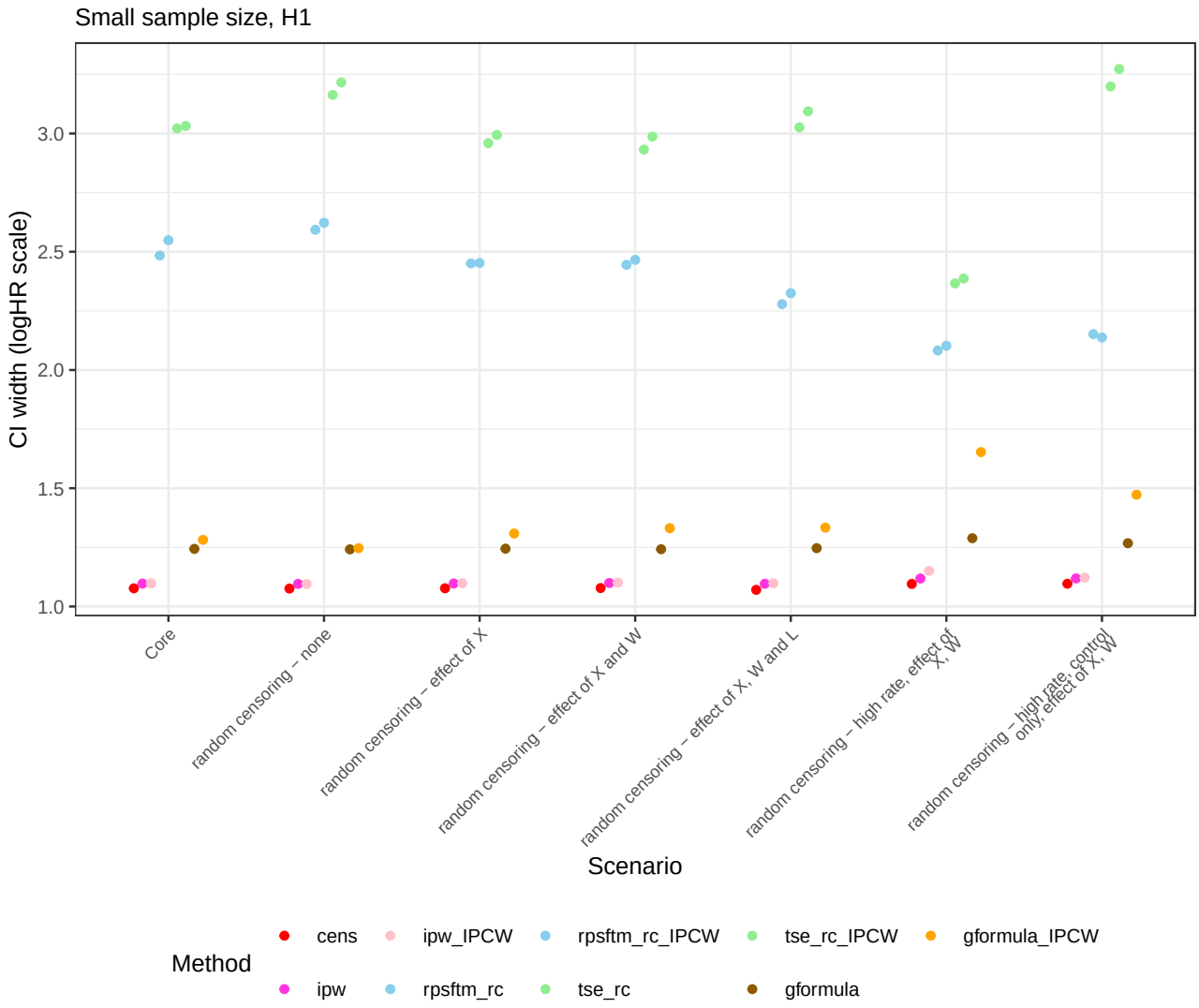


Figure 2.24: Width of confidence intervals for the log hazard ratio obtained through methods with and without additional inverse probability weighting for random censoring in treatment switching scenarios with small sample size under the alternative.

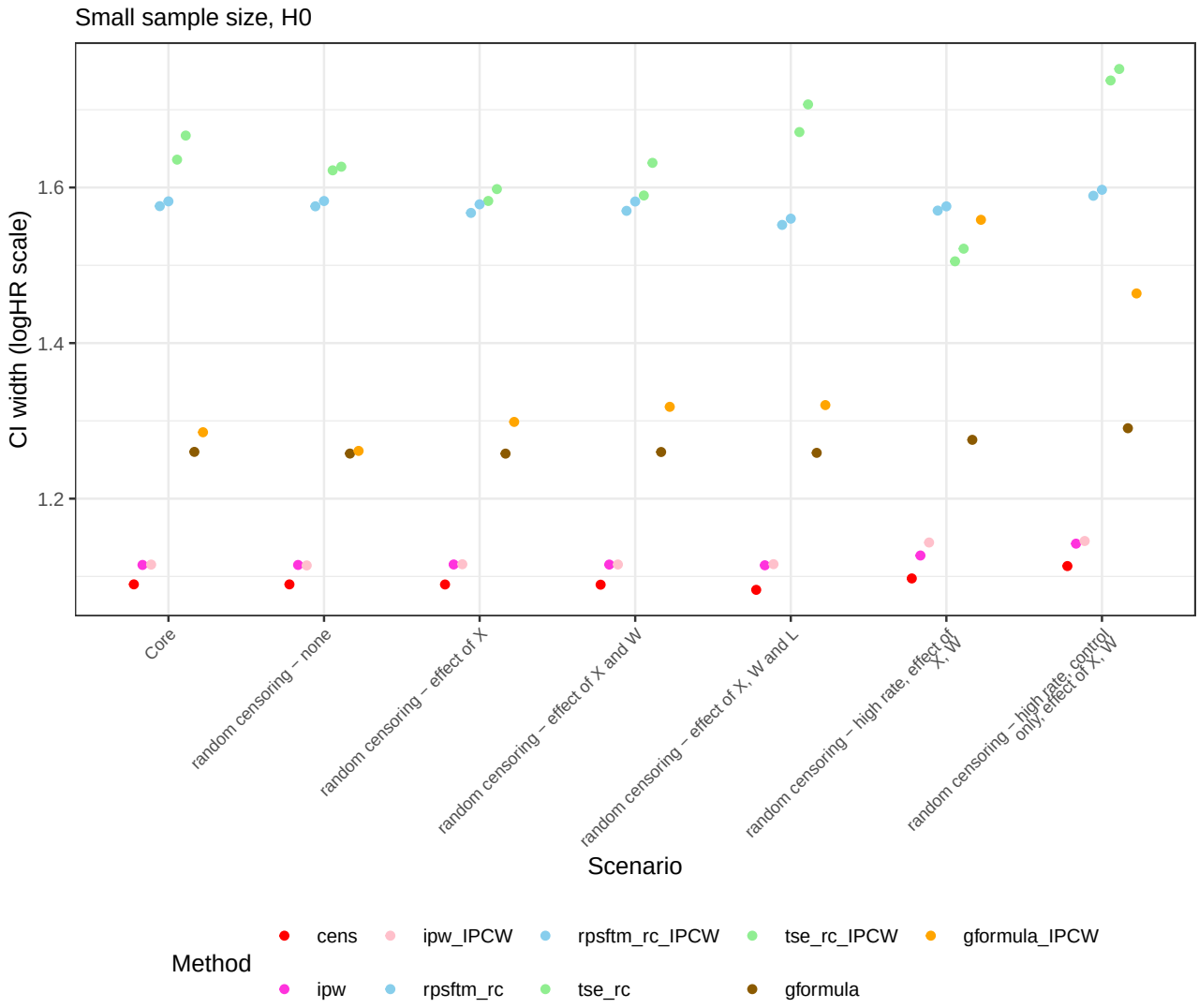


Figure 2.25: Width of confidence intervals for the log hazard ratio obtained through methods with and without additional inverse probability weighting for random censoring in treatment switching scenarios with small sample size under the null hypothesis.

2.4.6 Root mean squared error

The root mean squared error (RMSE) of the log hazard ratio estimates across scenarios and methods is shown in Figures 2.26 to 2.31. The RMSE is in line with findings for the bias and power as well as confidence interval width. Similar to those measures, uncertainty in terms of RMSE was smallest for IPW, censoring and TSE. The g-formula had similar RMSE to these methods. Recensoring increased the RMSE for RPSFTM and TSE and a high switching rate increased the RMSE for all methods. As before, the additional application of IPCW for random censoring had almost no impact on RMSE.

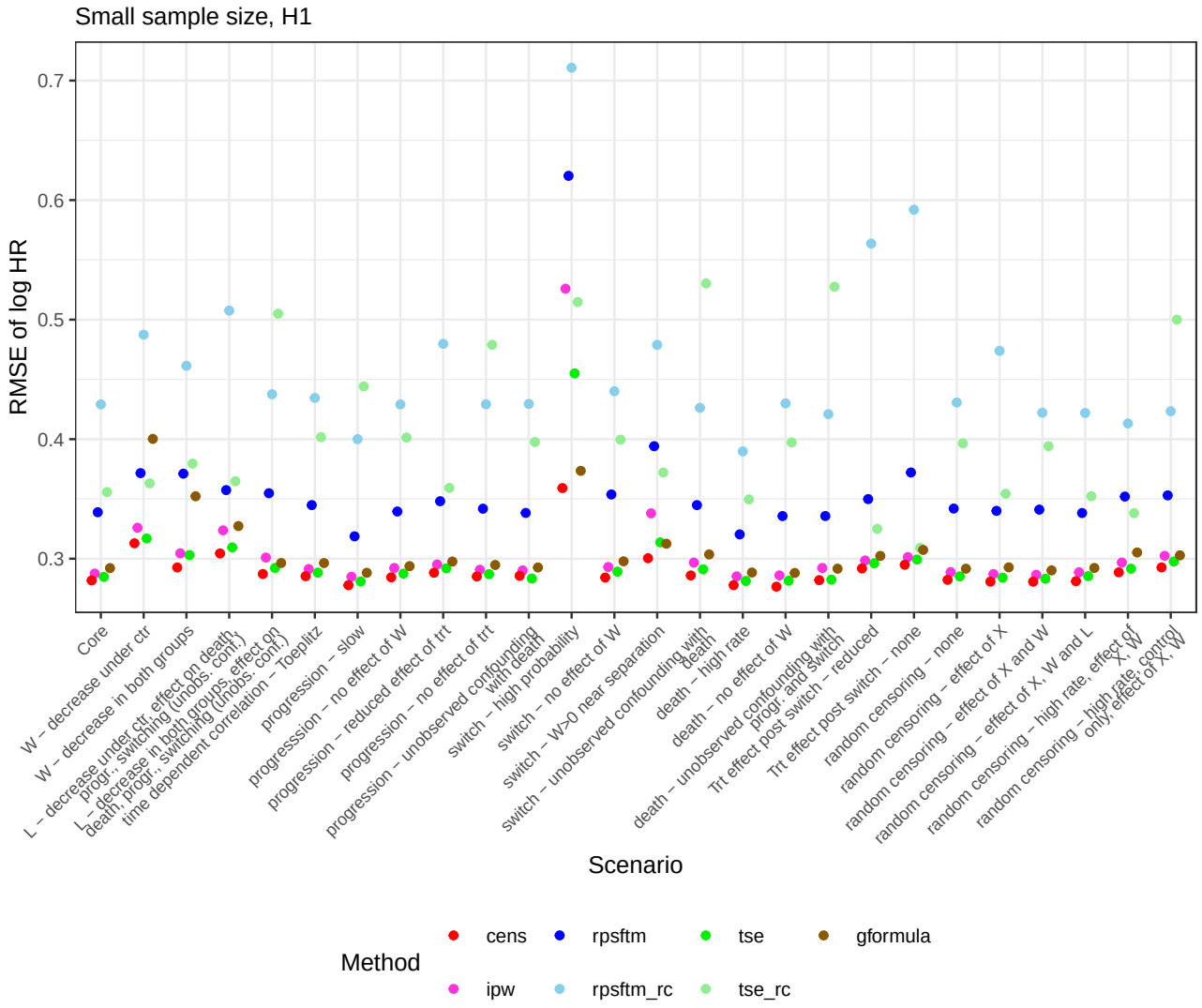


Figure 2.26: Root mean squared error (RMSE) of the estimated log hazard ratio in treatment switching scenarios with small sample size under the alternative.

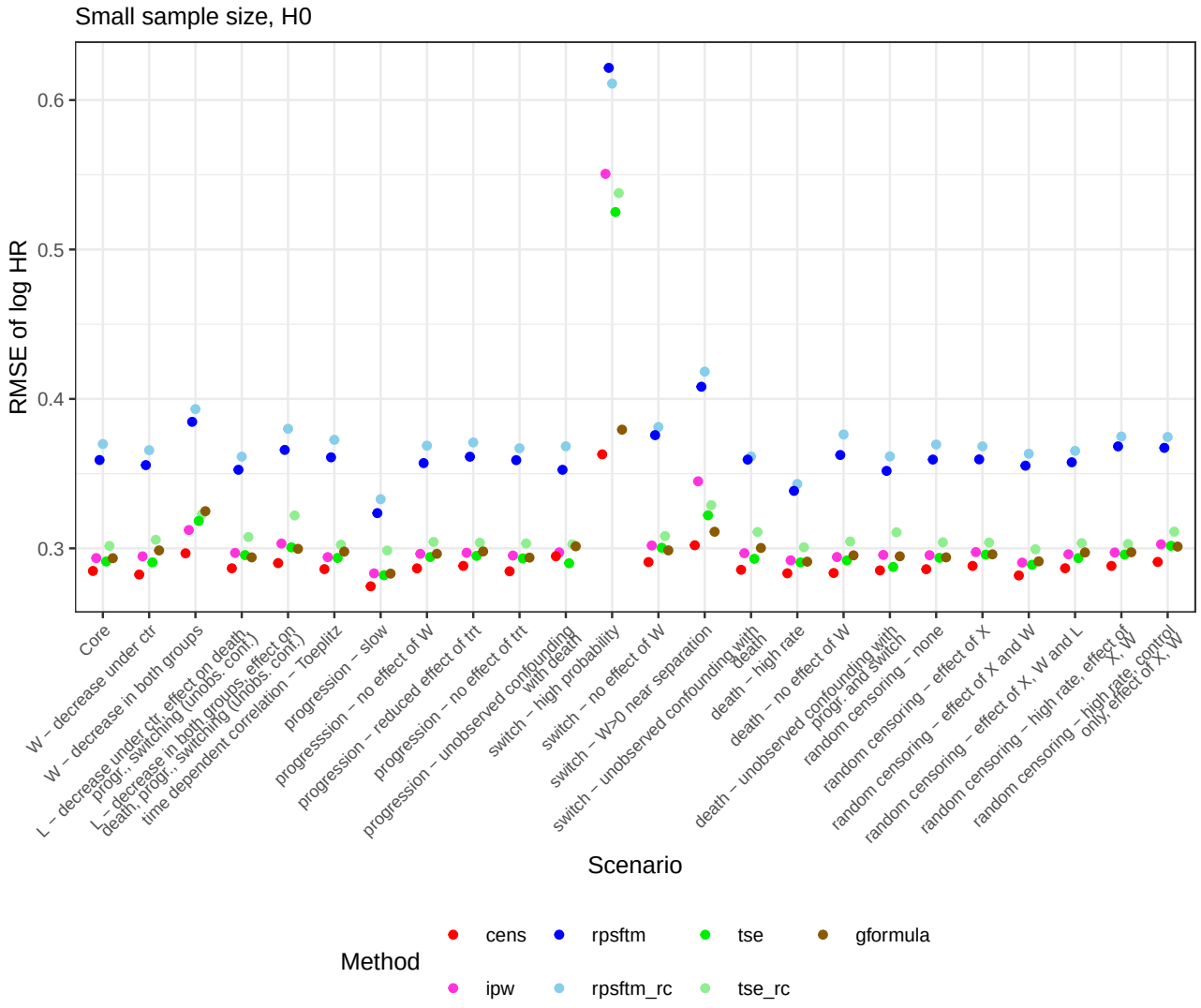


Figure 2.27: Root mean squared error (RMSE) of the estimated log hazard ratio in treatment switching scenarios with small sample size under the null hypothesis.

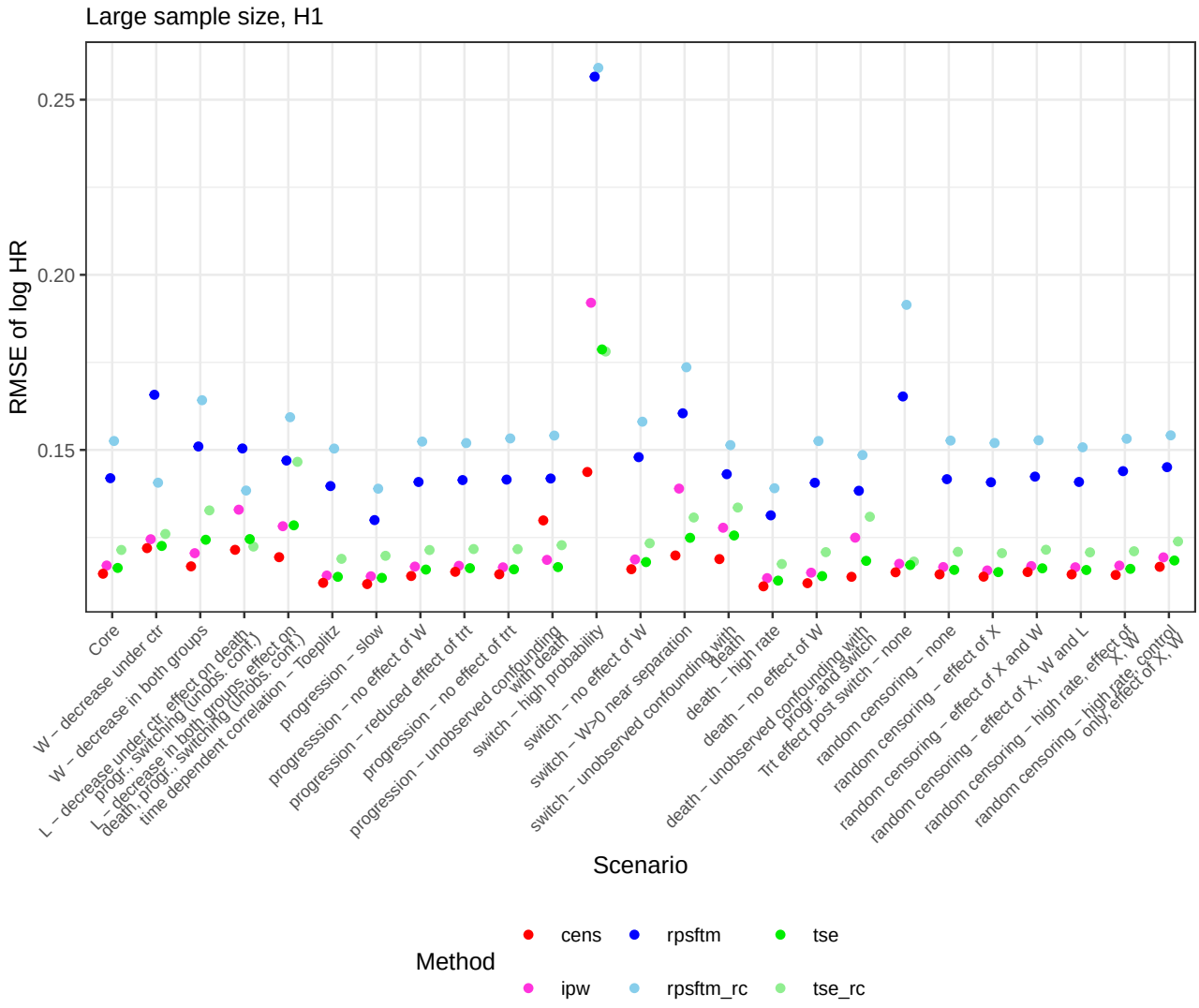


Figure 2.28: Root mean squared error (RMSE) of the estimated log hazard ratio in treatment switching scenarios with large sample size under the alternative.

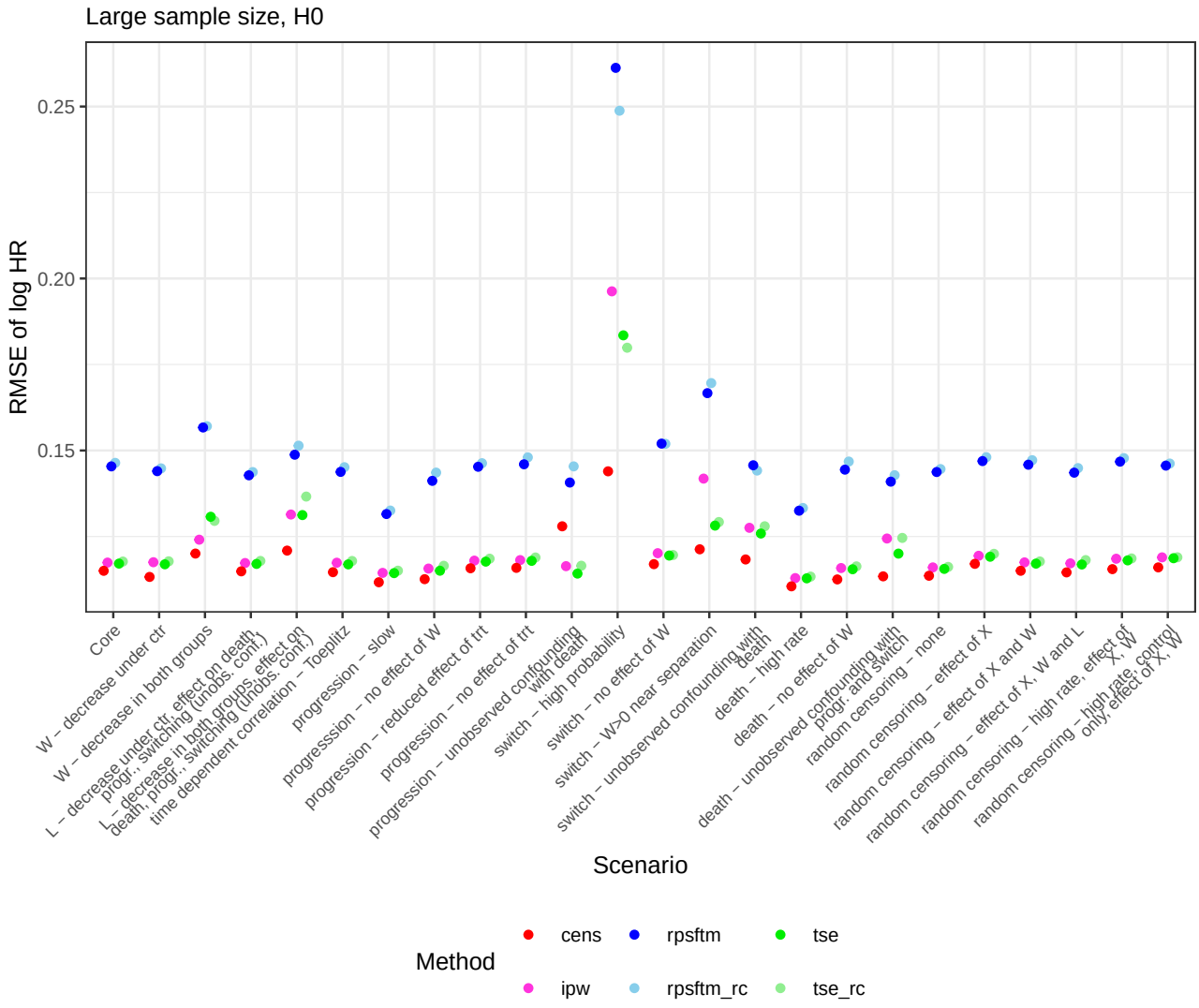


Figure 2.29: Root mean squared error (RMSE) of the estimated log hazard ratio in treatment switching scenarios with large sample size under the null hypothesis.

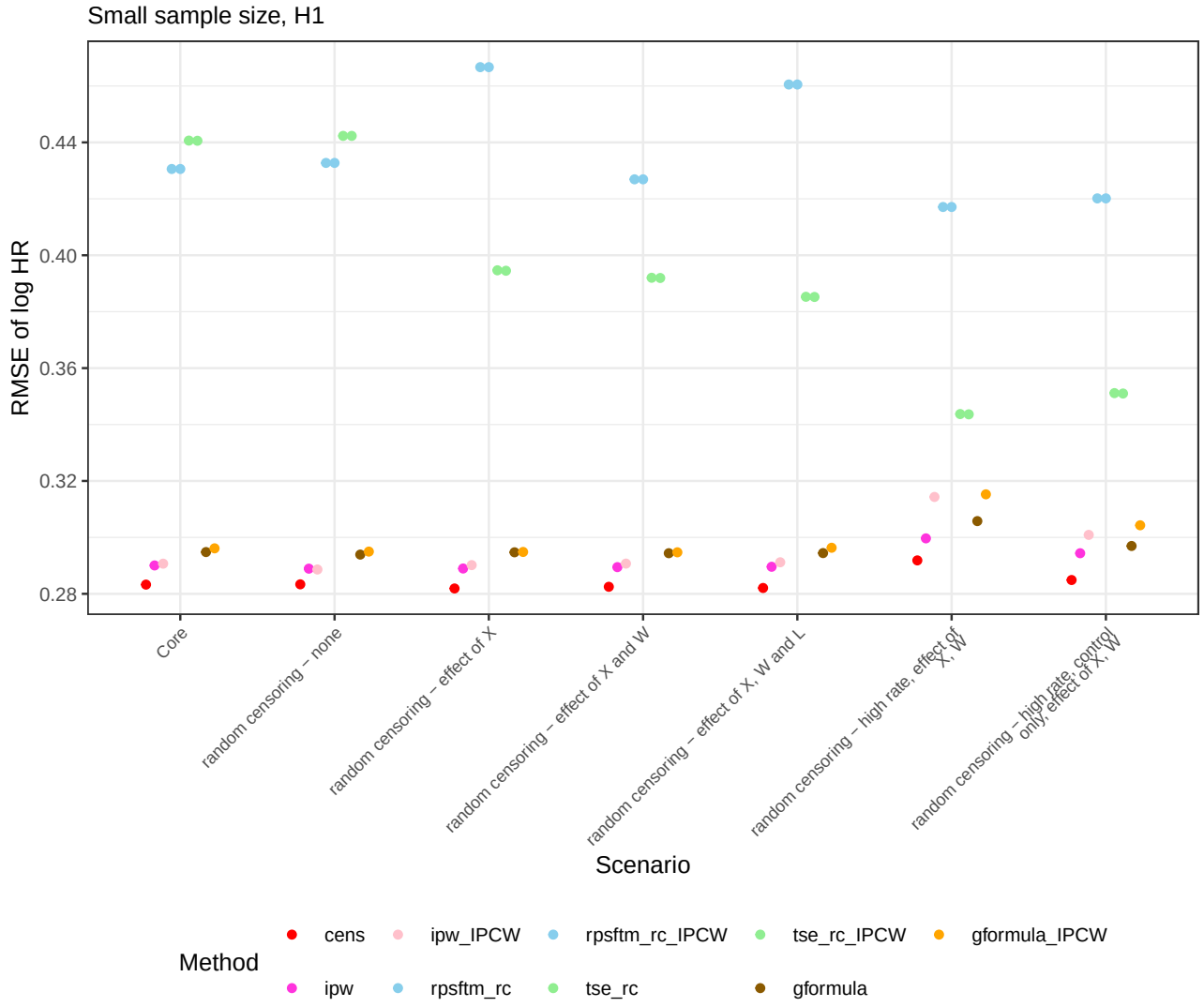


Figure 2.30: Root mean squared error (RMSE) of the log hazard ratio estimated through methods with and without additional inverse probability weighting for random censoring in treatment switching scenarios with small sample size under the alternative.

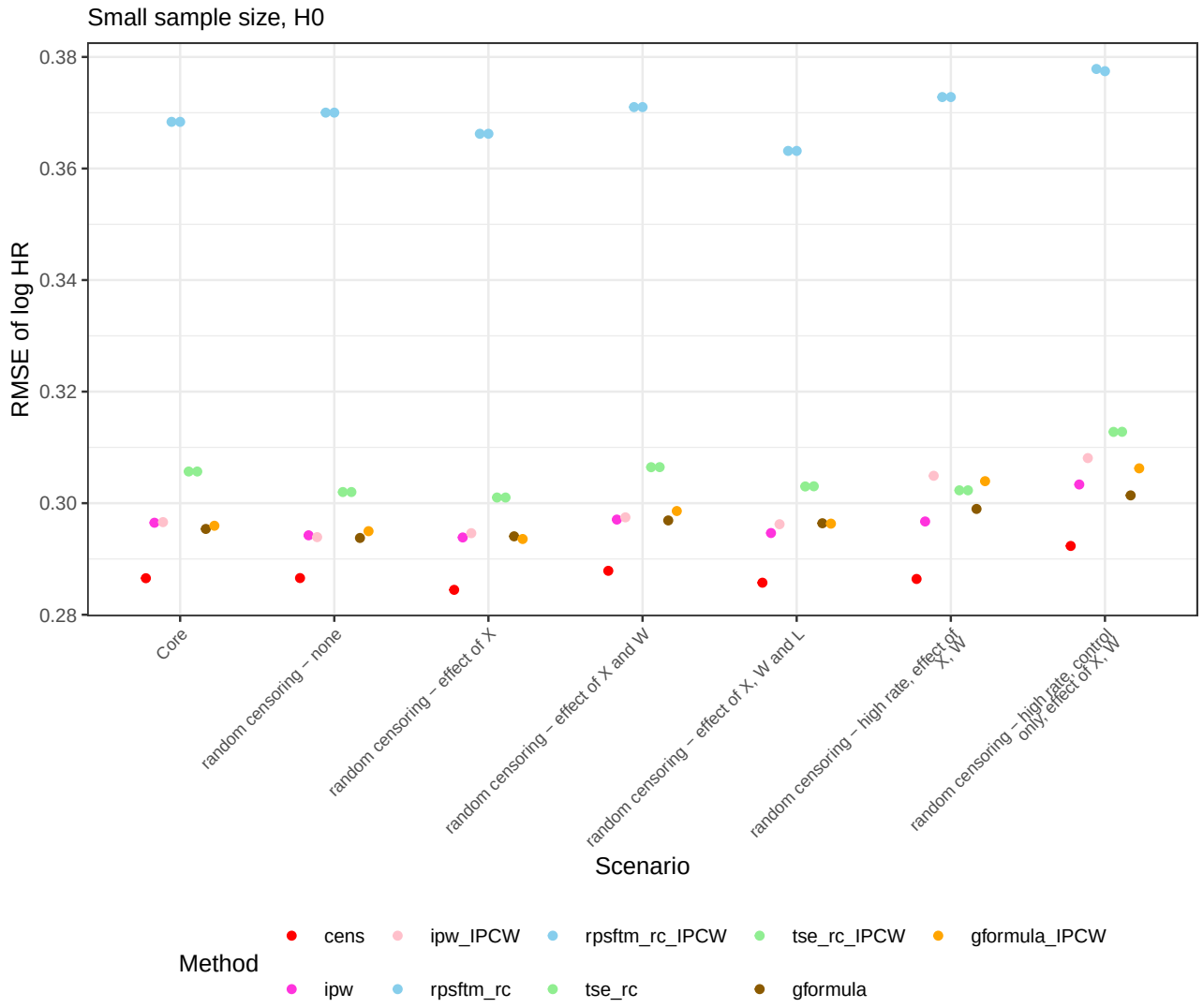


Figure 2.31: Root mean squared error (RMSE) of the log hazard ratio estimated through methods with and without additional inverse probability weighting for random censoring in treatment switching scenarios with small sample size under the null hypothesis.

2.5 Case studies

As case study for a trial with treatment switching and application of the included causal inference, we will study two specific synthetic data sets. The first was simulated under the core scenario, the second was simulated under the scenario in which the time dependent co-variate W has a decreasing time trend in the control group but stays stable under experimental treatment. The two scenarios were chosen to highlight differential behaviour of the methods depending on the underlying data generating mechanisms.

Both data sets were simulated under the alternative hypothesis with a strong treatment effect and the aimed for number of events was 66 on both cases. Other settings were as described in the Section on data generation. The used data sets are included in the supplementary file trt_switch_case_studies.Rdata.

In the case study analyses, all reported p-values are two-sided. Confidence intervals are two sided with nominal coverage 95%.

2.5.1 Case study 1 - core scenario

In this study, 96 patients were included in the control group and 105 patients in the treatment group. Overal 66 events were observed, with 40 events in patients initially assigned to control and 26 events in patients initially assigned to experimental treatment. Thirty-nine (41%) of the control group patients eventually switched to active treatment. The median (min - max) follow up times in this data set were 1.52 (0.03-3.04) years for the control group and 1.64 (0.03-3.05) years for the treatment group.

Table 2.11 shows summary statistics for the covariates X and W at baseline as well as the observed number of death, progression, random censoring and switching events per initial treatment group.

Table 2.11: Case study 1: Descriptive statistics by initial treatment group. Values are mean ± standard deviation or absolute (relative) frequencies.

Variable	Control (n=96)	Treatment (n=105)
X	-0.02 ± 0.87	0 ± 1.04
W	0.13 ± 0.98	0.24 ± 0.9
Death	40 (42%)	26 (25%)
Progression	79 (82%)	63 (60%)
Switching	39 (41%)	0 (0%)
Random censoring	1 (1%)	6 (6%)

Table 2.12 shows summary statistics for the covariates X and W at the time of progression. These statistics in particular serve to illustrate the properties at secondary baseline as used for the IPW and the TSE approach.

Table 2.12: Case study 1: Descriptive statistics for covariates observed at secondary baseline in the control group. Values are mean ± standard deviation.

Variable	Control	Treatment
X at secondary baseline	-0.17 ± 0.85	-0.22 ± 0.94
W at secondary baseline	-0.07 ± 0.96	0.02 ± 0.81

Figure 2.32 shows Kaplan-Meier curves for overall survival in the by initial group allocation. An analysis by a corresponding Cox model for overall survival with treatment group as well as X and W at baseline as covariates results in a hazard ratio (95% confidence interval) of 0.590 (0.359 to 0.969) with p=0.0373. These analysis corresponds to a treatment policy estimand and is included as a typical analysis aiming for a hypothetical estimand would likely in addition include an analysis targeting treatment policy.

The main aim is to estimate a hypothetical estimand under the assumption that no switching was possible. The following methods will be show cased: Inverse probability weighting (IPW), rank-preserving structural failure time model (RPSFTM), the simplified two-stage estimatio (TSE) and parametric g-formula. In addition, as

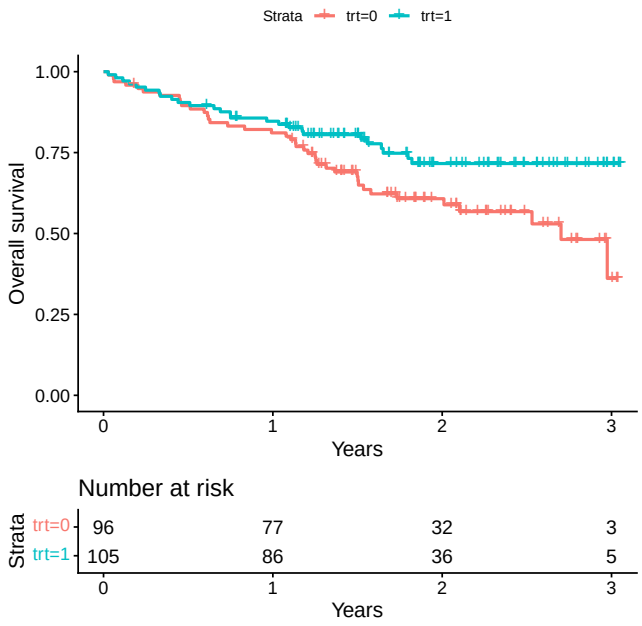


Figure 2.32: Case study 1: Kaplan-Meier estimation of overall survival by treatment group, corresponding to a treatment policy estimand strategy.

comparator method, a Cox model will be applied in which event times are censored at the time of switching. RPSFTM and TSE were performed without recensoring and with 1000 bootstrap samples for inference. G-formula was applied as described in Section 2.2. The number of bootstrap samples for RPSFTM, TSE and the g-formula was set to 1000.

Appendix A includes exemplary R code to fit the according analysis models.

The results of these analyses are shown in Table 2.13. The results of the applied methods are relatively homogeneous, with estimated hazard ratios close to 0.5 for all methods and comparable confidence intervals. The respective hypothesis test allow for rejection of the null hypothesis of no treatment effect in the hypothetical setting for all methods. RPSFTM here has the largest p-value of 0.0303 and the IPW method had the smallest p-value of 0.0084. Note that for all methods, the estimated effect for the hypothetical estimand is stronger than the effect estimated under the treatment policy strategy.

Table 2.13: Case study 1: Estimated hazard ratio (HR) and according two-sided 95% lower and upper confidence interval boundaries.

Method	HR	low	up	p
IPW	0.493	0.291	0.834	0.0084
RPSFTM	0.493	0.260	0.935	0.0303
TSE	0.485	0.280	0.840	0.0098
g-formula	0.484	0.271	0.863	0.0167
Censor after switching	0.530	0.311	0.902	0.0192

Figure 2.33 For further illustration, survival distributions were estimated based on the assessed analysis methods. For inverse probability weighting (IPW), survival times for control group patients who switched to active treatment were censored at the time of switching and non-switchers were up-weighted in the Kaplan-Meier estimation by the inverse probability of not switching. For the rank-preserving structural failure time model (RPSFTM) and the simplified two-stage estimation (TSE) counterfactual survival times calculated by the two methods were used to estimate Kaplan-Meier curves for the control group. For the g-formula approach, survival

times sampled from the fitted g-formula models were used for both groups. The resulting curves are shown in Figure 2.33. The different approaches result in relatively similar curves. Of note, the curves obtained through the g-formula approach are discrete at regular intervals due to the discretisation into 1 month intervals in this method. Compared to the survival curves estimated under the treatment policy estimand, the curves estimated under the hypothetical strategy showed differences between the treatment groups.

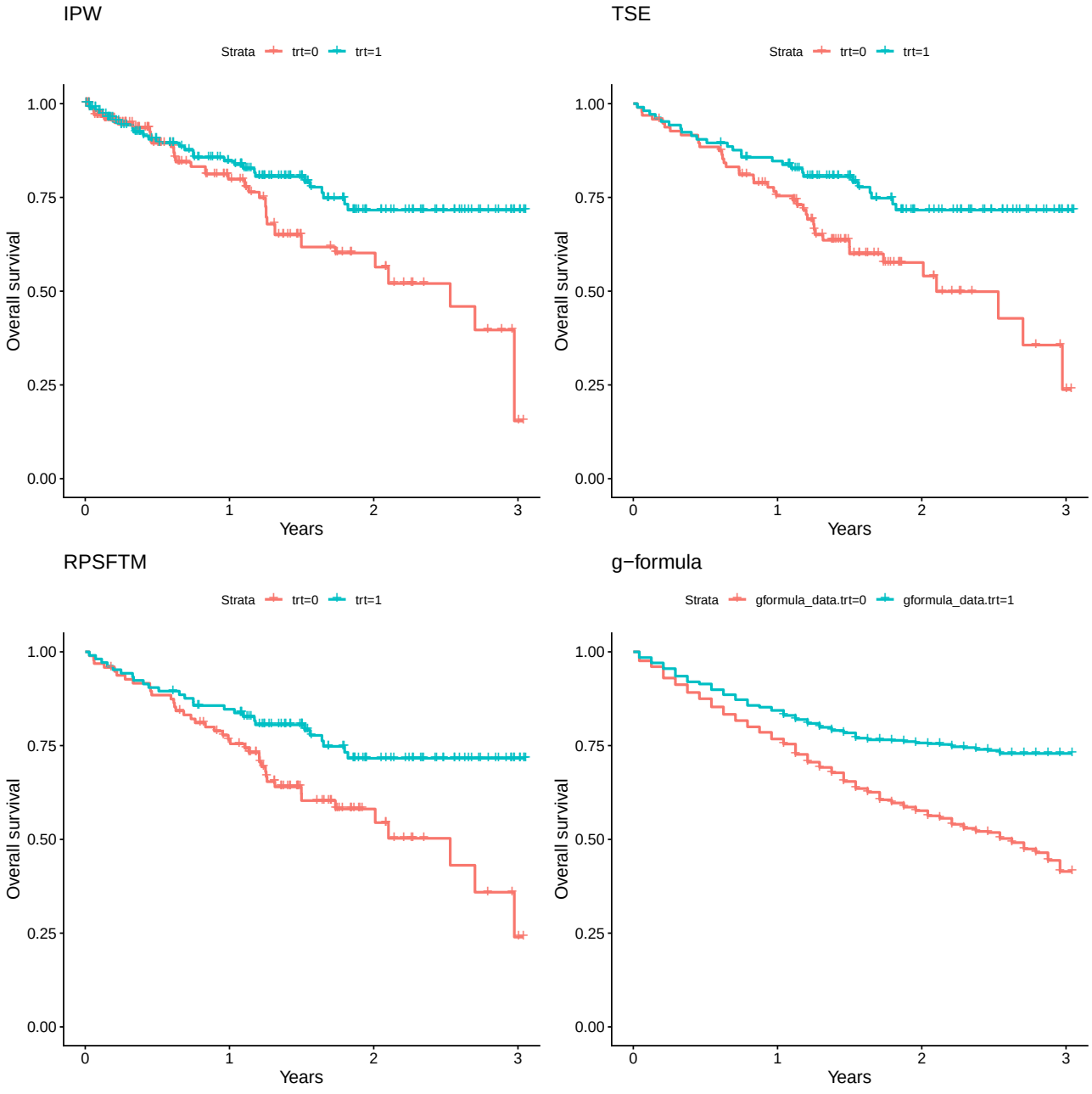


Figure 2.33: Case study 1: Survival distributions estimated under a hypothetical estimand strategy assuming no switching between treatment groups was possible, using different causal inference methods.

2.5.2 Case study 2 - differential pattern of time dependent covariate

In the second case study, 100 patients were included in the control group and 103 patients in the treatment group. Overall, 66 events were observed, with 44 events in patients initially assigned to control and 22 events in patients initially assigned to experimental treatment. In the control group, 35 (35%) patients switched to active treatment. The median (min - max) follow up times in the control and treatment group were 2.42 (0.02 - 4.19) and 2.81 (0.14 - 4.19), respectively.

Summary statistics for the study data are shown in Table 2.14.

Table 2.14: Case study 2: Descriptive statistics by initial treatment group. Values are mean ± standard deviation or absolute (relative) frequencies.

Variable	Control (n=100)	Treatment (n=103)
X	0.07 ± 1.1	-0.07 ± 1.03
W	0.44 ± 1.04	0.63 ± 0.94
Death	44 (44%)	22 (21%)
Progression	80 (80%)	74 (72%)
Switching	35 (35%)	0 (0%)
Random censoring	5 (5%)	7 (7%)

Table 2.15 shows summary statistics for the covariates X and W at the time of progression (secondary baseline). In contrast to case study 1, a differential pattern between treatment groups at the time of progression is visible. Patients under control who progressed had on average lower values of W, which is expected as the time dependent variable W is decreasing under control but not under treatment. Of note, larger values of X and W are associated with lower hazard for progression as well as for death. Patients under control who progressed had larger (better) values for X compared to the treatment group. This is likely explained as the patients who progressed are a selected subset of patients with less favourable covariate profile. In this scenario, treatment group patients have on average a favourable and stable profile of W, thus progression is likely mostly caused by low values of X. In contrast, control group patients have a declining profile for W, thus combinations of X and W that in sum are less favourable suffice to trigger progression.

Table 2.15: Case study 2: Descriptive statistics for covariates observed at secondary baseline in the control group. Values are mean ± standard deviation.

Variable	Control	Treatment
X	0.02 ± 1.06	-0.32 ± 0.91
W	0.02 ± 1.02	0.30 ± 0.98

Figure 2.34 shows Kaplan-Meier curves for overall survival in the by initial group allocation. An analysis under the treatment policy estimand results in a hazard ratio of 0.406 (95% confidence interval 0.242 to 0.678) with a p value of 0.0006.

The same methods as in case study 1 were applied to to estimate a hypothetical estimand under the assumption that no switching was possible. The results are shown in Table 2.16. In contrast to case study 1, there is more heterogeneity in the analysis results of different methods. RPSFTM and TSE estimated quite similar values for the hazard ratio of 0.312 and 0.315. IPW and The censoring comparator approach and IPW provided estimates slightly closer to 1 with values of 0.334 and 0.348, respectively. The estimate via g-formula was quite different with a hazard ratio of 0.421. This value was even closer to 1 than the hazard ratio of 0.406 obtained under the treatment policy estimand. resulted in a less pronounced effect estimate with a value. This is, however, in line with the simulation results which showed that the g-formula for the studied type of scenario results in biased estimates. In the given case study, the qualitative interpretation of the result does not change depending on the analysis method. The confidence intervals of all methods strongly overlap and all methods result in two-sided p-values smaller than 0.05, even though the p-value for the g-formula is considerably larger (0.0202) than the others, which are in the range of 0.0001 to 0.0002.

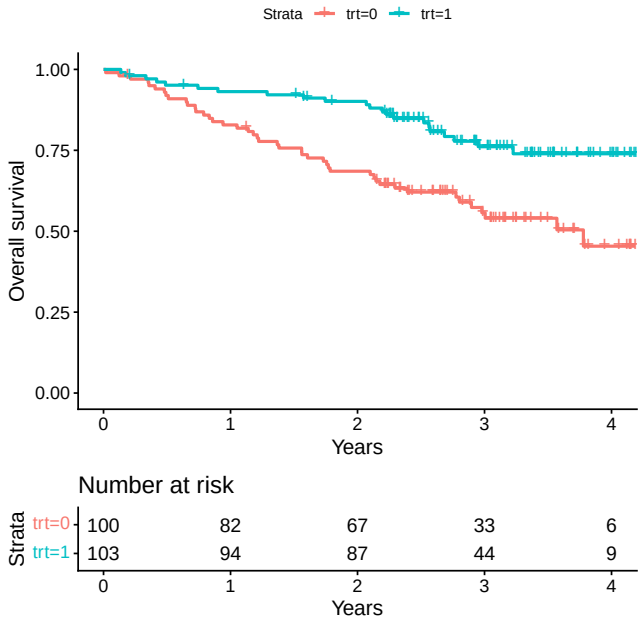


Figure 2.34: Case study 2: Kaplan-Meier estimation of overall survival by treatment group, corresponding to a treatment policy estimand strategy.

Table 2.16: Case study 2: Estimated hazard ratio (HR) and according two-sided 95% lower and upper confidence interval boundaries.

Method	HR	low	up	p
IPW	0.348	0.203	0.599	0.0001
RPSFTM	0.312	0.170	0.572	0.0002
TSE	0.315	0.179	0.555	0.0001
g-formula	0.421	0.206	0.860	0.0202
Censor after switching	0.334	0.195	0.571	0.0001

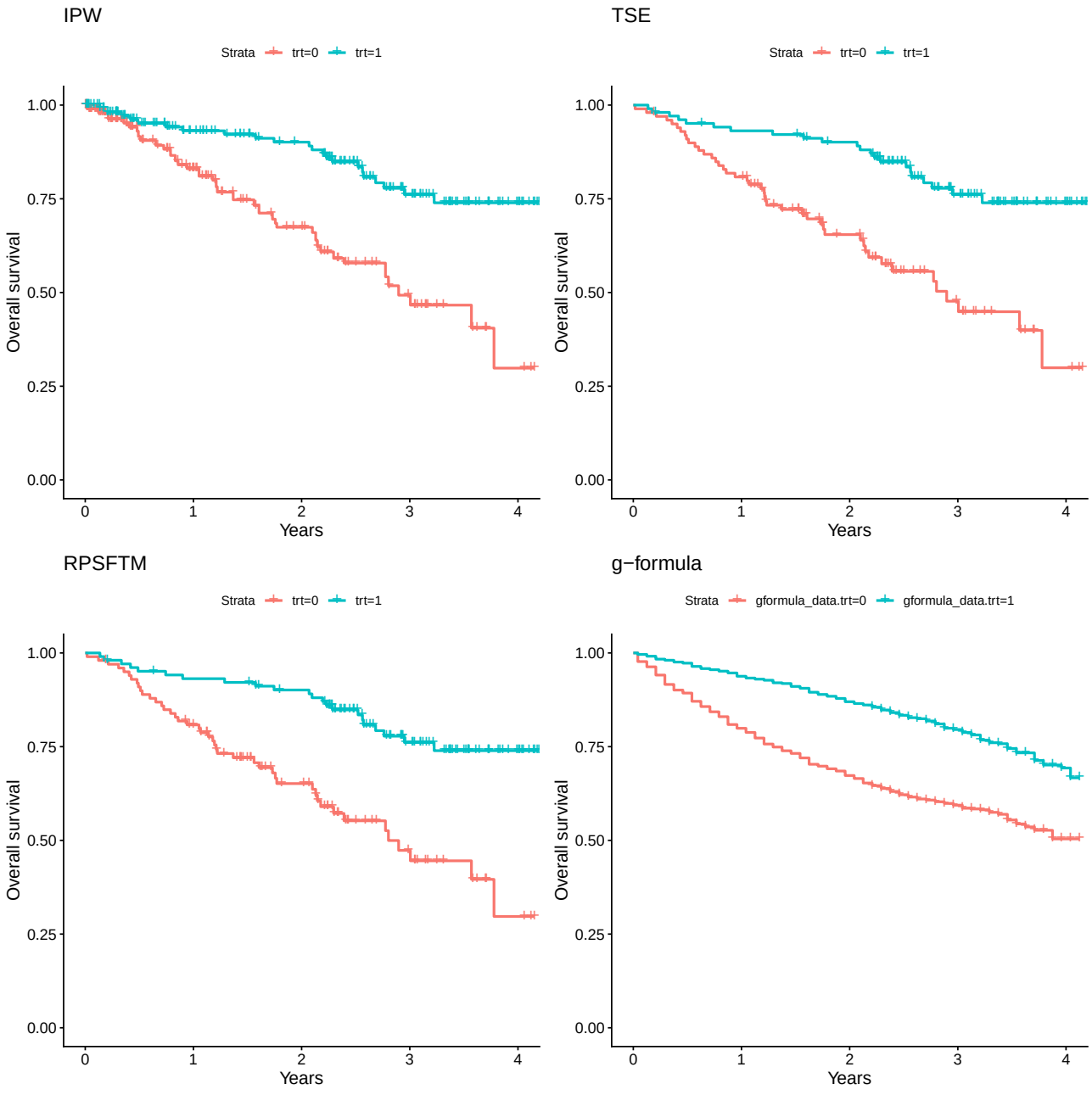


Figure 2.35: Case study 2: Survival distributions estimated under a hypothetical estimand strategy assuming no switching between treatment groups was possible, using different causal inference methods.

2.6 Discussion

Overall, the simulation results suggest that the studied causal inference methods are in principle applicable to account for treatment switching and estimate a treatment effect under the hypothetical estimand strategy. However, all studied methods were found to be sensitive with respect to their inherent model assumptions and violations of these assumptions may occur in usual study designs as was shown by different simulation scenarios.

IPW and the basic comparator using censoring after switching performed very similar, and both were most powerful. Note that the IPW approach is related to the censoring comparator, as it uses the same outcome model but adds weights to account for the switching propensity. Therefore, though, the IPW estimate has additional variability resulting from using estimated weights (which is accounted for by the robust variance estimation) and thus may need larger sample size to maintain the desired power. Both methods were affected by unobserved confounding. Among the studied scenarios, there was only one scenario where IPW clearly improved the censoring approach, that was unobserved confounding between progression and death. For the censoring approach, this setting implies confounding between switching and death, and results in a strong bias. IPW correctly accommodates for switching, as it is able to correctly estimate the switching probability conditional on progression having occurred, and whether progression in itself was confounded with death has no impact. However, in other settings where confounding extended towards the switching decision itself, IPW often performed worse than mere censoring.

TSE was also a powerful method, however equally affected by unobserved confounding as IPW.

RPSFTM was in general found to be the least favourable approach. Its model assumption of common treatment effect was found to be restrictive as violation of this assumption had much stronger impact than violations of assumptions for other methods. Also, its estimator seems to be the least efficient among the studied methods with power values far below those of other methods.

Recensoring for RPSFTM and TSE has been proposed to account for possible bias resulting from an association of calculated counterfactual times and censoring (White et al., 1999). However, in the simulation study, recensored variants of RPSFTM and TSE were found to have considerably less power than the unmodified variants while there was no obvious improvement of bias.

The g-formula is in principle a flexible approach though its implementation is less straightforward then for other methods. The, relatively simple, models considered in the g-formula implementation used here resulted in a method that was comparable in precision (especially in terms of RMSE) to IPW, censoring and TSE. However, its bias tended to by slightly larger and its power somewhat lower. Most strikingly, the g-formula failed in the setting where the observed time dependent covariate had a different time trend under control than under experimental treatment, with large bias and severe undercoverage, below 90%, of the nominal 95% confidence interval. As the g-formula models were fit separately per treatment group, this points to insufficient model specifications for this scenario. Possibly, by choosing more flexible models, these shortcomings could be remedied, however, in a real application the model complexity may be limited by the amount of available data and identifying the required complexity during the planning phase may be challenging. G-formula also is much more computation intense than the other methods, which may, however, be less limiting if only a single analysis is performed.

A small though significant was observed for IPW in the scenario where the time dependent covariate W had a declining time trend under control but stable mean under treatment (scenario "W - decrease under ctr"). Of note, this scenario did not include unobserved confounders. Instead, the bias could be due to patients who reach disease progression at different follow-up times having systematically different covariate profiles at the time of switching and different subsequent hazard functions. Simply put, when patients switch treatment, they should be implicitly replaced, from that time on, by similar patients who did not switch. However, since the IPW model does not take into account the time of switching, the "replacement" is a mixture of patients who experienced disease progression at a different time-points. While these patients may have a matching probability to switch, their future hazard for death may be different, depending on the current values of X and W(t) and the future trajectory of W(t). A possibly remedy could be to calculate weights not only depending on switching at the time of progression but to include the time-dependent probability to progress in the model for the weights. This mechanism warrants further exploration, as it may impact the validity of IPW even in settings where potential

confounders habe been observed.

The causal inference methods, more so than the censoring comparator, were found to be affected by too small sample sizes. It should thus also be kept in mind that these methods may either need some additional small sample adjustments, e.g. when estimating the robust variance for IPW, or should in general be applied to sufficiently large data sets.

Of note, the presented results should be interpreted within the context of the specific estimands and data-generating mechanisms. Generalisation to other settings may require additional investigations.

In conclusion, while correct estimation thus was found to be possible under appropriate assumptions, it will remain challenging to ensure that, mostly untestable, assumptions are met in an actual application. Future research may focus on developing sensitivity analysis to support the application of the studied methods.

3 Scenario class 2: Rescue medication, treatment policy and hypothetical strategy, continuous endpoint

Administration of rescue medication is a frequently observed intercurrent event in clinical trials. As a guiding example we assume a trial similar to Olarte Parra et al. (2025), comparing an experimental medication for Diabetes to a control with 1-year change in HbA1c as primary endpoint. In this example, insulin may be applied as rescue medication if blood glucose levels become too high. We further assume that primary endpoint data may be missing, and the probability for missing data may be larger following rescue medication (corresponding to a scenario of selective drop out). Furthermore, the HbA1c value is measured at regular intervals between baseline and the final assessment.

We consider both a treatment policy and a hypothetical strategy to address this intercurrent event:

Estimand with treatment policy strategy: The estimand of interest is the difference in means of one-year change from baseline in HbA1c regardless of the intake of rescue medication or discontinuation of treatment. The considered estimand has the following relevant attributes:

- Endpoint: Change in HbA1c
- Treatments: Experimental versus control
- Summary measure: Difference in means
- Population: Patients with type-2 diabetes
- Intercurrent events: Use of rescue medication and treatment discontinuation
- Intercurrent event strategy: Treatment policy strategy, ignoring the fact that some patients initiate rescue medication or discontinue treatment.

Under treatment policy, all the available data are included in the analysis and the focus of the study is on handling missing outcome data. The causal inference analysis method is inverse probability weighting. Conventional comparator methods are mixed models for repeated measures and multiple imputation conditional on rescue medication status.

Estimand with hypothetical strategy: The estimand of interest is the difference in means of of one-year change from baseline in HbA1c in the hypothetical scenario where rescue medication was not available, but regardless of discontinuation of treatment. The considered estimand has the following attributes:

- Endpoint: Change in HbA1c
- Treatments: Experimental versus control
- Summary measure: Difference in means.
- Population: Patients with type-2 diabetes
- Intercurrent events: Use of rescue medication and treatment discontinuation
- Intercurrent event strategy:
  - Hypothetical strategy, assuming that patients had not initiated rescue medication.
  - Treatment policy strategy, ignoring the fact that patients discontinue treatment.

Under the hypothetical strategy the focus is on modelling hypothetical outcomes assuming there was no rescue medication. The considered causal inference methods are

- Inverse probability weighting (IPW) with outcome after rescue medication set to missing
- De-mediation as described in Loh et al. (2020); Lasch et al. (2023)
- G-computation

In the core simulation, data to be analysed with a hypothetical strategy are simulated included a mechanism to generate missing data. In additional simulations, data are generated without missing data for comparison.

3.1 Data generation

As general study design we assume a randomised trial in diabetes with 1:1 allocation ratio comparing an experimental treatment to placebo. The primary outcome variable is HbA1c after one year. We assume that HbA1c is also measured at baseline and at regular monthly visits. We denote the HbA1c value measured for the i-th patient at month 0 (baseline) to month $k = 12$ (final assessment) by $y_{ij}, j = 0, \ldots, k$.

The assigned treatment $A_i \in 0, 1$, with $A_i = 1$ denoting treatment and $A_i = 0$ denoting control is sampled from a Bernoulli distribution with equal probability for 0 and 1, corresponding to an allocation ratio of 1:1.

We consider age as prognostic baseline covariate as literature on the HbA1c trajectories suggests that younger patients diagnosed with diabetes tend to have higher starting values (Bhattacharjee et al., 2025) and steeper increase in HbA1c (Nicolaisen et al., 2024).

In the simulation, age is sampled from a normal distribution with a mean of 60 years and a standard deviation of 10 years.

Trajectories of HbA1c are sampled for an individual patient i by the following algorithm. Parameter values for the core scenario and some deviations are discussed in the text. All considered parameter values are shown in Table 3.1.

1. The expected HbA1c values across visits $j = 0, \ldots, k$ for the i-th patient is defined via a population mean at baseline $\mu_0 = 8\%$ and an age depending slope s_{age} and a time depending treatment effect $\beta_{trt,j}$, an individual treatment response modifier $\alpha_{response,i} \sim \text{uniform}(0, 1)$ and the assigned treatment $A_i \in \{0, 1\}$ as

$$\mu_{ij} = \mu_0 + j/k * s_{age} + \beta_{trt,j} * \alpha_{response,i} * A_i.$$

   Note that % here and in the remainder of the Section is the absolute unit of HbA1c.

   The age dependent slope is defined as

$$s_{age} = 2 * \exp(-b_{age} * (age - 30))$$

   In the core scenario, $b_{age} = \log(2)/10$, such that the slope is exponentially decreasing with age and patients at age 30 have a mean increase in HbA1c of 2% per year and this effect is halved every 10 years such that patients at age 40 have an increase of 1% per year.

   The treatment effect is defined as

$$\beta_{trt,j} = -\delta * (1 - \exp(-\lambda * j))$$

   with $\delta = 1$ in the core scenario and the term $(1 - \exp(-\lambda * j))$ increasing the treatment effect from 0 at baseline ($j = 0$) to a value close to δ over time. The rate with which the treatment effect is established is set to $\lambda = \log(2)/2$, such that after 2 months, 50% of the treatment effect is present. This rate of decrease matches a lifespan of erythrocytes of around 120 days and a corresponding turnover rate of 0.83 1% per day and also corresponds to published trajectories of HbA1c in treated diabetes patients, see Bongaerts et al. (2023).

   To model heterogeneity between patients, and also allow for patients with unfavourable values under treatment such that administration of rescue medication may also occur for some patients under an effective treatment, the treatment effect is modified by an individual modifier $\alpha_{response,i}$, which is sampled from a uniform(0,1) distribution.

   The true treatment effect at the final visit $k = 12$ thus is $-\delta_{true} = -\delta/2 * (1 - \exp(-\lambda * k))$, which is -0.492 in the core scenario.

2. Residuals $r_{ij}, j = 0, \ldots, k$ are sampled from a multivariate normal distribution with mean 0 vector and covariance matrix such that $cor(r_{ij}, r_{ij'}) = \rho$, for $j \neq j'$ and $var(r_{ij}) = \sigma^2$ for all $j = 0, \ldots, k$. In the core scenario, $\rho = 0.5$ and $\sigma^2 = 1$.

   The HbA1c values without possible rescue medication are thus defined as

$$y_{ij} = \mu_{ij} + r_{ij}$$

3. At visits $j = 1, \ldots, k-1$, rescue medication is initiated with a certain probability $p_{rescue,ij}$ that increases with current HbA1c value and slightly decreases with age according to

$$\log(p_{rescue,ij}/(1 - p_{rescue,ij})) = \beta_{rescue,0} + (y_{ij} - 10) * \beta_{rescue,y} + (age - 60) * \beta_{rescue,age}$$

   with core scenario values $\beta_{rescue,0} = \log(0.05/(1 - 0.05))$, $\beta_{rescue,y} = \log(3)$ and $\beta_{rescue,age} = -\log(1.01)$.

   By these values, the marginal probability for rescue is approximately 10%, the probability for rescue is increasing notably, but not extremely steep, around a plausible threshold of 10% HbA1c. At age 60, probabilities are approximately 0.01, 0.02, 0.05, 0.14, 0.32 at a single visit with HbA1c = 8, 9, 10, 11, 12, respectively. In a further scenario, $\beta_{rescue,y} = \log(150)$ are used, which results in respective probabilities 0.0000 0.0004, 0.0500, 0.8876, 0.9992 and may result in situations close to violation of the positivity assumption of IPW methods.

   Furthermore, the probability for rescue is slightly decreasing with age, which could be argued, e.g., by assuming that the disease is more aggressive in younger patients.

4. Denote the visit at which rescue was initiated as j^*. We assume that rescue medication is given continuously instead of the study treatment, once it has been initiated. Thus, the values $y_{ij}, j > j^*$ are modified to have no effect due to study treatment but have an effect due to rescue, and now take values

$$y_{ij} = \mu_0 + j/k * s_{age} + \beta_{rescue,j-j^*} * \alpha_{response\_rescue,i} + r_{ij}.$$

   The effect of rescue is modelled similar to the effect of the experimental treatment as

$$\beta_{rescue,j} = -\delta_{rescue} * (1 - \exp(-\lambda_{rescue} * j))$$

   with $\delta_{rescue} = 0.75$ in the core scenario and $\lambda_{rescue} = \log(2)/1$, such that rescue has only 75% of the effect of experimental treatment in the core scenario, but acts faster with only 1 month until 50% effect. Faster rescue effect may usually not occur with HbA1c as endpoint, but is considered for the simulation in order to truly see an effect of rescue in the control group, even if it is initiated at later visits and provide a more general scenario similar to other disease areas where rescue effects are fast. Similar to the experimental treatment effect, the effect of rescue is modified by a response term $\alpha_{response\_rescue,i} \sim \text{uniform}(0, 1)$ that is sampled independently from $\alpha_{response,i}$. (Note that without the response modifying terms, rescue would always be a disadvantage for treatment group patients, unless rescue was more efficacious than experimental treatment.)

   As an addition to the core scenario, in which rescue medication is given instead of treatment, we also investigated a setting where rescue medication is given in addition to the current treatment. This is both due to the fact that some of the methods actually aim towards this setting, and then it would be interesting to see their performance, and also due to the fact that it is often seen in practice, that patients will have some medication in addition to the experimental treatment. Of course this setting will only impact the mean outcome in the treatment group, since there are no differences for the control group. With this addition implemented, the values $y_{ij}, j > j^*$ are given by

$$y_{ij} = \mu_0 + j/k * s_{age} + \beta_{rescue,j-j^*} * \alpha_{response\_rescue,i} + \text{setup} * \beta_{trt,j} * \alpha_{response,i} * A_i + r_{ij}.$$

5. Missing data is introduced assuming that patients may withdraw from the study, causing a monotone missing pattern. The probability to withdraw at visit j, $p_{withdraw,ij}$, is modelled to depend on the current

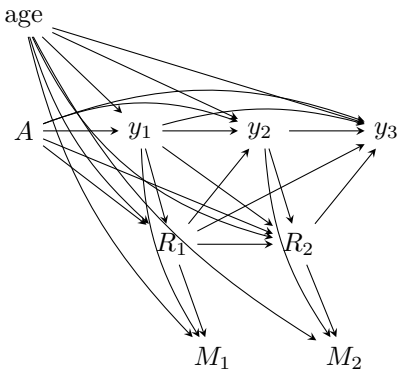


Figure 3.1: Directed acyclic graph (DAG) illustrating the assumed causal structure of the simulated longitudinal data, with the randomly assigned treatment (A), measured baseline covariate (age), indicator of rescue medication at visit j (R_j), indicator of missingness at visit j (M_j), and measurement of the outcome at visit j (y_j).

HbA1c value and the rescue medication status $R_{ij} \in 0, 1$, with $R_{ij} = 1$ indicating that rescue was initiated at or before visit j, via

$$\log\left(\frac{p_{withdraw,ij}}{1-p_{withdraw,ij}}\right) = \beta_{withdraw,0} + (y_{ij} - 10) * \beta_{withdraw,y} + (age - 60) * \beta_{withdraw,age} + R_{ij} * \beta_{withdraw,resc}$$

with core scenario values $\beta_{withdraw,0} = \log(0.02/(1-0.02))$, $\beta_{withdraw,y} = \log(3)$, $\beta_{withdraw,age} = \log(1.02)$ and $\beta_{withdraw,resc} = \log(1.5)$. Similar to the rescue model, a large value $\beta_{withdraw,y} = \log(150)$ is considered in an additional scenario to model close-to violation of positivity. The variable M_{ij} is drawn from a binomial distribution with probability $p_{withdraw,ij}$.

In the core scenario, data are set to missing after the visit of the withdrawal decision, such that data is missing at random. These core settings result in approximately 5% missing data at the final visit.

Figure 3.1 illustrates the causal structure for this scenario.

3.1.1 Sample size

For scenarios with different values for the treatment effect δ, the sample size is calculated to provide approximate power $1-\beta = 0.8$ at a two-sided significance level $\alpha = 0.05$ under the assumption of no rescue medication allowed and no missing data using the following equations. First, keeping in mind the analysis models include baseline HbA1c as covariate, the residual variance adjusted for baseline HbA1c is calculated as $\sigma^2_{adj} = \sigma^2 * (1-\rho^2)$. Then the sample size per group is calculated as $n = (z_{1-\alpha/2} + z_{1-\beta})^2 * \sigma^2_{adj}/\delta^2_{true} * 2$.

Of note, in the sample size calculation of an actual trial, the effect of covariates would typically be unknown and not be taken into account. However, in the simulation study, we aim for a range of power values sufficiently below 1 in order to clearly observe power differences between the investigated methods.

3.1.2 True treatment effect

As previously described, the true treatment effect can be derived analytically when the targeted estimand follows a hypothetical strategy. However, this derivation does not apply to estimands based on a treatment policy strategy. In this setting, the analysis methods aim to estimate a different, and typically smaller, treatment effect due to the inclusion of intercurrent events.

To quantify this effect, the true treatment effect under the treatment policy strategy is estimated empirically for each scenario using a simulated data set of size 4×10^7 generated without missing data.

Table 3.1: Parameter values for the data generating mechanism of the rescue medication scenarios.

Parameter	**Value in core scenario**	**Comment**	**Further values**	**Comment**
μ_0	8	Mean baseline HbA1c [%]		
σ	1	Std. dev. baseline HbA1c		
ρ	0.5	Correlation between repeated HbA1c measurements	0	No information across time points
δ	1	Maximal treatment effect	0, 0.5	No or weaker treatment effect.
λ	log(2)/2	Rate of increasing treatment effect		
δ_{rescue}	0.75	Maximal rescue effect		
λ_{rescue}	log(2)/1	Rate of increasing rescue effect		
$\beta_{rescue,0}$	log(0.05/(1-0.05))	appr. 10% rescue overall	log(0.2/(1-0.2))	appr. 30% rescue overall
$\beta_{rescue,y}$	log(3)		log(150)	Strong dependence of rescue on HbA1c threshold
$\beta_{rescue,age}$	-log(1.01)			
$\beta_{withdraw,0}$	log(0.02/(1-0.02))	appr. 10% missing	0, log(0.2/(1-0.2))	no missing or appr. 20% missing
$\beta_{withdraw,y}$	log(3)		0, log(150)	Strong dependence of withdrawal on HbA1c threshold
$\beta_{withdraw,age}$	log(1.02)		0	
$\beta_{withdraw,resc}$	log(1.5)		0	
setup	0	Rescue medication is taken instead of the treatment	1	Rescue medication is taken in addition to the treatment

3.2 Analysis methods for the treatment policy strategy

In this section the analysis methods in case of the treatment policy strategy are described, and they are summarised in Table 3.2 together with the methods for the hypothetical strategy.

3.2.1 Inverse probability weighting (IPW)

The probability for a patient to have non-missing outcome data at visit $j = 1, \ldots, k$ is estimated from a logistic regression model with explanatory variables treatment group, age at baseline, HbA1c at visit $j-1$ and rescue medication status at visit $j-1$. The models are fit using data from all patients still available at visit j. Denote the estimated probability for the i-th patient to have non-missing data at visit j as $\hat{\pi}_{ij}$. The inverse probability weight for a patient that remained in the study is calculated as $w_i = \frac{1}{\prod_{j=1}^{k} \hat{\pi}_{ij}}$. These inverse probability weights extracted from the final visit are applied to an ANCOVA model that includes only patients with non-missing primary outcome data at the final visit. The ANCOVA model includes treatment group, baseline HbA1c and age at baseline as covariates and HbA1c change from baseline as outcome. The weights are estimated using the `ipwtm` function from the `ipw` package in R, with the `type = "first"` option to compute weights based on the first occurrence of missingness. Robust HC2 standard errors are used for inference and implemented using `vcovHC(fit, type = "HC2")` via `coeftest` from `lmtest`. The IPW approach assumes a correctly specified propensity model and that the missing data satisfies the missing at random assumption.

3.2.2 Mixed model for repeated measures (MMRM)

An MMRM is fit to all the available longitudinal outcome data with predictors treatment group, baseline HbA1c, age at baseline, visit (as categorical variable) as well as visits by treatment and all visits by baseline covariate interactions. The model utilises an unstructured covariance matrix. The model is specified as:

$$Y_{ij} = \beta_0 + \beta_1 \text{trt}_i + \gamma_j \text{visit}_{ij} + \delta_j (\text{trt}_i \times \text{visit}_{ij}) + \alpha_1 y_{0,i} + \alpha_j (y_{0,i} \times \text{visit}_{ij}) + \eta_1 \text{age}_i + \eta_j (\text{age}_i \times \text{visit}_{ij}) + \varepsilon_{ij},$$

where:

- Y_{ij} denotes the outcome for subject i at visit j,
- trt_i is the treatment indicator,
- visit_{ij} is the categorical visit effect,
- $y_{0,i}$ is the baseline HbA1c value,
- age_i is the baseline age covariate,
- ε_{ij} denotes the within-subject residual error.

According to the Guideline on adjustment for baseline covariates in clinical trials (EMA/CHMP/295050/2013), when the baseline is included as a covariate in a standard linear model, the estimated treatment effects are identical for both 'change from baseline' (on an additive scale) and the 'raw outcome' analysis. Consequently if the appropriate adjustment is done, then the choice of endpoint becomes solely an issue of interpretability.

The within-subject covariance matrix is primarily modelled using an unstructured covariance matrix:

$\varepsilon_i \sim \mathcal{N}(0, \Sigma)$, where Σ is fully unstructured across visits.

If the unstructured covariance model fails to converge, a predefined fallback strategy is applied. The covariance structure is first simplified to compound symmetry and, if necessary, further simplified to a diagonal covariance structure. The covariance structure used in the final fitted model is recorded.

The model is fitted using the `mmrm` R package (Bell and Rabe, 2020) using restricted maximum likelihood (REML) estimation. Degrees of freedom and confidence intervals are based on Satterthwaite approximations.

To directly estimate the treatment effect at the final visit, the categorical visit factor is parametrised such that the final visit is the reference level. Under this parametrisation, the main treatment coefficient β_1 corresponds to the treatment comparison at the final scheduled visit.

Under the treatment policy strategy, all observed post-baseline outcome data is included in the analysis regardless of rescue medication initiation or treatment discontinuation.

Similar to the assumptions of the IPW, the MMRM also assumes correctly specified models and that the missing data satisfies the missing at random assumption.

Of note, in comparison to the multiple imputation method discussed below, the MMRM does not include rescue medication indicators at each visit. This is in line with the Guideline on adjustment for baseline covariates in clinical trials (EMA/CHMP/295050/2013) where it is stated that a covariate measured after randomisation such as use of rescue medication should not be included in the primary analysis of a confirmatory trial.

3.2.3 Multiple imputation conditional on rescue medication

Multiple imputation is applied in which the imputation is performed conditional on age, the rescue medication status, and past outcome values. The R package mice is used. The imputation includes the variables HbA1c and rescue medication status at each visit to be imputed if missing; observed HbA1c values and rescue medication status and age at baseline are used as predictors. The restriction that rescue medication is taken continuously once started is included using the passive imputation specification in `mice`. The imputation is performed for each treatment group separately. The imputation model includes:

- baseline HbA1c ($y_{0,i}$),
- baseline age,
- observed longitudinal HbA1c measurements,
- rescue medication indicators at each visit.

HbA1c values are imputed using predictive mean matching (PMM), while rescue medication indicators are imputed using logistic regression models.

For each post-baseline visit j, the imputation model for HbA1c includes adjacent visit history, such that the immediately preceding HbA1c measurement ($y_{i,j-1}$) is included as a predictor whenever available.

Rescue medication indicators are included as predictors in the HbA1c imputation models in order to preserve the association between rescue medication use and subsequent HbA1c trajectories under the treatment policy estimand. A number of 10 multiple imputations is used. For each imputed dataset, an ANCOVA model for change from baseline in HbA1c at the final visit is fitted:

$$\Delta Y_i = \beta_0 + \beta_1 \text{trt}_i + \beta_2 \text{age}_i + \beta_3 y_{0,i} + \varepsilon_i,$$

where:

- ΔY_i denotes the change from baseline in HbA1c for subject i,
- trt_i is the treatment indicator,
- age_i is age at baseline,
- $y_{0,i}$ is baseline HbA1c,
- ε_i is the residual error term.

Randomness arising from the multiple imputation procedure is controlled using seeds to ensure reproducibility of the analysis.

3.3 Analysis methods for the hypothetical strategy

In this section the analysis methods in case of the hypothetical strategy is described, and they are summarised in Table 3.2 together with the methods for the treatment policy strategy.

3.3.1 Inverse probability weighting (IPW)

This approach is similar as described for the treatment policy, with the difference that data after initiation of rescue medication is set to missing and rescue status is not used as predictor in the missingness model, similar as described in Olarte Parra et al. (2025). That is, the probability for a patient to have non-missing outcome data at visit $j = 1, \ldots, k$ is estimated from a logistic regression model with explanatory variables treatment group, age at baseline and HbA1c at visit $j - 1$. The models are fit using data from all patients still available at visit j. Denote the estimated probability for the i-th patient to have non-missing data at visit j as $\hat{\pi}_{ij}$. The inverse probability weight for a patient that remained in the study is calculated as $w_i = \frac{1}{\prod_{j=1}^{k} \hat{\pi}_{ij}}$. The weights are estimated using the `ipwtm` function in the `ipw` package in R. The weights are estimated using the `ipwtm` function from the `ipw` package in R, with the `type = "first"` option to compute weights based on the first occurrence of missingness. Robust HC2 standard errors are used for inference and implemented using `vcovHC(fit, type = "HC2")` via `coeftest` from `lmtest`.

3.3.2 De-mediation

The considered de-mediation approach was proposed by (Loh et al., 2020) and implemented in (Lasch et al., 2023; Lasch and Guizzaro, 2022; Olarte Parra et al., 2025) for different settings. A structural nested mean model is defined to model the effect ψ of a mediator (i.e. rescue medication) on the expected value of the outcome, conditional on the covariate history and treatment group. Outcomes after rescue medication are then adjusted by subtracting $\hat{\psi}$ and the adjusted outcome values are compared between groups.

A basic algorithm, assuming only one possible time point for rescue and that both covariates X and treatment group A are observed before this time point, is as follows:

- Estimate probability for rescue, $P(R = 1 \mid X, A)$ depending on treatment and covariates. To account for perfect separation, a firth corrected logistic regression is used.
- Fit a regression model for the outcome on covariates, treatment, rescue status and $P(R = 1 \mid X, A)$.
- Calculate $y^* = y - \hat{\psi}R$, where $\hat{\psi}$ is the regression coefficient for R in above model.
- Calculate the mean difference for y^* between the two treatment groups.

For an inherently longitudinal setting as envisaged in the simulation, an extension of this approach based on sequential application of the algorithm to all visits was described in Section 4.2.3 of (Olarte Parra et al., 2025) and is used in the simulation study.

The method assumes that no unobserved confounder between treatment and outcome and no unobserved confounders between the mediator and the outcome exist.

Missing data after treatment discontinuation is handled by applying multiple imputation prior to the actual analysis. The imputation is performed equally as presented in Section 3.2.3. It is performed conditional on age, rescue medication status, and past outcome values, and is performed using all available data. The R package mice is used. The imputation includes the variables HbA1c and rescue medication status at each visit to be imputed if missing; observed HbA1c values and rescue medication status and age at baseline are used as predictors. The restriction that rescue medication is taken continuously once started is included using the passive imputation specification in `mice`. The imputation is performed for each treatment group separately. HbA1c values are imputed using predictive mean matching (PMM), while rescue medication indicators are imputed using logistic regression models. The subsequent analysis is performed for each imputed data set and the results are pooled according to Rubin's rule.

3.3.3 G-computation

The idea of g-computation is to model the distribution of dependent variables, that being HbA1c trajectories and administration of rescue over time, by fitting respective models for data at each visit conditional on data at the previous visits. A custom function is used that fits a model for HbA1c at each visit (Y_j), using data from all patients still available at that visit. The models utilised in the G-computation are models of the shape:

$Y_j = A + R_j + y_{j-1} + age$ where the HbA1c value at visit j is modelled as a function of treatment group, rescue medication status at visit j, lagged HbA1c (i.e., HbA1c at visit $j-1$), and age.

A restriction is applied on the original data during data pre-processing such that once rescue medication is initiated (i.e., $R_{ij-1} = 1$), it remains active ($R_{ij} = 1$) for all subsequent visits, reflecting the clinical reality of continuous rescue use.

The fitted models for HbA1c at each visit j are used to sample for each treatment group a number of trajectories equal to the number of patients originally included, starting from baseline covariate values and treatment group, albeit under the assumption that rescue medication is not possible. The sampled data set does not contain missing values. The mean difference of the outcome between groups is estimated from the sampled data.

Bootstrap is applied to the whole procedure (fitting the models, g-computation under the assumption of no rescue medication allowed and calculation of the mean difference). The number of bootstrap samples is set to 200. Bootstrap 95% confidence intervals for the mean difference are calculated using the percentile method. The null hypothesis of no treatment effect is considered rejected if the confidence interval does not contain the value 0. A one-sided bootstrap-based p-value is computed to test the alternative hypothesis that the treatment effect is less than zero (i.e., the experimental treatment reduces HbA1c). The one-sided p-value is calculated as the proportion of the null distribution that is greater than or equal to zero following Rousselet et al. (2021). The null distribution of the treatment effect is constructed by centering the bootstrap estimates under the null hypothesis (i.e., subtracting the mean of the bootstrap samples from each bootstrap estimate).

Two approaches are supported through the setup argument. For setup = 1, treatment is kept as assigned in the observed data. For setup = 0, treatment is recoded to control from the time rescue starts. This recoding is done only once on the observed post-rescue data. During counterfactual simulation, rescue status is deterministically set to 0 for all individuals and visits so the treatments remain as randomised.

3.3.4 Mixed model for repeated measures (MMRM) as comparator method

For the MMRM, data after initiation of rescue medication is set to missing. The MMRM is subsequently applied in the same way as described for the treatment policy approach.

3.3.5 Multiple imputation as comparator method

Multiple imputation is used as further comparator method. As with MMRM, data after initiation of rescue medication is set to missing and multiple imputation is applied to obtain outcome values under the assumption that rescue was not available. For this reason rescue medication indicators at each visit are not included anymore in the imputation model. The HbA1c measurement obtained at the rescue visit itself (i.e. $j = r_i$) is retained in the analysis. HbA1c at each visit is imputed if missing, based on observed HbA1c values and age at baseline. The imputation is performed for each treatment group separately. HbA1c values are imputed using predictive mean matching (PMM). The imputation model includes:

- baseline HbA1c ($y_{0,i}$),
- baseline age,
- immediately preceding HbA1c visit history.

For each post-baseline visit j, the imputation model for HbA1c includes the immediately preceding HbA1c measurement ($Y_{i,j-1}$) whenever available. Under the hypothetical strategy, rescue medication indicators are not included as predictors in the HbA1c imputation models, in order to avoid conditioning on post-rescue information. The multiple imputation and ANCOVA are subsequently applied in the same way as described for the treatment policy approach.

Table 3.2: Summarising the methods used, along with a short description of each, the summary measure they report and how it is planned to implement them in the software.

Method	Description	Summary measure	Implementation
Treatment policy strategy			
Inverse probability weighting (IPW)	ANCOVA model with weights from IPW	Difference in means	lm, ipw::ipwtm
Mixed model for repeated measures (MMRM)	Visit by baseline interactions and unstructured covariance matrix	Difference in means	mmrm::mmrm
Multiple imputation	Comparator method	Difference in means	mice::mice
Hypothetical strategy			
Inverse probability weighting (IPW)	ANCOVA model with weights from IPW	Difference in means	lm, ipw::ipwtm
De-mediation	Subtracting the effect of the mediator through modelling	Difference in means	custom implementation using logistf, glm, lm, mice
G-computation	Modelling all nodes in the DAG	Difference in means	custom implementation using lm
Mixed model for repeated measures (MMRM)	Visit by baseline interactions and unstructured covariance matrix	Difference in means	mmrm::mmrm
Multiple imputation	Comparator method	Difference in means	mice::mice

3.4 Deviations from the study protocol

3.4.1 Data generation

A scenario in which the effect of rescue medication was added on top of the treatment as opposed to be given instead of treatment, was added.

A subsection on calculation of the true treatment effect is added, to accommodate the fact that using two different strategies to account for intercurrent event aims towards two different true values.

Through simulations, it was encountered that the missingness did not behave as intended, hence for the scenario with increased missingness the parameter value of $\beta_{withdraw,0}$ is changed from $\log(0.04/(1-0.04))$ to $\log(0.2/(1-0.2))$.

3.4.2 Analysis methods for treatment policy strategy

IPW: Added the use of the type = "first" option in the `ipwtm` function, which computes weights based on the first time a patient becomes missing (or is censored). Added details on the implementation of robust standard errors. This implementation is consistent with the protocol's intent even so it was not explicitly pre-specified.

MMRM: The MMRM can "fit" but with non-invertible Hessian, returns coefficients with NA / unstable SE give invalid df and return numbers that look fine but are actually numerically unreliable. This is why defensive checks

for coefficient existence / failed fits together with fallback covariance matrices were introduced. The primary covariance structure is unstructured (US). If model fitting fails due to convergence issues, simpler structures are considered sequentially like Compound symmetry (CS) and Diagonal (independence). This fallback mechanism ensures numerical stability in simulation settings while preserving the primary modelling intent whenever possible.

MI: The mice package lead a lot of "logged events" warning messages. To make the procedure more robust the following was implemented: HbA1c outcomes were imputed using predictive mean matching (PMM) implemented via the mice package. Imputation was performed separately within treatment groups. A monotone sequential predictor structure was used in which each HbA1c visit was imputed using baseline HbA1c, age, and the immediately preceding HbA1c measurement. Under the treatment-policy estimand, rescue indicators were additionally included as predictors. Rescue indicators themselves were imputed using logistic regression when applicable. Imputations used 5 PMM donor draws, ridge stabilization, and 10 MICE iterations. Randomness arising from the imputation procedure is controlled using `withr::with_seed()`, ensuring reproducibility of imputations without modifying the global random number generator state. This avoids unintended interference with simulation-based data generation across replicates.

3.4.3 Analysis methods for the hypothetical strategy

IPW: Added the use of the type = "first" option in the `ipwtm` function, which computes weights based on the first time a patient becomes missing (or is censored). Added details on the implementation of robust standard errors. This implementation is consistent with the protocol's intent even so it was not explicitly pre-specified.

Demediation: To account for perfect separation the firth corrected logistic regression is implemented, with the use of the `logistf` package in R.

G-computation: During initial testing, the runtime of the `gformula_continuous_eof` function from the `gfoRmula` package was found to be prohibitively high, presumably due to its internal handling of time-varying covariates, complex model specification, and repeated refitting across bootstrap samples. To ensure computational feasibility, a custom implementation was developed. Key differences between the custom function and `gformula_continuous_eof` include:

- Simplified model structure: The custom function fits visit-specific models only for HbA1c and the final outcome, omitting a separate rescue model in the simulation step. Since the intervention is a fixed "no-rescue" scenario, the simulation does not require a stochastic rescue model. Instead, rescue is deterministically set to 0 in all post-baseline visits. This omission of the rescue model is a simplification specific to this no-rescue intervention and would not be appropriate for natural-course simulations or alternative rescue interventions.
- Manual simulation loop: Instead of relying on the package's internal simulation engine, the counterfactual trajectories are simulated iteratively using a loop over time points, allowing full control over the intervention (e.g., setting $R = 0$) and avoiding redundant model evaluations.
- Efficient data handling: The data are reshaped once at the start, and model fitting and simulation are performed directly on the long-format data without intermediate object conversions.
- Bootstrap integration: The entire procedure (resampling, refitting, simulation, estimation) is wrapped in a single bootstrap loop, minimising overhead and enabling faster execution.

The model for the outcome was adjusted to include the HbA1c value at baseline as a predictor because this aligns more closely with the data generating mechanism.

A restriction is applied on the original data during data pre-processing such that once rescue medication is initiated (i.e., $R_{ij-1} = 1$), it remains active ($R_{ij} = 1$) for all subsequent visits, reflecting the clinical reality of continuous rescue use. No explicit restriction is implemented for the counterfactual estimation under the no rescue intervention because such a restriction is trivially satisfied as rescue is fixed to 0 at all visits.

The number of bootstrap samples was set to 200 because 500 samples were unfeasible in terms of computation time. Contrary to the protocol, a bootstrap-based one-sided p-value is computed.

A setup parameter was added to allow to accommodate the added scenario in which the effect of rescue medication is added on top of the treatment as opposed to be given instead of treatment. The parameter only influences the data pre-processing step but does not change the analysis function.

MMRM: Same as for the treatment policy strategy.

MI: Same as for the treatment policy strategy except that under the hypothetical estimand, rescue indicators were excluded to avoid conditioning on post-rescue information.

3.5 Results

In Table 3.3 the naming of the specific data generation scenarios are presented together with the specification that deviated from the core scenario. For the parameter values for the core scenario, see the second column of Table 3.1. Table B.1 provides an overview of the simulation settings, including sample size, proportion of patients receiving rescue medication, the degree of missingness, together with the true treatment effect under treatment policy strategy. Across scenarios, sample sizes were approximately balanced between treatment groups. The proportion of rescue medication ranged from approximately 6% to 39%, while missingness varied from 0% (NM scenario) to approximately 17% (HM scenario).

Table 3.3: Scenario-specific deviations from the core scenario

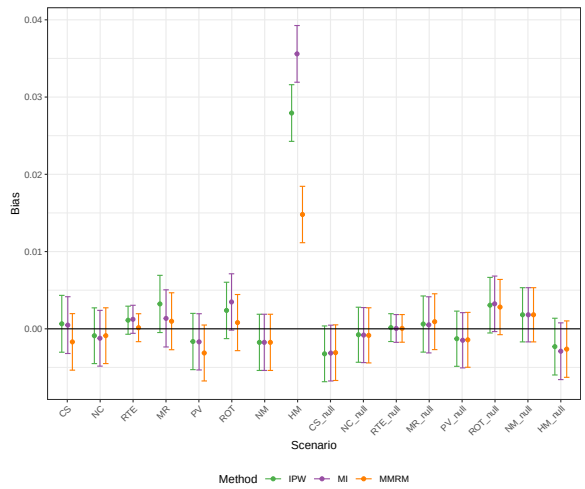
Scenario	Change from Core Scenario
CS	No change (core scenario)
NC	$\rho = 0$ (no correlation)
RTE	$\delta = 0.5$ (reduced treatment effect)
MR	Increased rescue probability (resc_0)
PV	Strong HbA1c effect on rescue ($\text{resc}_y = \log(150)$)
ROT	setup = 1 (rescue added on top of treatment)
NM	No missing data
HM	Higher missingness probability

Figures 3.2–3.7 summarise the performance of the investigated methods across the considered simulation scenarios for both the treatment policy and hypothetical estimands. The results are evaluated in terms of bias, RMSE, confidence interval (CI) coverage and width, type I error, and power. The shown error bars are normal approximated confidence intervals for the point estimates. In Tables B.2–B.6 the values for the assessed parameters are shown. Hypothesis tests are performed at the one-sided 2.5% significance level and confidence intervals are calculated as two-sided 95% confidence intervals. The tested one-sided null hypothesis is that HbA1c values do not decrease due to treatment.

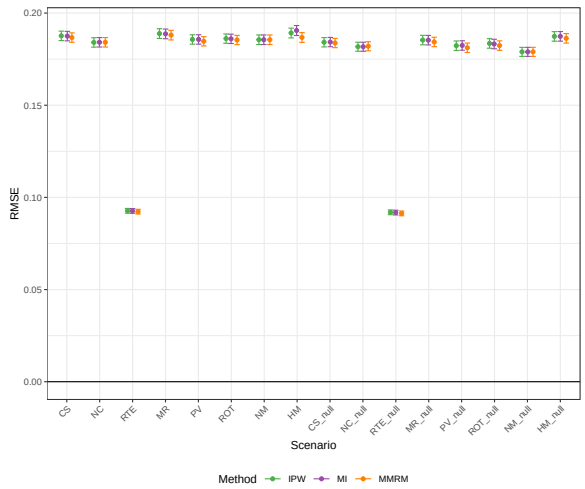
3.5.1 Treatment policy estimand

In the following figures, the results from the analyses targeting the treatment policy estimand are presented. Figures 3.2a, 3.2b, 3.3a, 3.3b, 3.4a and 3.4b summarise the performance (bias, RMSE, CI width, CI coverage, type I error rate, power) of the considered methods (IPW, multiple imputation (MI), and MMRM) under the treatment policy estimand for all 16 scenarios. For explanation of the naming see Table 3.3. As the true treatment effect under the treatment policy strategy is not analytically extractable, it is calculated as part of the simulation in large data sets separately for each scenario. The simulated true treatment policy effect can be found in the last column of Table B.1.

The bias plot shows that all considered methods yield estimates close to zero bias across most scenarios, indicating that the estimators are generally unbiased under the treatment policy estimand. Slight deviations from zero can be observed in more extreme scenarios, such as those involving strong rescue effects, missingness, or positivity violations. These deviations suggest increased sensitivity of the methods to violations of underlying assumptions. In these settings, multiple imputation (MI) tends to exhibit the largest deviations, followed by inverse probability weighting (IPW), whereas the mixed model for repeated measures (MMRM) generally shows the smallest bias.



(a) The bias of the methods.



(b) The root mean squared error of the methods.

Figure 3.2: The bias and RMSE of the methods used for the treatment policy strategy, along with the confidence interval as error bars.

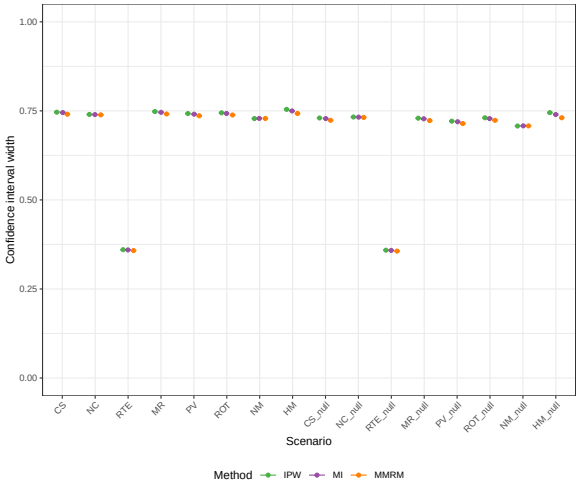
In scenarios without missingness (NM), all methods perform nearly identically and also with the missingness mechanism of the core scenario there is almost no difference. Considering the high missingness scenario (HM) MI has the largest bias, then IPW and the least bias is achieved with MMRM. The bias is markedly positive for the high missingness scenario which would lead to an underestimation of the treatment effect and conservative effect estimates. Overall, the results suggest that MMRM is slightly more robust to increased missingness, while all methods behave similarly when assumptions such as missing at random are well satisfied. For the exact values of the bias, see Table B.2.

Figure 3.2b presents the root mean squared error (RMSE) of the treatment effect estimates. Consistent with the bias results, all methods show similar RMSE across most scenarios, indicating comparable overall accuracy. Differences in RMSE are primarily driven by the data-generating scenario rather than by the choice of method. In particular, scenarios with reduced treatment effects (RTE) exhibit smaller RMSE due to larger effective sample sizes, whereas more complex scenarios (e.g. high missingness (HM) or increased rescue probability (MR)) are associated with increased RMSE. No method consistently dominates the others with respect to RMSE. Again, the largest differences are seen in the high missingness scenario where the MMRM outperforms IPW and MI. For the exact values of the RMSE, see Table B.3.

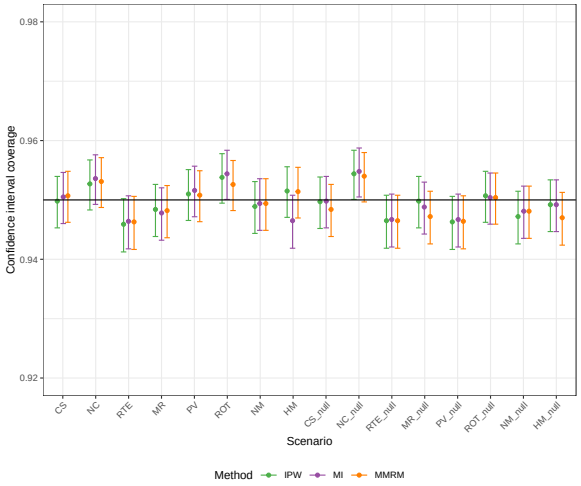
In Figure 3.3 the width of the confidence interval is presented together with the confidence interval coverage probability. Confidence interval widths are largely similar across methods, indicating comparable efficiency. Variations in precision are mainly driven by the scenario rather than the analytical approach. For example, scenarios with reduced treatment effect (RTE) exhibit narrower intervals due to larger required sample sizes.

All the methods generally retain the coverage probability of 95%, as seen in Figure 3.3b. For the exact coverage probabilities see Table B.4. One large difference between methods is seen for the high missingness scenario (HM) where IPW and MMRM retain the 95% coverage probability whereas MI has a slightly lower than 95% coverage probability.

Type I error is well controlled for all methods, with rejection rates under the null scenarios remaining close to the nominal significance level, see Figure 3.4a. A slight increase is observed under the PV scenario where there is a strong HbA1c effect on rescue. Power in Figure 3.4b is also comparable across scenarios, with expected reductions with increased rescue probability (MR) and increased missingness (HM). There seems to be a slight power gain for the PV scenario (Strong HbA1c effect on initiation of rescue medication) and the ROT scenario (rescue medication is added on top of treatment). The MMRM has consistently slightly higher power than IPW and MI which perform approximately similar for all considered scenarios. For the exact values of the type I

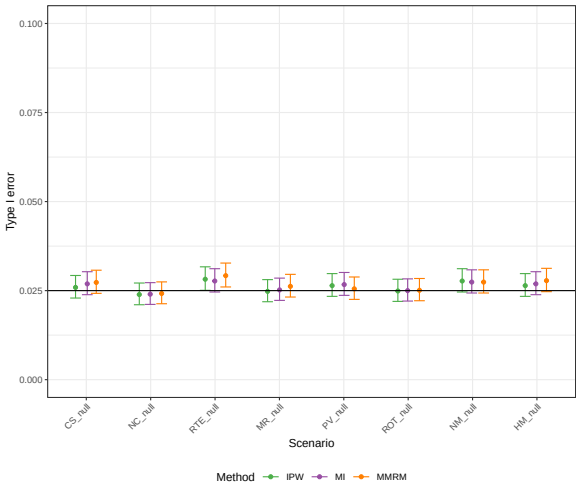


(a) The confidence interval width of the methods.

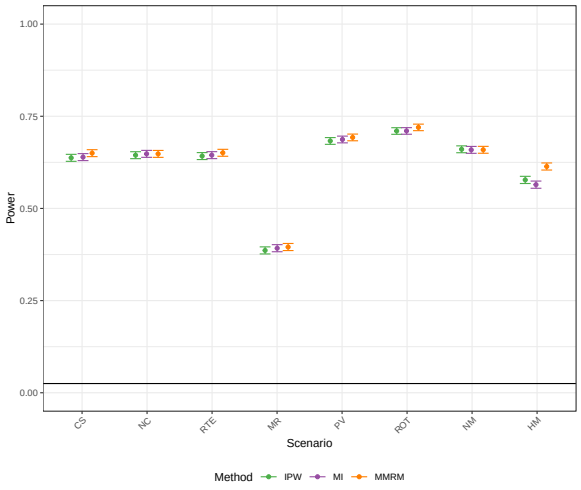


(b) The confidence interval coverage of the methods.

Figure 3.3: Confidence interval width and coverage of the methods used for the treatment policy strategy, along with the confidence interval as error bars.



(a) The type I error rate of the methods.



(b) The power of the methods.

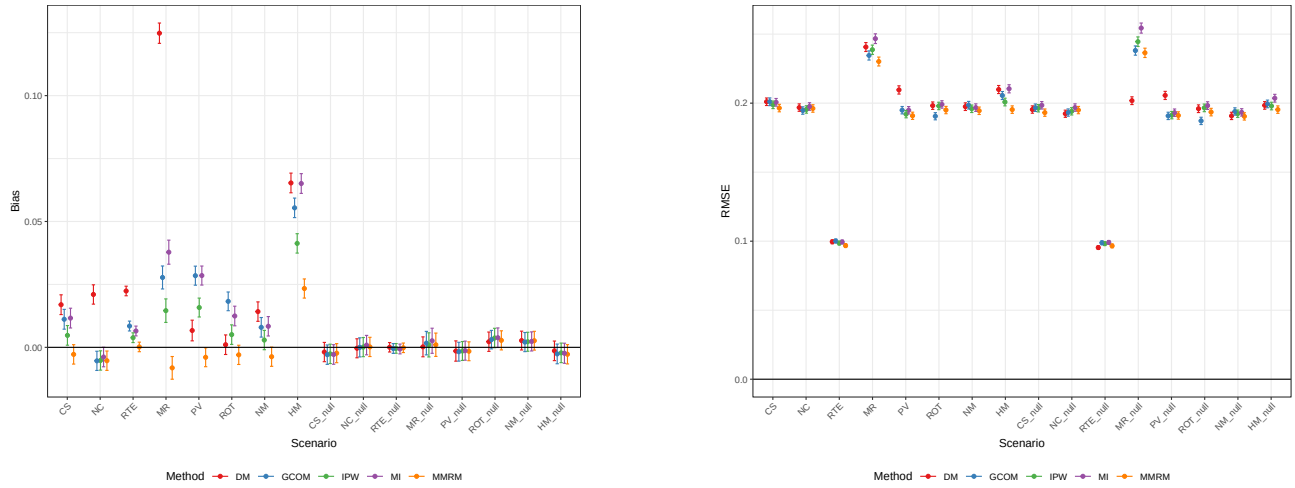
Figure 3.4: The type I error rate and power of the methods used for the treatment policy strategy, along with the confidence interval as error bars.

error and power, see Table B.5 and Table B.6.

Overall, under the treatment policy estimand all investigated methods perform robustly and similarly across most scenarios. In case of high missingness, the MMRM did perform slightly better than the other two methods concerning both bias and power. The results suggest that, provided assumptions such as missing at random (MAR) are approximately satisfied, method choice is less critical, and differences in performance are relatively minor compared to the impact of the underlying data-generating mechanism.

3.5.2 Hypothetical estimand

Figures 3.5, 3.6, and 3.7 present the corresponding results (bias, RMSE, CI coverage and width, type I error rate, and power) for the hypothetical estimand, including the methods de-mediation (DM) and g-computation (GCOM), alongside IPW, MI, and MMRM, as only applicable to a hypothetical estimand. In contrast to the treatment policy setting, performance varies more substantially across methods. The hypothetical estimand requires stronger assumptions (e.g., correct modelling of the post-rescue counterfactual trajectory), and this is reflected in increased variability of bias, coverage, and power across scenarios.



(a) The bias of the methods.

(b) The RMSE of the methods.

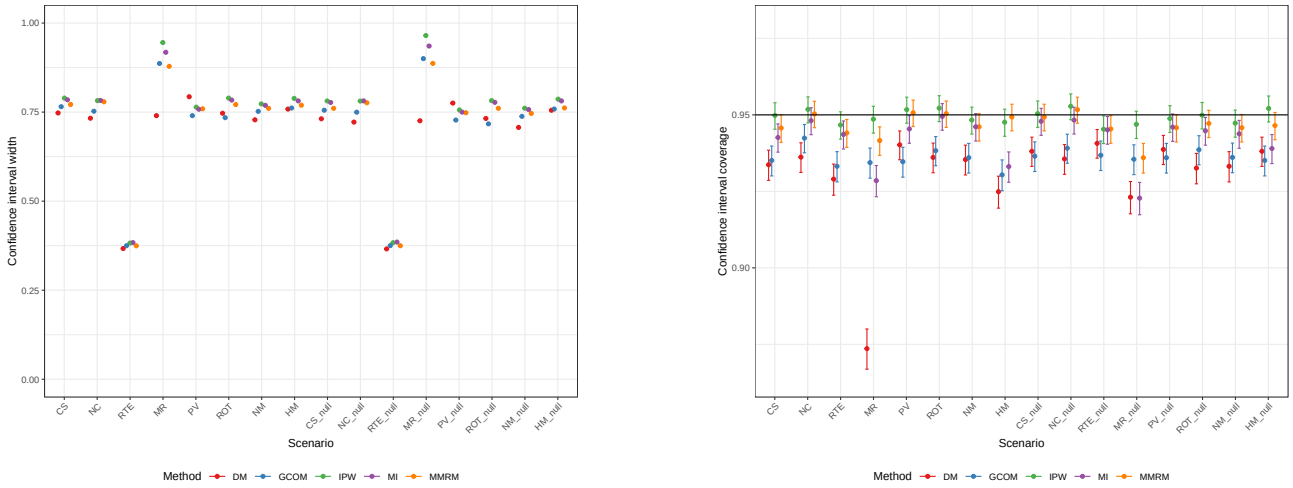
Figure 3.5: The bias and RMSE of the methods used for the hypothetical strategy, along with the confidence interval as error bars.

Under the hypothetical estimand (Figure 3.5a), variability in bias is more pronounced. Among the considered methods, MMRM generally exhibits the smallest bias across most scenarios, however, the bias is often in the negative direction, especially with increased rescue probability (MR), leading to anti-conservative behaviour. For high missingness (HM) the bias of DM is high and similar to the bias of MI. DM often has the highest bias except when there is a strong HbA1c effect on rescue (PV) and rescue added on top of treatment (ROT), for the latter DM is even the best performing method. IPW shows a moderate positive bias in most scenarios. GCOM often has a similar or slightly less positive bias compared to MI except for rescue added on top of treatment (ROT). In settings with no correlation (NC), where model assumptions are less compatible with the data-generating mechanism, all methods except DM have a negative bias. For the exact values of the bias, see Table B.2.

Interestingly, DM shows a markedly large positive bias for increased rescue probability (MR). This performance can be explained by conflicts between the data generating mechanism and the analysis method. Whenever patients is in need of rescue medication they will be taken off the experimental treatment and receive rescue medication instead. In this scenario we turned up the amount of rescue medication used, which implies that patients initiate rescue earlier than in the core scenario, and as a result, patients have less time on the experimental treatment. This is an issue for the de-mediation method, since it consider the final outcome and then for every visit where rescue medication is used, the estimated visit specific effect is subtracted. This leaves us with only the effect of the experimental treatment, as intended, but this will be lower than expected, since the treatment

was not administered for the full trial period. With these considerations, this is maybe not deemed the most reasonable mechanism to implement for rescue medication as an intercurrent event. These considerations also tell, that applying de-mediation in a setting where the treatment is not administered for the full trial period is not appropriate.

Figure 3.5b presents the root mean squared error (RMSE) of the estimators. Consistent with the bias results, RMSE varies considerably across methods and scenarios for the hypothetical estimand. Like for the treatment policy estimand the expected lower RMSE is observed for the reduced treatment effect scenario with a higher sample size. However, the increased rescue probability scenarios (MR and MR_null) show a substantial increase in RMSE. Under the null hypothesis, DM obtains the lowest RMSE with increased rescue probability (MR_null). Concerning differences between methods, RMSE seems highest for MI except for the strong HbA1c effect on rescue (PV and PV_null) where DM performs worst. For high missingness, MI and DM have similar but slightly higher RMSE than GCOM, IPW and MMRM. In most scenarios MMRM is among the methods with the lowest RMSE. MMRM, GCOM and IPW tend to achieve lower RMSE in many scenarios, reflecting a balance between bias and variance. DM shows competitive performance only in some settings, particularly where the mediator modelling is well aligned with the data-generating mechanism. For the exact values of the RMSE, see Table B.3.



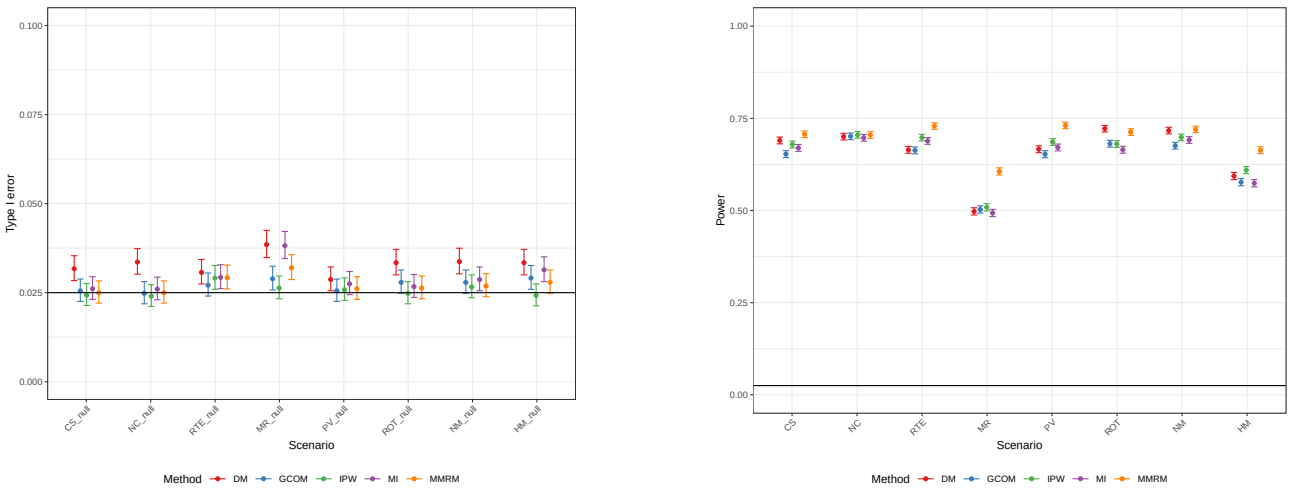
(a) The confidence interval width of the methods.

(b) The confidence interval coverage of the methods.

Figure 3.6: Confidence interval width and coverage of the methods used for the hypothetical strategy, along with the confidence interval as error bars.

Figure 3.6 displays the confidence interval widths and coverage probabilities. Compared to the treatment policy setting, greater variability in both interval width and coverage is observed across methods. In particular, IPW tends to produce wider intervals compared to the other methods consistent across scenarios. The CI width is markedly lower for the reduced treatment effect scenario (RTE) across all methods due to the higher sample size. DM has a markedly lower CI width than the other methods for the increased rescue probability scenario (MR) and has consistently the lowest CI width across scenarios. MI has wider CIs than MMRM and GCOM. GCOM has the lowest CI width after DM except for the scenario with increased rescue probability (MR) where it is outperformed by MMRM.

Confidence interval coverage in Figure 3.6b remains close to nominal 95% levels, especially for IPW and MMRM, which perform markedly better than DM, GCOM, and MI with almost always lower than the nominal coverage probability. DM often has the lowest coverage probability which is particularly low for the scenario with increased rescue probability (MR and MR_null). But also for high missingness (HM) and a reduced treatment effect (RTE) the coverage probability is particularly low for DM and GCOM. This is consistent with the low CI width observed for DM and GCOM where the low width comes at the cost of a low coverage probability. DM has a substantial drop in coverage probability for the increased rescue probability scenario (MR) consistent with the high observed bias. For the exact values of the coverage, see Table B.4.



(a) The type I error rate of the methods.

(b) The power of the methods.

Figure 3.7: The type I error rate and power of the methods used for the hypothetical strategy, along with the confidence interval as error bars.

Figure 3.7 presents the empirical type I error rates and power. Figure 3.7a shows that there are mostly only slight increases in type I error rate. The inflation in type I error is highest for the increased rescue probability scenario (MR_null) but below 0.04 (see Table B.5) and generally lower for MMRM and IPW than for MI, DM or GCOM. DM and GCOM are slightly inflated across scenarios. IPW performs best in controlling type I error rate across scenarios while under increased rescue medication (MR) there also is an inflation for MMRM. For the exact values of the type I error rate, see Table B.5.

Power is depicted in Figure 3.7b. The methods perform very similar concerning power with MMRM having a slight but consistent advantage over all scenarios. Only in the rescue on top (ROT) scenario MMRM is slightly outperformed by DM. The higher power of MMRM is most pronounced in the scenarios with increased rescue probability (MR) and high missingness (HM) where the other methods display considerably decreased power. For the exact values of the power, see Table B.6.

Considering only the methods applied for both the treatment policy estimand and the hypothetical estimand, modelling assumptions, such as missing at random (MAR), are more prone to violation under the hypothetical estimand. This is due to the fact that the missingness is higher since outcomes after initiation are set to missing as well. This is the general pattern, since data for the treatment policy estimand always contains less than or equal amount of missingness compared to the data for the hypothetical scenario. However, in cases with high missingness, model assumptions are also of concern when applying the treatment policy estimand. This discussion also highlights the fact that the missingness due to the intercurrent event of receiving rescue medication and the actual missingness due to withdrawal are pooled and treated as one single censoring process.

The DM can provide improved bias or power properties for rescue on top scenarios (ROT) and strong HbA1c effect on rescue scenarios (PV) when correctly specified, but may be more sensitive to misspecification. When the proportion of intercurrent events is moderate and the underlying assumptions are approximately satisfied, MMRM and IPW may provide adequate estimates, with results broadly comparable to those obtained using more complex causal inference approaches.

The findings from this part of our study can be compared to the results presented in the paper *Estimating hypothetical estimands with causal inference and missing data estimators in a diabetes trial case study* by Olarte Parra et al. (2025). A key finding of Olarte Parra et al. (2025) is that the different estimators (MMRM, MI, IPTW, G-formula, and G-estimation) yielded broadly similar treatment effect estimates and standard errors across methods. This is consistent with the results observed in the present simulation study, where all methods show very similar performance in terms of bias, coverage, and type I error, indicating robustness when assumptions

Table 3.4: This table shows the messages received through the simulation run with 10,000 iterations, presented by their frequency.

Scenario	logistf	non convergence	separation	logged events
CS	2	3	1018	20000
NC	0	0	635	19995
RTE	0	0	0	17000
MR	0	4	540	18804
PV	3	4	870	20000
ROT	0	1	1075	19999
NM	0	0	7	20000
HM	1	0	68	20000
CS_null	0	1	668	19998
NC_null	0	0	272	19982
RTE_null	0	0	0	15973
MR_null	0	2	393	18102
PV_null	8	3	580	20000
ROT_null	0	1	655	20000
NM_null	0	0	1	19998
HM_null	0	0	3	20000

are approximately satisfied. In the edge cases of the assumptions, for instance in the MR, PV and HM scenario, the performance of the methods are less robust, producing unbiased estimates and have increased RMSE. In Olarte Parra et al. (2025), IPTW produces estimates similar to MI and MMRM but with slightly larger standard errors due to the variability introduced by weighting. This pattern is also observed in our simulation results, where IPTW tends to show increased variability. However, in our simulation study, greater differences between methods are observed under the hypothetical estimand. While Olarte Parra et al. (2025) also report some variability across estimators, particularly for certain implementations of the G-formula, the overall agreement in their case study is stronger. This difference may be explained by the relatively low proportion of intercurrent events and missing data in their trial dataset, limiting the impact of modelling assumptions.

3.5.3 Warning messages encountered during the simulation run.

Even though it was aimed to eliminate warnings by introducing fallback options for example, we still received some messages from the simulation run. Table 3.4 reports the frequency of messages encountered during the simulation runs, most notably those related to model fitting procedures.

The most pronounced message received during the simulation run is the message categorised as logged events in Table 3.4, and this is a message from the `mice` package. Whenever an imputation model is used, and two of the predictors are collinear, it collects this in the logged events. It is not surprising, that this occurs a lot of times, given that for every data set the multiple imputation by chained equations is performed 3 times.

The second most pronounced message is categorised as separation in Table 3.4, and it reads: `glm.fit: fitted probabilities numerically 0 or 1 occurred`. This occurs in the IPW implementation, but an investigation of this warning showed that this is only the case for non-observed individuals. That is, the weights that are carried to the analysis does not explode due to this message.

The other warnings occur less than 10 times in 10000 iterations, and are hence not of a concern. Furthermore, other warnings, such as non-convergence of models or issues arising in penalised regression procedures (i.e. Firth logistic regression fits), occurred only rarely and were not systematic across scenarios. Overall, the low frequency of these warnings indicates that the implemented estimation procedures are numerically stable in the vast majority of cases, although care is required in scenarios approaching positivity violations.

Table 3.5: Descriptive table with visit specific rescue medication use and missingness

Treatment	Visit	n	Mean HbA1c	Rescue in %	Missingness in %
0	0	51	8.14	0.00	0.00
0	1	51	8.18	0.00	1.96
0	2	51	8.11	2.00	1.96
0	3	51	8.26	4.00	1.96
0	4	51	8.39	4.00	1.96
0	5	51	8.22	4.00	1.96
0	6	51	8.51	8.00	1.96
0	7	51	8.34	8.00	1.96
0	8	51	8.44	10.00	1.96
0	9	51	8.23	10.20	3.92
0	10	51	8.20	12.24	3.92
0	11	51	8.54	14.58	5.88
0	12	51	8.37	14.58	5.88
1	0	47	8.08	0.00	0.00
1	1	47	7.99	2.13	0.00
1	2	47	7.91	2.13	0.00
1	3	47	7.87	4.25	0.00
1	4	47	7.87	4.25	0.00
1	5	47	7.82	4.25	0.00
1	6	47	7.71	6.38	0.00
1	7	47	7.98	6.38	0.00
1	8	47	7.92	8.70	2.13
1	9	47	7.81	8.70	2.13
1	10	47	7.99	8.89	4.25
1	11	47	7.86	13.33	4.25
1	12	47	7.89	13.33	4.25

3.6 Case study

Below is a presentation of one instance of the data generating mechanism under the core scenario and the application of the presented analysis methods. Data is generated with the parameters from the core scenario. Summary of visit specific variables grouped by visit and treatment received, such as the proportion of patients receiving rescue medication, the proportion of missing observations, and the mean of the outcome variable, can be found in Table 3.5.

By the final visit, the cumulative amount of rescue medication use reaches approximately 14%. Similarly, the proportion of missing data increases gradually over time and remains moderate, reaching approximately 5% at the final visit. From the table it is clear to see the monotone mechanisms in both use of rescue medication and missingness.

The groups are fairly equal measured on the randomisation, with 51 control group patients and 47 treatment group patients. For a visual insight in how the mean HbA1c progress over time, see Figure 3.8. This figure also contains all the individual trajectories of the patients. It is clear to see that there is a lot of variation in the individual paths, which is of course also something that the models are required to handle. This would not be expected in a real conducted diabetes trial, due to the natue of the HbA1c biomarker, which can be interpreted as the mean level of blood sugar over the last three months. However, since this is just a trial with a continuous endpoint, and one can imagine a more versatile endpoint, this is not impossible in practice.

In Table 3.6 the results of the methods are presented. The code for the analysis functions can be found in the public project repository at `https://github.com/CI-RCT-Sim/CI.RCT.Sim`, in the files named analyse_diabetes_XX.R, where XX is the analysis method.

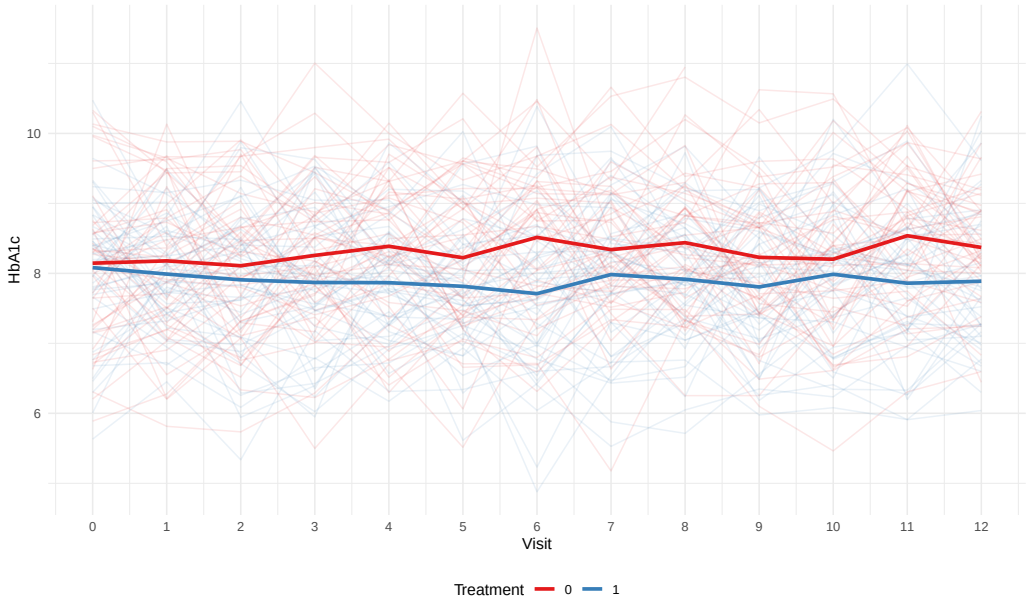


Figure 3.8: Trajectories for the patients in the treatment and control group respectively.

Table 3.6: Table with analysis output for each method

method	Estimate	Standard error	95% CI lower	95% CI upper	p-value
ipwtp	-0.485	0.183	-0.848	-0.121	0.005
mmrmtp	-0.483	0.179	-0.839	-0.127	0.008
mitp	-0.465	0.181	-0.825	-0.106	0.006
ipwhyp	-0.553	0.200	-0.951	-0.155	0.004
dmhyp	-0.573	0.178	-0.923	-0.224	0.001
gcomhyp	-0.568	0.206	-0.948	-0.122	0.010
mmrmhyp	-0.580	0.191	-0.959	-0.200	0.003
mihyp	-0.511	0.200	-0.913	-0.109	0.007

The true treatment effect in the treatment policy setting is -0.442 in the core scenario, as can be seen in Table B.1, and the true hypothetical treatment effect is -0.492. What is seen in this illustrative data set is, that all estimates are more negative than the simulated effect. They are quite similar in standard error. It is not surprising that the more negative estimates of the true treatment effect in this case also leads to rejection of the null hypothesis by all the analysis methods, as well as none of the confidence intervals contains an estimate of 0. From the trajectories in Figure 3.8 it is also clear to see that the two groups split quite early. One could maybe imagine two groups that are more similar, and then the treatment effect could be harder to detect.

3.7 Discussion

Taken together, the results indicate that under the treatment policy estimand, all investigated methods provide approximately unbiased and comparably precise estimates for moderate missingness. In contrast, under the hypothetical estimand, both bias and variability increase and differ more strongly between methods. Performance deteriorates particularly in scenarios with high missingness (HM) for both estimands or increased rescue probability (MR) for the hypothetical estimand. The results suggest that MMRM is slightly more robust in these scenarios than the other methods. These results are aligned to what was observed in Siddiqui (2011) for missing data, where the MMRM approach appears to be a better choice in maintaining statistical properties of a test (e.g. type I error rate, power) as compared to the MI approach in dealing with ignorable missing data of clinical trials.

For both estimands it was observed that the methods struggle more and hence show more varying performance in the scenarios with higher missingness. An investigation showed that the 5% missingness that was aimed for was reached, as one could also tell from Table B.1. But this investigation also showed that the distribution across arms was similar, ranging from 20-45% more prevalent in the control group compared to the treatment group, both when it comes to missing data and rescue medication initiation. When the impact is similar in the two groups, the adjustment and modelling has little relevance, which could possibly explain the lack of differentiation in the performance of the methods.

Type I error rate control seems adequate for all methods for the treatment policy estimand and for MMRM, MI and IPW for the hypothetical estimand except under the increased rescue probability (MR) scenario where only IPW achieves adequate control. For the hypothetical estimand DM and GCOM seem consistently slightly inflated. Methods perform very similar concerning power. For the hypothetical estimand DM has a slight advantage in the rescue on top scenario (ROT), else the MMRM method achieves highest power and IPW performs second best in most scenarios.

The findings of this simulation study can be contrasted with the results reported by Olarte Parra et al. (2025), who investigated estimation methods for a hypothetical estimand in a diabetes trial case study. Similar to our findings they observed that the different estimators (MMRM, MI, IPTW, G-formula, and G-estimation) yielded broadly similar treatment effect estimates and standard errors across methods. However, in our simulation study, greater differences between methods are observed for the hypothetical estimand. Parra et al. concluded that MI or causal methods are generally preferable when time varying covariates matter, however, without proof in their simulations (e.g. higher power). In our simulations the MMRM always performs better, even with high rescue, which could be explained by the sparse data generating mechanism, that did not include many time varying covariates except for the outcome under investigation. The specification of the data generating mechanism can influence the interpretation as well, since the functional forms of the models were inspired by the data generating mechanism and vice versa. As previously mentioned, one should be aware that the results of this simulation study should be interpreted within the context of the specific estimands and data-generating mechanisms considered. They are not intended to generalise to other settings where the causal structure, intercurrent events and handling strategies, or modelling assumptions differ.

Especially in the rescue on top (ROT) scenario DM is performing very well with almost no bias and the highest power. This is in agreement with findings in Lasch et al. (2023). However, the overall and general conclusion from this study which often favours MMRM is not in agreement with this source. Differences between the present results and previous studies should be interpreted in light of differences in estimand definitions, data-generating assumptions, and modelling strategies, rather than as contradictions.

In summary, the results of this simulation shows the similarity of estimator performance under favourable conditions, especially for the treatment policy estimand. For the hypothetical estimand, MMRM often has smaller bias and higher power, while IPW has better type I error control in the increased rescue probability scenario. Therefore, IPW might be preferable when type I error control under increased rescue probability is prioritized; MMRM is competitive for bias, RSME, and power in many scenarios.

4 Scenario 3: Preventive vaccine efficacy trial, principal stratum strategy, binary endpoint

In this scenario class, we consider a randomised, placebo-controlled preventive vaccine efficacy trial for a seasonal infection. The vaccination regimen consists of two doses administered 14 days apart (a prime–boost schedule).

In this setting we consider the impact of the intercurrent event that some participants may not complete the two-dose regimen for various reasons (e.g. due to reactogenicity following the first dose). This results in a mixture of participants receiving only dose 1 versus participants completing the full regimen. We allow for *partial vaccine efficacy* after dose 1 and a larger *full-regimen vaccine efficacy* after dose 2. The focus of the simulation is on evaluating causal inference methods that target an effect among participants who would complete the two-dose regimen irrespective of randomised assignment, while allowing for realistic dependence between regimen completion and outcomes through the ICE process.

We limit our investigation to a setting in which the *force of infection* is effectively zero during the first two weeks after dose 1, to avoid additional complications from pre-dose 2 infections and to isolate the impact of compliance effects; this corresponds, for example, to a trial initiated in a pre-season period when infection pressure is negligible, with the force of infection increasing to a low but non-zero level after day 14 for the remainder of follow-up.

Estimand: The considered estimand has the following relevant attributes:

- Endpoint: Indicator of whether infection occurred during trial follow-up.
- Treatments: Vaccination versus placebo.
- Summary measure: Vaccine efficacy (VE), defined as 1 minus the risk ratio.
- Population: Participants who would complete the two-dose regimen under either treatment assignment (the principal stratum of *always compliers*).
- Intercurrent event: Non-completion of the vaccination regimen due to failure to receive dose 2 by day 14 (e.g., following an AE).
- Intercurrent event strategy: Principal stratum strategy, targeting the effect among those who would complete the two-dose regimen under either treatment assignment.

The following causal inference methods are applied in the simulation:

- Instrumental variable (IV) approach using a two-stage Poisson regression (Bowden et al., 2021).
- Principal score weighting (Jo and Stuart, 2009), in two variants: with and without inclusion of the principal-score covariates in the outcome model.

As comparator method we estimate vaccine efficacy using a binomial model (Nauta, 2020) restricted to the (observed) study population of participants that fully adhere to the vaccine schedule. In practice, this is what is frequently used as a primary analysis to estimate the biologic effect (Beckers et al., 2025) as suggested in the Guideline on the Clinical Evaluation of vaccines (European Medicines Agency (EMA), 2023). In line with Horne et al. (2000), we refer to this analysis as per-protocol analysis, acknowledging that in practice other deviations (e.g. use of protocol prohibited medication) may occur that might warrant a different approach than simple exclusion from the analysis set (Beckers et al., 2025).

Principal strata and interpretation: Regimen completion is defined as receipt of the second dose at day 14. The principal strata are defined in terms of potential regimen completion status under both treatment assignments:

- Always compliers: individuals who would complete the two-dose regimen under both vaccine and placebo.
- Vaccine compliers: individuals who would complete the regimen under vaccine but not under placebo.
- Placebo compliers: individuals who would complete the regimen under placebo but not under vaccine.

	Completion under placebo	
Completion under vaccine	**Yes**	**No**
Yes	Always compliers	Vaccine compliers
No	Placebo compliers	Never compliers

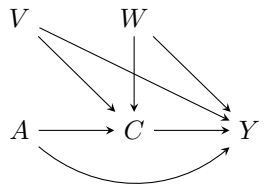


Figure 4.1: Assumed causal structure linking the randomly allocated treatment (A), confounders (V, W), compliance variable (C), and the binary outcome (Y).

- Never compliers: individuals who would fail to complete the regimen regardless of assignment.

The estimand of primary interest targets always compliers, as this is the group for whom the causal effect of completing the full two-dose regimen is most directly interpretable.

In the data-generating model, the force of infection is set to zero up to day 14, so infections do not occur before the regimen completion decision is made. This avoids additional complications arising from early infections prior to the second dose.

Identification and assumptions: Membership in the principal strata is not directly observable, since each participant's regimen completion status is only observed under the assigned treatment. Identification of vaccine efficacy within the always complier stratum therefore relies on assumptions implied by the specified data-generating model and the chosen parameter values. Compliance is modelled dependent on baseline marker status and treatment assignment (e.g. imagining some marker related AE as a trigger of the intercurrent event of non-compliance). In the simulation, this marker-dependent compliance mechanism is the primary source of information available to analytically distinguish the principal strata and to estimate the targeted principal stratum estimand using the proposed methods (e.g. via principal score modelling and IV-type approaches), conditional on the selected parameter values. In addition, we assume in the main scenarios that infection status is ascertained without diagnostic misclassification (i.e. perfect sensitivity and specificity of case detection). While outcome misclassification can be relevant in vaccine trials, it is not considered in the present simulation study.

Missing data: In this simulation scenario, we do not model missing primary endpoint data (e.g., loss to follow-up or withdrawal of consent) and assume complete ascertainment of infection status at the end of follow-up. This choice is made to maintain focus on the study objective—quantifying the impact of (informative) non-compliance with the two-dose regimen and comparing estimators for the corresponding principal stratum estimand—without introducing additional mechanisms that would require separate parametrisation and could confound interpretation of compliance-related effects.

4.1 Data generation

For data generation in this scenario, we consider a simple vaccine trial set-up with a binary primary endpoint (infection by end of follow-up), derived from an underlying time-to-event process, and the intercurrent event of not receiving the second dose. The corresponding causal structure is illustrated in Figure 4.1.

We assume that participants $i = 1, \ldots, n$ are randomised between a vaccine and a control treatment. Treatment assignment is indicated as $A_i \in \{0, 1\}$, with $A_i = 1$ indicating vaccine treatment and $A_i = 0$ indicating control.

Dose 1 is administered at time $t = 0$. The second dose is scheduled 14 days later, i.e. at $t = 14$ days. Participants are followed until administrative end of study at time τ (e.g., $\tau = 180$ days), or until infection occurs.

Confounders

The simulation includes two binary baseline markers $V_i, W_i \in \{0, 1\}$, which may act as a prognostic factor and/or an effect modifier. We generate

$$V_i \sim \text{Bernoulli}(p_V), \quad W_i \sim \text{Bernoulli}(p_W),$$

independently.

We consider V_i to potentially be prognostic for both intercurrent event and infection risk and W_i to, additionally, have effect modification potential. To evaluate method performance under various situations of observed and unobserved confounding and/or effect modifying variables, we consider scenarios where either V_i, W_i, or both are observed.

Force of infection

The force of infection (i.e. baseline incidence rate) is represented through a baseline hazard that is zero in the first two weeks and constant thereafter. We define

$$\lambda_0(t) = \begin{cases} 0, & 0 < t \leq 14, \\ \lambda_{\text{post}}, & 14 < t \leq \tau, \end{cases}$$

where λ_{post} is chosen to reflect realistic incidence scenarios (e.g., corresponding to 1 in 500, 1 in 1000, or 1 in 2000 patient-years).

Compliance

Compliance with the vaccination schedule is defined as taking both doses, i.e. receiving dose 2 at day 14. Let $C_i \in \{0, 1\}$ denote compliance, where $C_i = 1$ indicates that participant i receives dose 2 and $C_i = 0$ indicates non-compliance. Conceptually, non-compliance can be viewed as being triggered by some event occurring between dose 1 and dose 2 (e.g. reactogenicity or other adverse events), but for data generation we model compliance directly as a binary outcome.

We model compliance as a deterministic function of the potential treatment condition $a \in \{0, 1\}$, the baseline covariates (V_i, W_i), and a participant-specific latent variable U_i. Specifically, let $p(a, V, W)$ be a compliance-propensity function taking values in $[0, 1]$ and $C(p, U)$ a compliance function taking values in $\{0, 1\}$, and write

$$p_i(a) := p(a, V_i, W_i), \qquad C_i(a) := C\big(p_i(a),\, U_i\big),$$

where higher propensities imply a higher probability to comply. We refer to $C_i(a)$ as *potential* compliance under a (possibly counterfactual) treatment status $a \in \{0, 1\}$. The propensity function is specified by the logistic model

$$\text{logit}\{p_i(a)\} = \gamma_0 + \gamma_W W_i + \gamma_V V_i + \gamma_A a + \gamma_{AW} a W_i. \tag{1}$$

Note that the compliance propensity is deterministic and can be evaluated for either potential treatment condition a for each individual given their covariate information. The parameters γ_A, γ_W, γ_V, and γ_{AW} allow compliance propensities to differ between treatment groups, marker status, and risk class, respectively; for simplicity we only include an interaction between potential treatment condition a and W_i.

As an example interpretation, V_i may represent membership in a specific risk group (e.g. healthcare workers) which may be associated with higher infection risk but also higher willingness to comply with the vaccination schedule (modelled through $\gamma_V > 0$). W_i may represent an (unobserved) immune-responsiveness (reactogenicity propensity) trait, where $W_i = 1$ indicates higher reactogenicity (reducing compliance in vaccinated, e.g., $\gamma_W = 0$ and $\gamma_{AW} < 0$) and potentially stronger vaccine-induced protection (effect modification). While such immune-responsiveness may typically be unobserved, in practice it could be approximated by observed baseline characteristics that are associated with both reactogenicity and immune response, for example sex (and/or younger age group).

Coupled generation of potential compliance outcomes. We generate potential compliance outcomes using a participant-specific latent threshold variable

$$U_i \sim \text{Uniform}(0,1),$$

generated once per participant and held fixed across potential outcomes. Potential compliance is then generated as

$$C_i(a) = \mathbb{1}\{U_i < p_i(a)\}, \qquad a \in \{0,1\}. \tag{2}$$

This construction ensures that, conditional on (V_i, W_i), $C_i(a) \sim \text{Bernoulli}\{p_i(a)\}$ marginally over U_i (i.e. $\Pr(C_i(a) = 1 \mid a, W_i, V_i) = p_i(a)$), while the pair $(C_i(0), C_i(1))$ is coupled through the common U_i. The observed compliance indicator is

$$C_i = C_i(A_i).$$

This ensures that for $p_i(0) > p_i(1)$ the corresponding potential compliance outcomes satisfy $C_i(0) \geq C_i(1)$ (and vice versa), which under appropriate parameter settings induces monotonicity (e.g. assuming that participants who would comply under treatment would also comply under control).

Infection hazard

The individual infection hazard $\lambda_i(t)$ is specified via a proportional hazards model driven by $\lambda_0(t)$. The baseline hazard $\lambda_0(t)$ is modelled as a piece-wise constant function with $\lambda_0(t) = 0$ for $t < 14$ and $\lambda_0(t) = \lambda_{\text{post}} > 0$ for $t \geq 14$. Vaccine protection is allowed to differ depending on whether only dose 1 is received or the full two-dose schedule is completed. Let C_i denote compliance with dose 2 (i.e. receipt of dose 2 at day 14). We write

$$\lambda_i(t) = \lambda_0(t) \exp\Big(\beta_V V_i + \beta_W W_i + A_i\, \theta_i(t)\Big), \tag{3}$$

$$\theta_i(t) = \theta^{(1)}(W_i) + \mathbb{1}(t \geq 14)\, C_i \Big\{\theta^{(2)}(W_i) - \theta^{(1)}(W_i)\Big\}, \tag{4}$$

$$\theta^{(d)}(W_i) = \beta_A^{(d)} + \beta_{AW}^{(d)}\, W_i, \qquad d \in \{1,2\}. \tag{5}$$

Here, $\theta^{(d)}(W_i)$ denotes the dose-specific treatment effect on the log-hazard scale for subjects with baseline marker value W_i, allowing the main vaccine effect to differ after dose 1 versus after dose 2 through $\beta_A^{(1)}$ and $\beta_A^{(2)}$. Effect modification by the marker is governed by the interaction parameter $\beta_{AW}^{(d)}$. For simplicity we consider only two types of scenarios where either $\beta_A^{(1)} = \beta_{AW}^{(1)} = 0$ and $\beta_{AW}^{(2)} = \beta_{AW}$, indicating that only subjects receiving both doses achieve some protection, or $\beta_A^{(1)} < 0$ and $\beta_{AW}^{(1)} = \beta_{AW}^{(2)} = \beta_{AW}$, assuming that effect modification does not depend on dose status. In particular, for dose status $d \in \{1,2\}$ we have $\theta^{(d)}(0) = \beta_A^{(d)}$ for marker-negative subjects and $\theta^{(d)}(1) = \beta_A^{(d)} + \beta_{AW}$ for marker-positive subjects.

The time-varying term $\theta_i(t)$ equals $\theta^{(1)}(W_i)$ prior to day 14 and switches to $\theta^{(2)}(W_i)$ from day 14 onwards only for participants who receive dose 2 ($C_i = 1$); otherwise it remains $\theta^{(1)}(W_i)$. The parameters β_V and β_W represent prognostic effects of V_i and W_i, respectively; if these prognostic effects are not required they are set to zero.

Event-time generation and observed endpoint

Event times for infection and the intercurrent event are generated using inverse-CDF sampling under the models specified above.

For each participant, we simulate baseline markers V_i and W_i. Compliance with dose 2 C_i is then generated according to the logistic model above and the infection time T_i under the corresponding infection hazard $\lambda_i(t)$.

The observed binary infection endpoint is defined as

$$Y_i = \mathbb{1}(T_i \leq \tau),$$

i.e. $Y_i = 1$ if infection occurs before administrative end of follow-up at time τ, and $Y_i = 0$ otherwise.

4.1.1 Sample size

The per-group sample size was chosen to provide approximately 80% power for the super-superiority test of $\mathrm{VE}_{\mathrm{AC}} > 30\%$, assuming a true vaccine efficacy of 70% after the second dose, under the post-day-14 incidence rate λ_{post}. At incidence levels of 1/500, 1/1000, and 1/2000 per patient-month, this resulted in 3815, 7639, and 15287 participants per treatment group, respectively.

4.1.2 True treatment effect

Define the principal stratum of *always-compliers* as

$$\mathcal{S}_{\mathrm{AC}} = \{C(0) = 1,\ C(1) = 1\}.$$

Under the coupled compliance generation $C(a) = \mathbb{1}\{U < p(a \mid V, W)\}$ with $U \sim \mathrm{Unif}(0,1)$, where

$$\mathrm{logit}\,\{p(a \mid v, w)\} = \gamma_0 + \gamma_W w + \gamma_V v + \gamma_A a + \gamma_{AW} a w,$$

we have

$$\Pr(\mathcal{S}_{\mathrm{AC}} \mid v, w) = m(v, w), \qquad m(v, w) := \min\{p(0 \mid v, w),\, p(1 \mid v, w)\}.$$

Hence, using $\Pr(V = 1) = p_V$, $\Pr(W = 1) = p_W$ and independence of V and W, the stratum distribution is

$$\Pr(v, w \mid \mathcal{S}_{\mathrm{AC}}) = \frac{\Pr(V = v)\Pr(W = w)\, m(v, w)}{\sum_{v' \in \{0,1\}} \sum_{w' \in \{0,1\}} \Pr(V = v')\Pr(W = w')\, m(v', w')}.$$

Let $\Delta = \tau - 14$. Because $\lambda_0(t) = 0$ for $t \le 14$ and $\lambda_0(t) = \lambda_{\mathrm{post}}$ for $14 < t \le \tau$, the infection risk by τ for the control group $a = 0$ is given by

$$\pi_0(v, w) := \Pr\{Y(0) = 1 \mid v, w, \mathcal{S}_{\mathrm{AC}}\} = 1 - \exp\Big(- \lambda_{\mathrm{post}} \Delta\, e^{\beta_V v + \beta_W w} \Big).$$

For $a = 1$ (vaccine), always-compliers satisfy $C(1) = 1$ and therefore receive the dose-2 effect from day 14 onward, i.e. $\theta^{(2)}(w) = \beta_A^{(2)} + \beta_{AW}^{(2)} w$, yielding

$$\pi_1(v, w) := \Pr\{Y(1) = 1 \mid v, w, \mathcal{S}_{\mathrm{AC}}\} = 1 - \exp\Big(- \lambda_{\mathrm{post}} \Delta\, e^{\beta_V v + \beta_W w + \theta^{(2)}(w)} \Big).$$

The true principal-stratum risks are obtained by standardisation over (V, W) within $\mathcal{S}_{\mathrm{AC}}$:

$$\Pr\{Y(a) = 1 \mid \mathcal{S}_{\mathrm{AC}}\} = \sum_{v \in \{0,1\}} \sum_{w \in \{0,1\}} \pi_a(v, w)\, \Pr(v, w \mid \mathcal{S}_{\mathrm{AC}}), \qquad a \in \{0, 1\}.$$

The true relative risk (RR) in always-compliers is then

$$\mathrm{RR}_{\mathrm{AC}} = \frac{\Pr\{Y(1) = 1 \mid \mathcal{S}_{\mathrm{AC}}\}}{\Pr\{Y(0) = 1 \mid \mathcal{S}_{\mathrm{AC}}\}},$$

and the corresponding vaccine efficacy is computed as $\mathrm{VE}_{\mathrm{AC}} = 1 - \mathrm{RR}_{\mathrm{AC}}$.

4.1.3 Simulation scenarios and parameter ranges

We organise simulations into a small number of scenario classes that vary key features affecting identifiability and robustness, while avoiding an exhaustive factorial combination of all parameter values. The parameter values shown in Table 4.1 should be viewed as interpretable suggestions; where needed, they were calibrated within the implemented data-generating models to achieve realistic target quantities such as overall compliance and cumulative incidence. Where informative, we also expanded the parameter ranges in Table 4.1 beyond the primary, realistic settings to construct "stress-test" scenarios expected to challenge assumptions and induce method failure. The parameter values actually used in each scenario class are reported in the corresponding subsection of the results (subsection 4.4); the full set of investigated configurations, including additional scenarios not shown there, is provided as supplemental spreadsheets (`results_vaccine_scenarios_[scenarioname].xlsx`).

- **Scenario A1: Benchmark (no heterogeneity; compliance independent of treatment).** Baseline markers are inactive and compliance is generated to be similar across randomised arms. This scenario provides a reference setting where reference analysis methods are expected to perform well, and operating characteristics are primarily driven by the incidence and efficacy settings.
- **Scenario A2: Compliance driven by vaccination (treatment-dependent compliance, no baseline confounding).** Baseline markers remain inactive, but compliance differs by randomised assignment, e.g. representing intercurrent events that occur primarily in the vaccine arm. This scenario isolates the impact of treatment-dependent compliance on estimation and testing, without introducing baseline covariate–driven confounding.
- **Scenario B: Prognostic-only baseline heterogeneity via V.** V is active as a prognostic factor for infection risk and may also influence compliance, while W remains inactive. This scenario class evaluates method performance under baseline risk heterogeneity that is not intrinsically related to differential compliance or vaccine benefit. A realistic motivating example is a subgroup with higher exposure and infection risk but also higher compliance (e.g., healthcare workers), which can be represented by allowing V to increase.
- **Scenario C: Predictive-only heterogeneity via W (effect modification without baseline prognostic impact).** W is active only through heterogeneity in vaccine protection and may impact compliance in vaccinated individuals, but is assumed not to affect baseline infection risk and compliance. This reflects settings where a baseline factor is predictive of vaccine benefit and potentially compliance under vaccination, while not being prognostic for infection in the absence of vaccination. This scenario can be motivated by a subset with higher propensity for adverse events (e.g., reactogenicity) that reduces dose-2 compliance, while simultaneously being predictive of stronger vaccine-induced protection (i.e., higher vaccine efficacy in that subset).
- **Scenario D: Combined $V+W$ scenario (concurrent prognostic and predictive heterogeneity).** A targeted combined scenario was included where V drives baseline risk and compliance heterogeneity and W drives protection and compliance heterogeneity. This scenario was intended as a representative stress-test and was not implemented across the full combination of all parameter options.
- **Additional complex and stress-test scenarios.** Additional scenarios were analysed outside the main A–D structure to evaluate robustness under more challenging or less central conditions. These include lower infection incidence, near-positivity violations due to very high or very low compliance in marker-defined subgroups, stronger treatment-dependent non-compliance, stronger dose-one protection among non-compliers, and combinations of unobserved prognostic and predictive markers.
- **Observed-data variants.** Robustness to limited covariate availability was assessed by reanalysing the same simulated datasets under different observed covariate sets (e.g., observing both markers, only one, or neither). This induces residual confounding and/or effect-modification, without requiring separate data generation.

Across these scenario classes, incidence and efficacy settings (and, where applicable, compliance patterns) were varied in a small number of interpretable configurations, using Table 4.1 as a starting point and calibrating as needed to maintain realistic operating characteristics consistent with the causal data-generating assumptions in Figure 4.1. Scenarios B and C are used to develop case studies for illustration, as described in Section 6.2 in Ristl et al. (2026).

4.2 Analysis methods

Let S_{AC} denote the principal stratum of always compliers. The vaccine efficacy in this stratum is

$$\mathrm{VE}_{\mathrm{AC}} = 1 - \frac{\Pr\{Y(1) = 1 \mid S_{\mathrm{AC}}\}}{\Pr\{Y(0) = 1 \mid S_{\mathrm{AC}}\}}.$$

This estimand represents the causal proportionate reduction in infection probability among those who would comply with the vaccination schedule under either assignment.

Table 4.1: Simulation parameters for the preventive vaccine trial data-generating model.

Parameter	**Short description**	**Scenario values**
n	Total sample size	Chosen to provide $\approx 80\%$ power to demonstrate $> 30\%$ VE assuming a vaccine efficacy of 70% under given incidence rate λ_{post}
p_V	Prevalence of baseline marker V_i	(0, 0.10, 0.30)
p_W	Prevalence of baseline marker W_i	(0, 0.10, 0.30)
λ_{post}	Baseline hazard after day 14	Chosen to match incidence (patient-month) level: (1/500, 1/1000, 1/2000)
β_V	Prognostic effect of V on infection hazard (log-HR)	$\log(1.5) = 0.405$, or 0 for no prognostic effects.
β_W	Prognostic effect of W on infection hazard (log-HR)	$\log(1.2) = 0.182$, or 0 for no prognostic effects.
$\beta_A^{(2)}$	Two-dose vaccine efficacy after day 14 among compliers ($C_i = 1$)	Chosen to yield VE of (0%, 50%, 70%, 90%).
$\beta_A^{(1)}$	One-dose vaccine efficacy (pre-day 14 for all; post-day 14 for non-compliers)	Chosen to yield (never-complier) VE of (0%, 30%, 50%, 60%) for corresponding $\beta_A^{(2)}$, or 0% throughout for scenarios assuming no effect of Dose 1
β_{AW}	Effect modification by W (common across dose status)	$\log(0.8) = -0.223$, or 0 for no effect modification.
γ_0	Baseline compliance	Chosen to achieve typical two-dose completion ($\approx 95\%$ overall), given the remaining γ parameters.
γ_A	Treatment effect on compliance (log-odds)	-0.357 (Odds Ratio ≈ 0.7) or 0 for no treatment effect on compliance.
γ_V	Effect of V on compliance (log-odds)	0.5 (Odds Ratio ≈ 1.65), or 0 for no impact on compliance.
γ_W	Effect of W on compliance (log-odds)	-0.8 (Odds Ratio ≈ 0.45), or 0 for no impact on compliance.
γ_{AW}	Interaction of treatment and W on compliance (log-odds)	-0.3 (Odds Ratio ≈ 0.74), or 0 for effect modification.

For all methods, the analysis is restricted to information actually available in a given scenario: when a marker is treated as unobserved, it is not included in any analysis model. Scenarios with different observed/unobserved configurations are analysed in analogy.

Testing procedure (super-superiority test). All three methods test the super-superiority hypothesis

$$H_0 : \mathrm{VE}_{\mathrm{AC}} \leq 0.3 \quad \text{versus} \quad H_1 : \mathrm{VE}_{\mathrm{AC}} > 0.3$$

on the log relative-risk scale, one-sided at the 2.5% significance level. The 0.3 threshold matches the addition of VE = 30% to the parameter grid (see subsubsection 4.1.3). Empirical type I error is therefore evaluated under scenarios with true $\mathrm{VE}_{\mathrm{AC}} = 0.3$, and empirical power under scenarios with $\mathrm{VE}_{\mathrm{AC}} > 0.3$. Note that, as a consequence, scenarios with true $\mathrm{VE}_{\mathrm{AC}} < 0.3$ (including $\mathrm{VE}_{\mathrm{AC}} = 0$) are expected to yield empirical rejection rates well below the nominal level.

4.2.1 Instrumental variable (IV)

The instrumental variable approach uses randomised treatment allocation as the instrument for actual receipt of the full vaccination regimen. This approach identifies principal stratum effects under the assumptions of relevance, randomisation, the exclusion restriction, and monotonicity, and does not rely on no unmeasured confounding (Bowden et al., 2021; Angrist et al., 1996; Jiang and Ding, 2021; Frangakis and Rubin, 2002).

Let $T_i = \mathbb{1}(A_i = 1, C_i = 1)$ denote the endogenous treatment variable indicating receipt of the full two-dose regimen, with A_i used as the instrument. Where available, the observed baseline markers V_i and W_i enter the analysis as exogenous covariates.

In the protocol, a linear two-stage IV estimator using `ivreg::ivreg` was proposed. In preparatory runs this estimator was found to be systematically biased for the risk-ratio target on which vaccine efficacy is defined, suspected to be due to using a linear model for a binary outcome on the risk-ratio scale. The estimator was therefore replaced by a custom two-stage implementation with a Poisson outcome model in the second stage (Mullahy, 1997). Specifically:

1. **Stage 1.** A linear regression (via `lm.fit`) of T_i on the instrument A_i and the included exogenous covariates yields fitted values $\widehat{T}_i$. The intercept and the exogenous covariates are passed through stage 1 unchanged (i.e. they enter stage 2 as themselves).
2. **Stage 2.** A Poisson regression with log link of Y_i on $\widehat{T}_i$ and the same exogenous covariates is fitted. Vaccine efficacy is estimated by contrasting the marginal predicted risks at $\widehat{T}_i = 1$ and $\widehat{T}_i = 0$, obtained via `emmeans::emmeans` on the response (risk-ratio) scale.

The point estimate is $\widehat{\mathrm{VE}}_{\mathrm{IV}} = 1 - \widehat{\mathrm{RR}}_{\mathrm{IV}}$, and confidence interval limits are obtained by transforming the asymptotic confidence limits for $\log \widehat{\mathrm{RR}}_{\mathrm{IV}}$ returned by `emmeans`. The super-superiority test is performed by `emmeans` with `null`$= \log(1 - 0.3)$ on the log-RR scale, one-sided.

4.2.2 Principal score weighting

The principal score approach (Jo and Stuart, 2009) approximates the causal effect among always compliers by reweighting observed compliers in the control arm. The approach relies on principal ignorability (stratum membership independent of potential outcomes conditional on the observed covariates) and monotonicity. Following Jo and Stuart (2009), monotonicity is imposed in the direction of *no vaccine compliers*: vaccination does not induce regimen completion in anyone who would not also complete under control, i.e. $C_i(1) \leq C_i(0)$, corresponding to $\gamma_A + \gamma_{AW} W_i \leq 0$ in the data-generating model. Under this assumption the treated compliers ($A_i = 1,\ C_i = 1$) are exactly the always compliers, whereas the control compliers ($A_i = 0,\ C_i = 1$) are a mixture of always compliers and placebo compliers. The always-complier effect is therefore identified by taking the treated compliers as observed and reweighting the control compliers to the always-complier covariate distribution.

The principal score model is a logistic regression of compliance C_i on the observed baseline markers, fit using treated participants only (in whom C_i corresponds to $C_i(1)$):

$$\mathrm{logit}\{\Pr(C_i = 1 \mid X_i)\} = \beta_0 + \beta^\top X_i,$$

where X_i contains the observed subset of $\{V_i, W_i\}$. An intercept-only model is used when neither marker is observed. The fitted model predicts $\widehat{p}_i = \widehat{\Pr}(C_i(1) = 1 \mid X_i)$ — under the assumed monotonicity, the probability of always compliance — and the corresponding principal-score odds $\widehat{o}_i = \widehat{p}_i/(1-\widehat{p}_i)$ for all participants. Weights are assigned as

$$w_i = \begin{cases} 1, & A_i = 1,\ C_i = 1, \\ 0, & A_i = 1,\ C_i = 0, \\ 0, & A_i = 0,\ C_i = 0, \\ \widehat{o}_i, & A_i = 0,\ C_i = 1. \end{cases}$$

Two variants of the outcome model are considered, both fit as weighted logistic regressions of Y_i on A_i with weights w_i:

1. **Unadjusted outcome model.** The outcome model contains only the intercept and A_i. This is the procedure as originally specified in the protocol.
2. **Adjusted outcome model.** The outcome model additionally contains the same observed baseline markers as the principal-score model. This variant was added during implementation and follows the spirit of doubly-robust weighted-regression approaches (Bang and Robins, 2005).

Marginal predicted infection risks under $A = 0$ and $A = 1$ are obtained from the weighted outcome model using `emmeans`, with the contrast taken on the log scale to give the estimated risk ratio $\widehat{\mathrm{RR}}_{\mathrm{PS}}$. Vaccine efficacy is $\widehat{\mathrm{VE}}_{\mathrm{PS}} = 1 - \widehat{\mathrm{RR}}_{\mathrm{PS}}$. Confidence intervals are obtained by transforming the asymptotic limits for $\log \widehat{\mathrm{RR}}_{\mathrm{PS}}$, and the super-superiority test is performed on the log-RR scale as in the IV analysis.

4.2.3 Per-protocol analysis

In practice, the primary analysis for vaccine efficacy is usually based on the per-protocol set (Horne et al., 2000; Baden et al., 2021; Polack et al., 2020). As comparator method, we, therefore, estimate vaccine efficacy restricted to participants who adhere to vaccination according to their observed compliance status, using a Poisson model for the risk-ratio. Such an analysis, limited to "observed compliers", can be considered a "pragmatic" approach to approximate the principal stratum estimand implied by the "biological effect" (Beckers et al., 2025; Michiels et al., 2022; Fay et al., 2026). Considering, it is subject to selection bias unless restrictive and implausible assumptions apply (e.g. assuming that participants who received the two doses in one arm would have done so in other arm). Evaluating the potential bias of such an analysis under various scenarios of non-compliance therefore represents a practically relevant objective.

4.3 Deviations from the study protocol

The implemented simulation deviates from the originally proposed parameter grid in three respects. First, the protocol proposed VE values of 0%, 50%, 70% and 90%. A value of 30% was added, since this is the VE at the margin of the super-superiority hypothesis being tested (see subsection 4.2). Second, the grid of parameter values was expanded such that all VE values were investigated for all considered combinations of the remaining parameters. Third, an additional scenario with lower overall compliance of 50% was included.

For the comparator method an analysis with a Poisson model was used instead of the exact binomial model proposed in Nauta (2020). The latter is an exact binomial test for the probability of an infected participant belonging to the treatment group conditional on the total number of events. Confidence bounds and hypotheses tests are performed using the exact Clopper-Pearson method. Preference was given to the Poisson model as it corresponds to the second stage of the IV regression. Vaccine efficacy was estimated with `emmeans` as one minus the marginal mean risk ratio.

The model for the second stage of the instrumental variable regression was changed to a Poisson regression model after strong bias was observed for the initially proposed linear-probability model in early simulations. The implementation was subsequently changed from using the package `ivreg` to a manual implementation. (See Appendix C)

Table 4.2: Analysis methods, with a short description, the summary measure reported, the label used in the results figures, and the software implementation.

Method	Label	Description	Summary measure	Implementation
Instrumental variable (IV)	`iv`	Allocated treatment as instrument for full-regimen receipt; two-stage estimator	Vaccine efficacy	`stats::lm.fit` (stage 1) + `stats::glm` (Poisson, stage 2), `emmeans::emmeans`
Principal score weighting, adjusted outcome model	`ps_cov`	Weights from principal-score model (Jo and Stuart, 2009); outcome model additionally includes the baseline markers	Vaccine efficacy	`stats::glm` (logistic score and weighted outcome model), `emmeans::emmeans`
Principal score weighting, unadjusted outcome model	`ps_nocov`	As above, but the outcome model contains only the intercept and treatment	Vaccine efficacy	`stats::glm` (logistic score and weighted outcome model), `emmeans::emmeans`
Per-protocol analysis	`pp`	Comparator method; Poisson model, restricted to observed compliers	Vaccine efficacy	`stats::glm`, `emmeans::emmeans`

For both the IV and the principal-score methods, point estimates, confidence intervals, and the super-superiority test are obtained from `emmeans` applied to the fitted second-stage / outcome model. This is a deviation from the protocol, which specified "robust variance covariance matrix estimates" and "delta-method" inference since `emmeans` uses the model-based variance covariance estimates per default. Furthermore `emmeans` returns one-sided confidence intervals when using a one-sided test which led to one-sided 95% confidence intervals being calculated instead of two-sided 95% confidence intervals. The implications for empirical coverage and rejection probabilities are discussed in section 4.4.5.

The labelling of the scenario classes in this report differs slightly from the protocol: the protocol's classes *A0* and *A1* are referred to as *A1* and *A2* here. In addition, scenario classes B, C, and D appear as *B1*, *C1*, and *D1* in some figures and parameter tables, reflecting the labels used when generating the parameter configurations and carried through to plot generation; the trailing "1" carries no further meaning.

The additional stress-test scenarios (subsubsection 4.4.5) are single parameter perturbations of the category D baseline scenario: the direction of an individual compliance, prognostic, or predictive effect is reversed, the dose-one effect is switched off (so that dose 1 confers no protection, against a protective dose-one effect in the baseline), or overall completion is reduced to 50%. The magnitudes of the marker effects on compliance, infection risk, and vaccine efficacy were not themselves varied beyond the values used in the main scenarios.

4.4 Results

Results are presented graphically by scenario category. Within each category, scenarios are identified on the x-axis by their parameter configuration. Numerical summaries for all scenarios and methods are provided as a supplementary table.

The four analysis methods of subsection 4.2 are referred to throughout by short labels: `iv` (instrumental variable), `pp` (per-protocol), and the two principal-score weighting variants `ps_cov` (adjusted outcome model, additionally including the baseline markers) and `ps_nocov` (unadjusted outcome model). All target the always-complier vaccine efficacy VE_{AC} defined in subsection 4.2; this is the quantity reported as VE_{PS} (principal-stratum vaccine efficacy) in the parameter tables, the two notations being equivalent.

4.4.1 Scenario category A

Scenario category A assessed estimator behaviour under benchmark conditions (A1) and under treatment-dependent compliance (A2), without baseline marker-driven compliance, prognosis or effect modification. Figures 4.2–4.3 present the operating characteristics for both scenario classes.

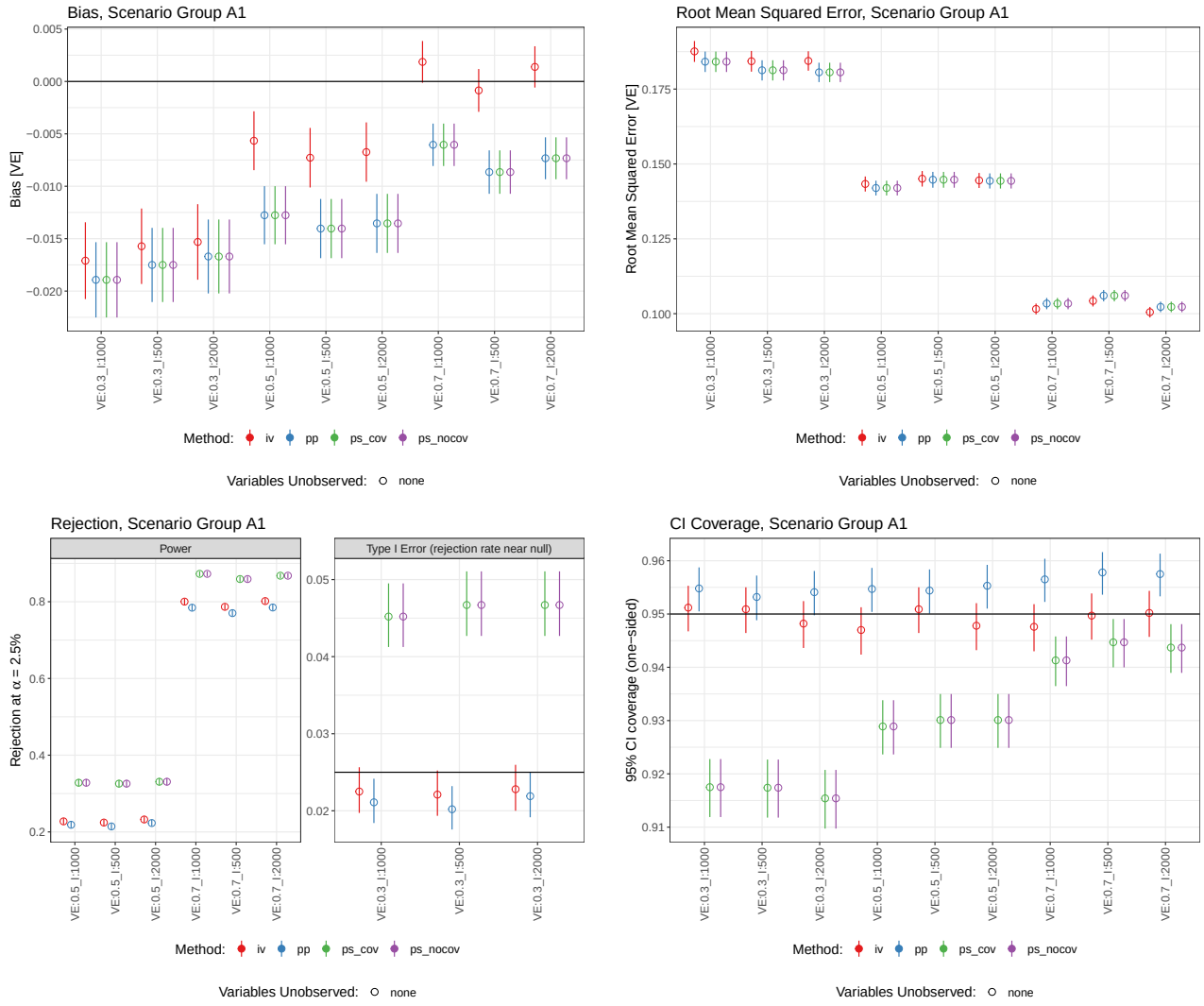


Figure 4.2: Operating characteristics for scenario category A1 (benchmark). Bias (top left), RMSE (top right), rejection probability including power and type I error (bottom left), and 95% CI coverage (bottom right). Reference lines denote zero bias, nominal 2.5% type I error, and nominal 95% coverage respectively.

Scenarios A1: Under benchmark conditions (A1), the causal structure did not include marker-driven selection into compliance, and no subjects are V- or W-positive in any of the scenarios. The identifying assumptions of all estimators therefore hold, and specifications that include covariate effects reduce to intercept-only models. Performance consequently shows no dependence on observation status, as expected, so we report only the results in which all covariates are treated as observed.

This is the only scenario class where we show results from scenarios assuming different incidence rates. The scenario label I:n indicates the baseline *monthly* incidence after day 14, i.e. 1 infection per n subjects per month. We limit presentation to values of $\beta_{A2} \in \{0.3, 0.5, 0.7\}$ (i.e. full-dose hypothetical vaccine efficacy) to keep presentation compact.

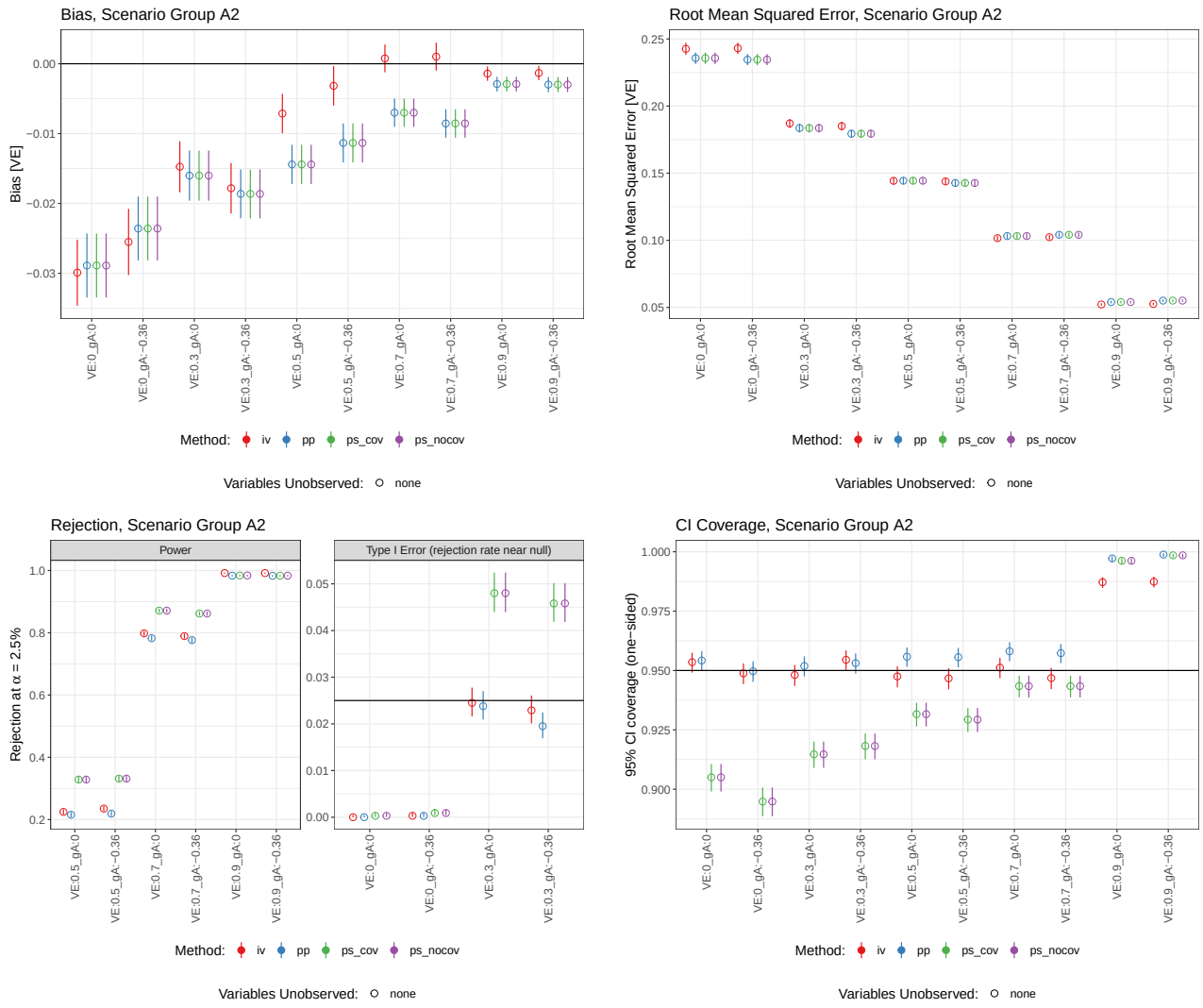


Figure 4.3: Operating characteristics for scenario category A2 (treatment-dependent compliance). Bias (top left), RMSE (top right), rejection probability including power and type I error (bottom left), and 95% CI coverage (bottom right).

Table 4.3 summarises the parameter configurations included in the results for scenario group A1. Each row corresponds to one scenario, identified by its label in the first column. The incidence column gives the baseline infection rate expressed as the number of subjects per infection per month. Parameters that vary across scenarios within the group are shown in bold; fixed parameters are shown for reference. The final column gives the true principal stratum vaccine efficacy VE_{PS} for each scenario configuration.

Table 4.3: Parameter configurations for scenario group A1.

Scenario	**Incidence**	β_{A2} **(VE Scale)**	C	δ_1	γ_A	γ_V	γ_W	γ_{AW}	β_V	β_W	β_{AW}	p_V	p_W	VE_{PS}
VE:0.3_I:1000	**1000**	**0.3**	0.95	T	0	0	0	0	0	0	0	0	0	**0.299**
VE:0.3_I:2000	**2000**	**0.3**	0.95	T	0	0	0	0	0	0	0	0	0	**0.300**
VE:0.3_I:500	**500**	**0.3**	0.95	T	0	0	0	0	0	0	0	0	0	**0.299**
VE:0.5_I:1000	**1000**	**0.5**	0.95	T	0	0	0	0	0	0	0	0	0	**0.499**
VE:0.5_I:2000	**2000**	**0.5**	0.95	T	0	0	0	0	0	0	0	0	0	**0.500**
VE:0.5_I:500	**500**	**0.5**	0.95	T	0	0	0	0	0	0	0	0	0	**0.499**
VE:0.7_I:1000	**1000**	**0.7**	0.95	T	0	0	0	0	0	0	0	0	0	**0.699**
VE:0.7_I:2000	**2000**	**0.7**	0.95	T	0	0	0	0	0	0	0	0	0	**0.700**
VE:0.7_I:500	**500**	**0.7**	0.95	T	0	0	0	0	0	0	0	0	0	**0.699**

Bias: Bias on the VE scale was modest but noticeably different from zero for all methods. Bias is negative indicating underestimation of true principal stratum vaccine efficacy. It is decreasing in size with increasing vaccine efficacy. The largest absolute bias is around negative 3% in settings with zero vaccine efficacy (not shown). Around the null hypothesis ($\beta_{A2} = 0.3$) the bias is around -2 to -1.5 percentage points VE for all approaches. The extent of (absolute) bias decreases with increasing VE. Baseline incidence does not have a noticable impact on bias.

The instrumental variable approach `iv` appears to have a slightly smaller (absolute) bias if VE is larger than zero. This is noticeable for vaccine efficacies above 0.3. Principal scoring approaches `ps_cov` (which uses covariates in the outcome model) and `ps_nocov` which uses covariates only for estimating the principal score, perform almost identical to the per-protocol estimator `pp`.

RMSE: RMSE decreases with increasing VE with minimally larger values for the instrumental variable approach for vaccine efficacies up to 0.5 after which RMSE appears numerically smaller. This aligns with the close to zero bias of this method for scenarios with vaccine efficacy 0.7. The other methods, per-protocol estimator, principal score estimator - with and without covariates in the main model - appear to perform identical in terms of RMSE. Observed status of covariates has no noticable impact on method performance in terms of RMSE (not shown).

Rejection rates: In terms of power there is little difference between the per protocol estimator, and the instrumental variable approach with a slight numerical advantage for the latter. Both provide about 80% power for the scenario with vaccine efficacy of 0.7. Which is in line with planning assumptions. The principal score methods have a noticable power advantage (around 7% point increase). For scenarios close to the null hypotheses (see limitations) instrumental variable and per-protocol estimators control the rejection rate at or just below nominal $\alpha = 0.025$. The principal score methods show rejection rates above 4% indicating a sizable Type I error inflation.

Coverage probability: The instrumental variable approach appears to provide close to nominal coverage with deviations from 95% within simulation error. The per protocol estimator appears to provide conservative coverage in excess of the 95% nominal level especially for larger vaccine efficacies. The principal score methods do not appear to control coverage probabilities with simulated coverage significantly below the nominal level. For the scenarios close to the null (see limitations) non-coverage is around 8% points which would approximately correspond to a 4% Type I error inflation for a corresponding 2.5% level one-sided test in this setting, which aligns with results shown for Type I error.

Scenarios A2: The treatment-dependent compliance scenarios (A2) isolated the effect of randomisation-group differences in compliance on estimation of the always-complier estimand.

In this scenario we show results for all investigated vaccine efficacy settings, i.e. $\beta_{A2} \in \{0.0, 0.3, 0.5, 0.7, 0.9\}$.

Again, we omit results for settings with only partially or completely unobserved covariates V and W. We only show scenarios assuming a baseline incidence rate of 1 infection per 1000 subjects per month, due to the finding that this parameter has little impact on performance.

The scenario configurations for scenario group A2 are summarised in Table 4.4, with varying parameters shown in bold and the true principal stratum vaccine efficacy VE_{PS} given in the final column.

Table 4.4: Parameter configurations for scenario group A2.

Scenario	Incidence	β_{A2} **(VE Scale)**	C	δ_1	γ_A	γ_V	γ_W	γ_{AW}	β_V	β_W	β_{AW}	p_V	p_W	VE_{PS}
VE:0.3_gA:-0.36	1000	**0.3**	0.95	T	**-0.36**	0	0	0	0	0	0	0	0	**0.299**
VE:0.3_gA:0	1000	**0.3**	0.95	T	**0**	0	0	0	0	0	0	0	0	**0.299**
VE:0.5_gA:-0.36	1000	**0.5**	0.95	T	**-0.36**	0	0	0	0	0	0	0	0	**0.499**
VE:0.5_gA:0	1000	**0.5**	0.95	T	**0**	0	0	0	0	0	0	0	0	**0.499**
VE:0.7_gA:-0.36	1000	**0.7**	0.95	T	**-0.36**	0	0	0	0	0	0	0	0	**0.699**
VE:0.7_gA:0	1000	**0.7**	0.95	T	**0**	0	0	0	0	0	0	0	0	**0.699**
VE:0.9_gA:-0.36	1000	**0.9**	0.95	T	**-0.36**	0	0	0	0	0	0	0	0	**0.900**
VE:0.9_gA:0	1000	**0.9**	0.95	T	**0**	0	0	0	0	0	0	0	0	**0.900**
VE:0_gA:-0.36	1000	**0**	0.95	T	**-0.36**	0	0	0	0	0	0	0	0	**0.000**
VE:0_gA:0	1000	**0**	0.95	T	**0**	0	0	0	0	0	0	0	0	**0.000**

Treatment related compliance does not have a substantial impact on operating characteristics compared to scenarios with unrelated compliance. For Bias, RMSE, Type I error, and Power only minor differences within the range of simulation error are observed between comparable scenarios with and without treatment related non-compliance.

4.4.2 Scenario category B

Scenario category B assessed prognostic heterogeneity through marker V, which affected baseline infection risk treatment compliance and regimen completion. Marker W was inactive.

All scenarios shown include treatment-dependent compliance ($\gamma_A \neq 0$) and an active marker V, in one of three configurations: V affecting both infection risk and compliance, only infection risk (a prognostic effect alone), or only compliance. Covariate-observation variants are shown for the principal-score methods, while IV and per-protocol results are shown for the fully observed setting; the baseline incidence is fixed at 1 in 1000 per month and the VE values are limited to 0.3, 0.5, and 0.7. The remaining configurations are reported in the full results table provided as a supplement.

The scenario configurations are summarised in Table 4.5, with varying parameters shown in bold and the true principal stratum vaccine efficacy VE_{PS} given in the final column.

Table 4.5: Parameter configurations for scenario group B1.

Scenario	Incidence	β_{A2} **(VE Scale)**	C	δ_1	γ_A	γ_V	γ_W	γ_{AW}	β_V	β_W	β_{AW}	p_V	p_W	VE_{PS}
VE:0.3_gV:0.5_bV:0	1000	**0.3**	0.95	T	-0.36	**0.5**	0	0	**0**	0	0	0.3	0	**0.299**
VE:0.3_gV:0.5_bV:0.41	1000	**0.3**	0.95	T	-0.36	**0.5**	0	0	**0.41**	0	0	0.3	0	**0.299**
VE:0.3_gV:0_bV:0.41	1000	**0.3**	0.95	T	-0.36	**0**	0	0	**0.41**	0	0	0.3	0	**0.299**
VE:0.5_gV:0.5_bV:0	1000	**0.5**	0.95	T	-0.36	**0.5**	0	0	**0**	0	0	0.3	0	**0.499**
VE:0.5_gV:0.5_bV:0.41	1000	**0.5**	0.95	T	-0.36	**0.5**	0	0	**0.41**	0	0	0.3	0	**0.499**
VE:0.5_gV:0_bV:0.41	1000	**0.5**	0.95	T	-0.36	**0**	0	0	**0.41**	0	0	0.3	0	**0.499**
VE:0.7_gV:0.5_bV:0	1000	**0.7**	0.95	T	-0.36	**0.5**	0	0	**0**	0	0	0.3	0	**0.699**
VE:0.7_gV:0.5_bV:0.41	1000	**0.7**	0.95	T	-0.36	**0.5**	0	0	**0.41**	0	0	0.3	0	**0.699**
VE:0.7_gV:0_bV:0.41	1000	**0.7**	0.95	T	-0.36	**0**	0	0	**0.41**	0	0	0.3	0	**0.699**

Bias: For the per-protocol estimator `pp`, the instrumental variable approach `iv`, and the principal score method `ps_cov` that uses covariates both in the score and outcome model patterns are similar to the benchmark scenario A1. The instrumental variable approach has slightly reduced absolute bias, which, however, is slightly positive (overestimation of VE) for settings with VE above 0.7. The per-protocol estimator and principal score method

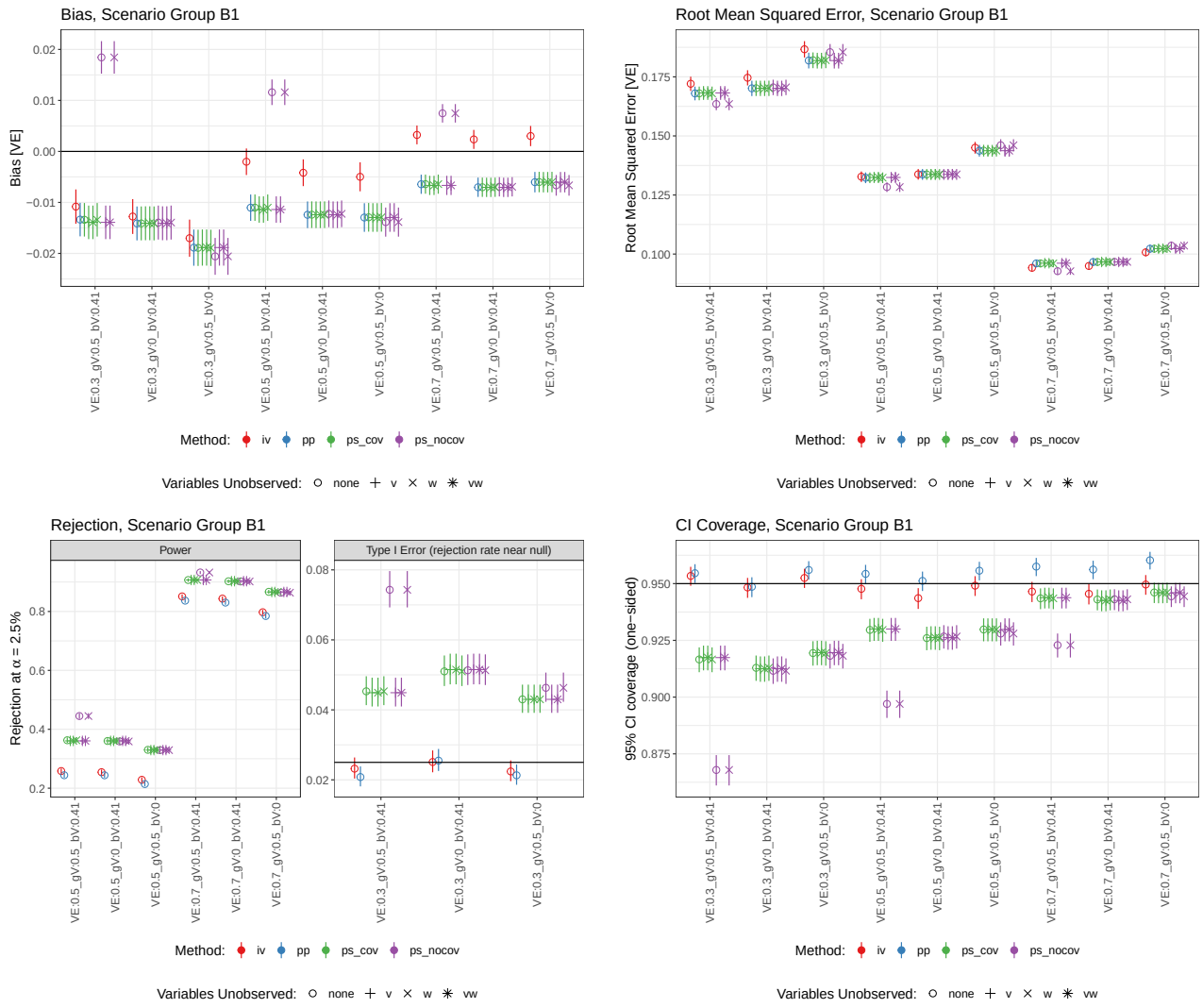


Figure 4.4: Operating characteristics for scenario category B (prognostic heterogeneity through V). Bias (top left), RMSE (top right), rejection probability including power and type I error (bottom left), and 95% CI coverage (bottom right).

perform almost identically. Observation status of covariates (i.e. whether either V, W, or both were available to the model) has no noticeable impact - results differ in the 5th digit behind the comma which is not visible in the figures. Consequently, only settings where both covariates are observed are shown for these methods to keep presentation compact. Overall the extent of bias seems to be slightly reduced compared to scenarios A1 and A2 but still different to zero taking into account simulation error. Numerically the instrumental variable approach appears to reduce the bias best for scenarios where V has an impact on both compliance and infection risk - though this result within simulation error and observed best in the VE 0.3 and VE 0.5 settings.

The PS method that does not include covariates in the outcome performs noticeably worse if either both covariates or only V were observed and V has an impact on both compliance and infection risk. The bias becomes anti-conservative at a comparable absolute scale. Whereas all other methods underestimate vaccine efficacy by about 1% point, `ps_nocov` overestimates it by about 2% points in the setting with vaccine efficacy close to the null of 0.3. This might be expected as in this case the outcome model is misspecified and omits active covariates in the outcome model, while using them in the score model. In scenarios where V only affects compliance bias is numerically increased (within simulation error) which is surprising and warrants further inspection.

RMSE: RMSE follows similar patterns as in the reference scenarios A1 for all methods beside `ps_nocov`. The latter appears to have slightly reduced RMSE in settings where V affects both compliance and infection risk and V is observed, but larger RMSE where V only affects compliance. `iv` has again slightly larger RMSE for scenarios up to VE 0.5 and smaller RMSE above.

Rejection: Again `iv` and `pp` have similar power, close to the planned 80% around VE of 0.7. `ps_cov` has noticably larger power regardless of covariate observation status.`ps_nocov` has similar power to the principal score method with covariates in the outcome model. However, in scenarios where V affects both compliance and infection risk power is noticeably increased, potentially reflecting the positive bias observed in these scenarios. A similar pattern is observed for scenarios close to the null, where `pp` and `iv` reject close to nominal levels. As in the reference scenarios A1, the principal score methods are far from nominal levels, with worst results for `ps_nocov` reaching around 7% if both V affects both compliance and infection risk and is observed.

Coverage: Same pattern as above. `pp` provides conservative coverage in excess of the nominal levels. `iv` has close to nominal coverage with slight undercoverage in some scenarios which might be due to simulation error but follows the pattern observed for bias regarding observation status of V. Principal score methods generally do not provide adequate coverage with worst results for settings with misspecified outcome model (`ps_nocov` when V affects both compliance and infection risk).

4.4.3 Scenario category C

Scenario category C assessed predictive heterogeneity through marker W, which modified vaccine efficacy and compliance under vaccination. Marker V was inactive.

As for the preceding categories, covariate-observation variants are shown only for the principal-score methods, the baseline incidence is fixed at 1 in 1000 per month, and the VE values of 0 and 0.9 are omitted. We also leave out scenarios in which a marker acts only through an interaction term, affecting compliance or infection risk without a corresponding main (prognostic) effect, these are reported in the full results table provided as a supplement. Together these choices keep the presentation compact. The scenario configurations are summarised in Table 4.6, with varying parameters shown in bold and the true principal stratum vaccine efficacy VE_{PS} given in the final column.

Bias and RMSE: Almost identical pattern to the scenario above. For `ps_nocov` the bias is now increased in the negative direction when W is observed, with a correspondingly increased RMSE in these settings. This reflects that an observed W enters the score model, and hence the weights, while being omitted from the outcome model (see corresponding discussion on sensitivity to model specification section 4.4.5). We have not investigated scenarios where W only affects compliance but not infection risk, or vice versa, and only show settings where W affects compliance or treatment effect through the corresponding interaction terms γ_{AW} and β_{AW} with treatment.

Rejection rates: Rejection rates in settings close to the null are now above $\alpha = 0.025$ unless for per-protocol and

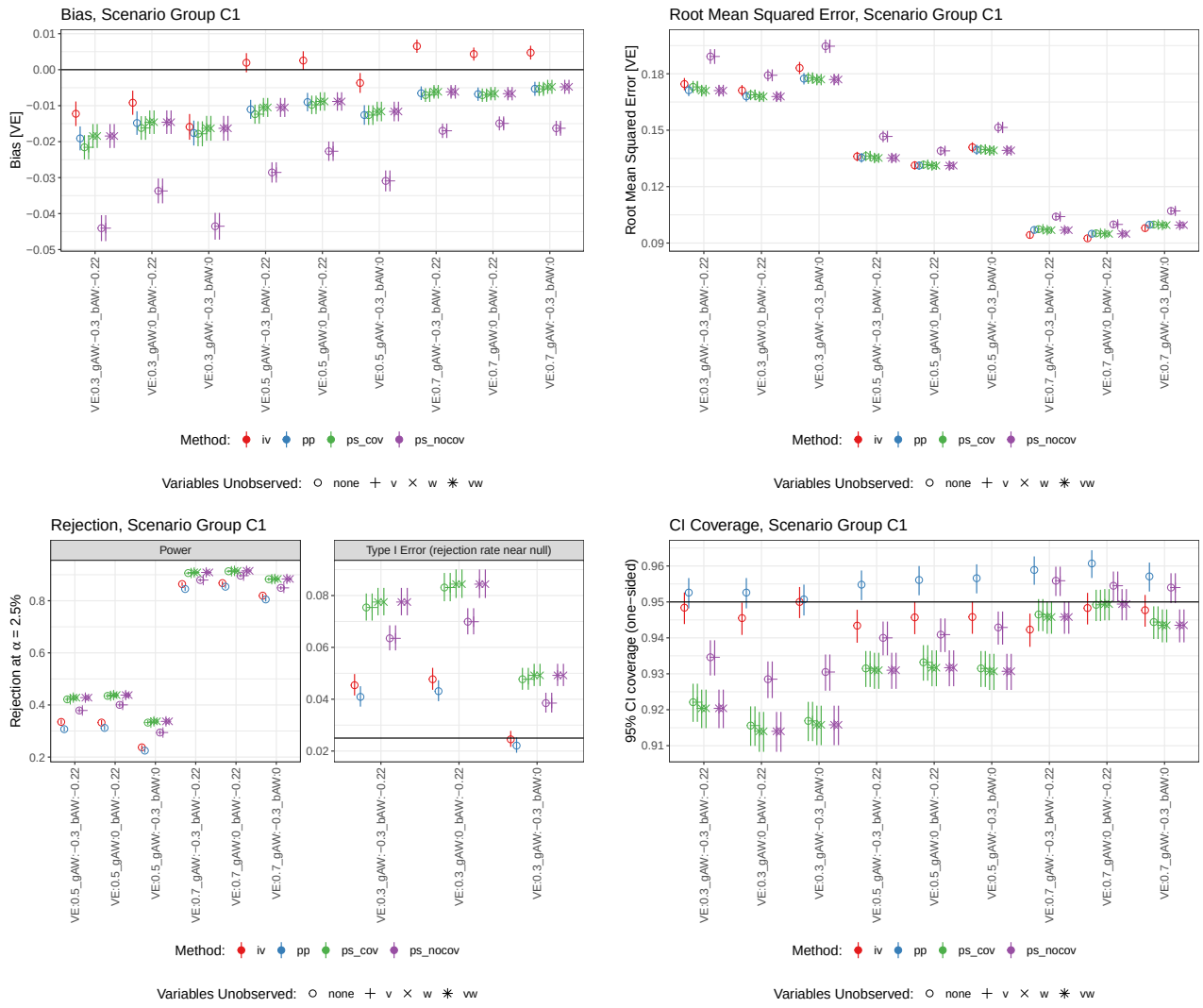


Figure 4.5: Operating characteristics for scenario category C (predictive heterogeneity through W). Bias (top left), RMSE (top right), rejection probability including power and type I error (bottom left), and 95% CI coverage (bottom right).

Table 4.6: Parameter configurations for scenario group C1.

Scenario	Incidence	β_{A2} **(VE Scale)**	C	δ_1	γ_A	γ_V	γ_W	γ_{AW}	β_V	β_W	β_{AW}	p_V	p_W	VE_{PS}
VE:0.3_gAW:-0.3_bAW:-0.22	1000	**0.3**	0.95	T	-0.36	0	-0.8	**-0.3**	0	0.18	**-0.22**	0	0.3	**0.344**
VE:0.3_gAW:-0.3_bAW:0	1000	**0.3**	0.95	T	-0.36	0	-0.8	**-0.3**	0	0.18	**0**	0	0.3	**0.299**
VE:0.3_gAW:0_bAW:-0.22	1000	**0.3**	0.95	T	-0.36	0	-0.8	**0**	0	0.18	**-0.22**	0	0.3	**0.345**
VE:0.5_gAW:-0.3_bAW:-0.22	1000	**0.5**	0.95	T	-0.36	0	-0.8	**-0.3**	0	0.18	**-0.22**	0	0.3	**0.531**
VE:0.5_gAW:-0.3_bAW:0	1000	**0.5**	0.95	T	-0.36	0	-0.8	**-0.3**	0	0.18	**0**	0	0.3	**0.499**
VE:0.5_gAW:0_bAW:-0.22	1000	**0.5**	0.95	T	-0.36	0	-0.8	**0**	0	0.18	**-0.22**	0	0.3	**0.532**
VE:0.7_gAW:-0.3_bAW:-0.22	1000	**0.7**	0.95	T	-0.36	0	-0.8	**-0.3**	0	0.18	**-0.22**	0	0.3	**0.719**
VE:0.7_gAW:-0.3_bAW:0	1000	**0.7**	0.95	T	-0.36	0	-0.8	**-0.3**	0	0.18	**0**	0	0.3	**0.699**
VE:0.7_gAW:0_bAW:-0.22	1000	**0.7**	0.95	T	-0.36	0	-0.8	**0**	0	0.18	**-0.22**	0	0.3	**0.719**

instrumental variable estimators in the setting without treatment effect modification. This is not surprising as treatment effect modification increases the true principal score vaccine efficacy substantially above 0.3.

Coverage: Coverage is therefore the more relevant metric, which shows close to nominal levels for `pp` and `iv` in settings close to the null. For VE around 0.5 confidence intervals for `iv` indicate slight undercoverage. Principal scoring methods, provide coverage substantially below the nominal 95% level (up to 9% non-coverage). If W is observed the `ps_nocov` method has slightly better coverage than the other principal scoring approach, which might reflect the large negative bias (underestimation of VE) observed for these settings.

4.4.4 Scenario category D

Scenario category D combined prognostic heterogeneity through V and predictive heterogeneity through W, representing the most complex main scenario class.

The scenario configurations are summarised in Table 4.7, with varying parameters shown in bold and the true principal stratum vaccine efficacy VE_{PS} given in the final column.

Table 4.7: Parameter configurations for scenario group D1.

Scenario	Incidence	β_{A2} **(VE Scale)**	C	δ_1	γ_A	γ_V	γ_W	γ_{AW}	β_V	β_W	β_{AW}	p_V	p_W	VE_{PS}
VE:0.3_gAW:-0.3_bV:0.41_bAW:-0.22	1000	**0.3**	0.95	T	-0.36	0.5	-0.8	**-0.3**	**0.41**	0.18	**-0.22**	0.3	0.3	**0.344**
VE:0.3_gAW:-0.3_bV:0.41_bAW:0	1000	**0.3**	0.95	T	-0.36	0.5	-0.8	**-0.3**	**0.41**	0.18	**0**	0.3	0.3	**0.299**
VE:0.3_gAW:-0.3_bV:0_bAW:-0.22	1000	**0.3**	0.95	T	-0.36	0.5	-0.8	**-0.3**	**0**	0.18	**-0.22**	0.3	0.3	**0.344**
VE:0.3_gAW:0_bV:0.41_bAW:-0.22	1000	**0.3**	0.95	T	-0.36	0.5	-0.8	**0**	**0.41**	0.18	**-0.22**	0.3	0.3	**0.345**
VE:0.5_gAW:-0.3_bV:0.41_bAW:-0.22	1000	**0.5**	0.95	T	-0.36	0.5	-0.8	**-0.3**	**0.41**	0.18	**-0.22**	0.3	0.3	**0.531**
VE:0.5_gAW:-0.3_bV:0.41_bAW:0	1000	**0.5**	0.95	T	-0.36	0.5	-0.8	**-0.3**	**0.41**	0.18	**0**	0.3	0.3	**0.499**
VE:0.5_gAW:-0.3_bV:0_bAW:-0.22	1000	**0.5**	0.95	T	-0.36	0.5	-0.8	**-0.3**	**0**	0.18	**-0.22**	0.3	0.3	**0.531**
VE:0.5_gAW:0_bV:0.41_bAW:-0.22	1000	**0.5**	0.95	T	-0.36	0.5	-0.8	**0**	**0.41**	0.18	**-0.22**	0.3	0.3	**0.532**
VE:0.7_gAW:-0.3_bV:0.41_bAW:-0.22	1000	**0.7**	0.95	T	-0.36	0.5	-0.8	**-0.3**	**0.41**	0.18	**-0.22**	0.3	0.3	**0.719**
VE:0.7_gAW:-0.3_bV:0.41_bAW:0	1000	**0.7**	0.95	T	-0.36	0.5	-0.8	**-0.3**	**0.41**	0.18	**0**	0.3	0.3	**0.699**
VE:0.7_gAW:-0.3_bV:0_bAW:-0.22	1000	**0.7**	0.95	T	-0.36	0.5	-0.8	**-0.3**	**0**	0.18	**-0.22**	0.3	0.3	**0.719**
VE:0.7_gAW:0_bV:0.41_bAW:-0.22	1000	**0.7**	0.95	T	-0.36	0.5	-0.8	**0**	**0.41**	0.18	**-0.22**	0.3	0.3	**0.719**

Results for scenario category D are presented for baseline incidence of 1 in 1000 per month and vaccine efficacy values of 0.3, 0.5, and 0.7. Covariate observation variants are shown for principal score methods only; IV and per-protocol results are shown for the fully observed covariate setting. Scenarios where an interaction effect on compliance or infection risk is present without the corresponding main effect are excluded, as are scenarios without treatment-dependent compliance. The full numerical results across all parameter configurations are provided in the supplementary table. The scenario configurations are summarised in Table 4.7, with varying parameters shown in bold and the true principal stratum vaccine efficacy VE_{PS} given in the final column.

Bias: The per-protocol and instrumental variable estimators follow the same pattern as in the simpler scenarios, with `iv` showing the smallest bias (close to zero across all settings). When the markers are observed, `ps_cov` performs comparably to per protocol estimator. `ps_nocov` is far more variable, and its bias is governed by three things: the compliance interaction γ_{AW}, the effect modification β_{AW}, and which markers are observed. With both markers observed, bias is close to zero only when there is no compliance interaction ($\gamma_{AW} = 0$). A compliance interaction ($\gamma_{AW} = -0.3$) introduces a negative bias that is further exacerbated when the same

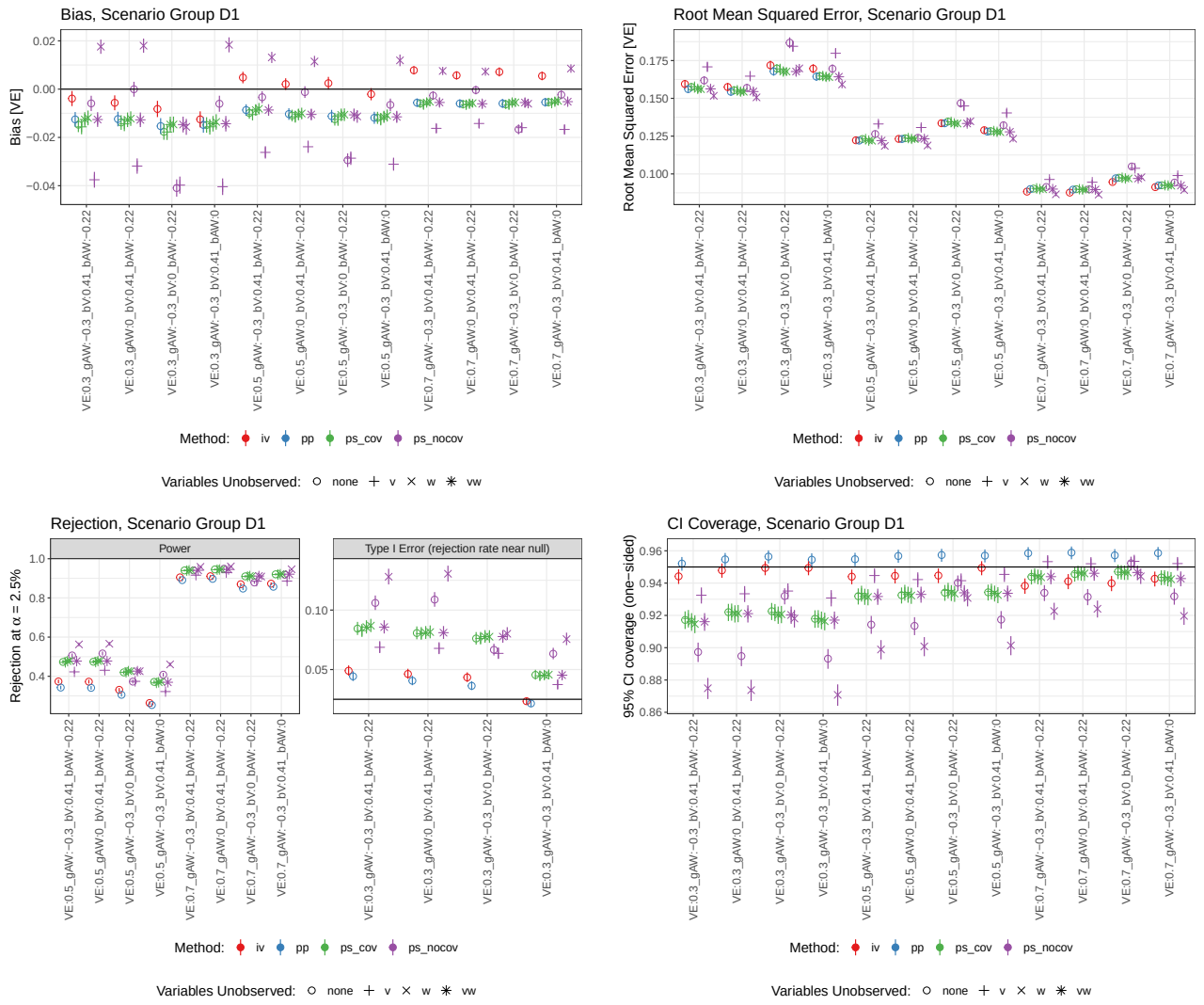


Figure 4.6: Operating characteristics for scenario category D (combined prognostic and predictive heterogeneity through V and W). Bias (top left), RMSE (top right), rejection probability including power and type I error (bottom left), and 95% CI coverage (bottom right).

marker also modifies infection risk ($\beta_{AW} \neq 0$). Leaving a single marker unobserved worsens bias in opposite directions: overestimation of VE when W is unobserved, and underestimation when V is unobserved. With neither marker observed `ps_nocov` reduces to the same model as `ps_cov` and matches its (and the per-protocol) performance.

These differences are not driven by the score model, which is fit in treated participants only and is correctly specified across all settings; they arise because `ps_nocov` uses the observed markers in the score model but omits them from the outcome model, so the two models carry different explanatory variables (unlike `ps_cov`, which uses the same set in both). `ps_nocov` is therefore highly sensitive to outcome-model specification: its bias is removed only when all active markers are also included in the outcome model, at which point performance is no better than per protocol. The precise drivers of the residual pattern are not fully resolved and would require dedicated scenarios isolating each form of misspecification.

RMSE: Same pattern as above and in other scenarios. Positive bias increases RMSE, negative decreases it compared to other methods.

Rejection: Again power advantage for principal score methods according to bias direction. Rejection rates in excess of nominal levels for all methods if treatment effect modification is part of the data generating process. Rates for principal scoring approaches, however, far from nominal levels (up to 15%). Reported "Type I Error" results are difficult to interpret as the true principal-stratum VE exceeds 0.3, so these are not genuine null settings. Coverage therefore the more relevant metric.

Coverage: Close to nominal coverage for `iv` with a tendency for slight undercoverage. `pp` provides coverage close to but in excess of nominal level. Principal scoring with at times substantial undercoverage. For settings where large negative bias was observed coverage is closer. Note that since one-sided coverage was evaluated this needs to be interpreted with caution. Interestingly scenarios where `ps_nocov` showed negligible bias coverage is substantially below nominal levels and worse than for related methods with larger bias.

4.4.5 Additional complex and stress-test scenarios

The additional stress-test scenarios were designed to evaluate estimator robustness under conditions that deviate from the assumptions underlying the main scenario categories. Starting from a baseline configuration (*Base*) corresponding to the full category D setting: treatment-reducing compliance ($\gamma_A < 0$), negative compliance effects of both V and W on completion, prognostic effects of both markers on infection risk, a positive effect modification of vaccine efficacy by W, and a protective dose-one effect.

Each stress-test scenario modifies a single aspect of the data-generating mechanism. Specifically, the scenarios consider reduced overall compliance (*OC50*), reversed direction of the compliance effect of the predictive marker W (*CW+*), reversed direction of the compliance effect of the prognostic marker V (*CV-*), treatment-increasing rather than treatment-reducing compliance (*CA+*), reversed direction of the compliance interaction between treatment and W (*CAW+*), reversed direction of the prognostic effect of V on infection risk (*PV+*), reversed direction of the prognostic effect of W on infection risk (*PW+*), and absence of a dose-one effect (*D1F*). The scenario *CA+* is of particular interest as it violates the monotonicity assumption underlying the instrumental variable estimator, since treatment assignment increases rather than decreases the probability of non-completion. The remaining scenarios represent perturbations that challenge the principal score and per-protocol estimators through unexpected directions of covariate effects on compliance and infection risk.

The scenario configurations are summarised in Table 4.8, with varying parameters shown in bold. For extra scenarios we only highlight individual cells where the specific scenario deviates from the base scenario and the true principal stratum vaccine efficacy VE_{PS} given in the final column.

Bias: Overall bias remains moderate around absolute levels of 1% points and generally below 2.5% points. Notable exceptions are scenario `OC50` where the instrumental variable approach appears to break down introducing extreme positive bias (overestimation of VE) of around 15% points. Interestingly, in the scenario where receiving only one dose of vaccine does not confer any protective effect results in a moderate negative bias ($\sim -2.5\%$ points) in estimates based on the instrumental variable approach. This is notable as in all other settings in this group, `iv` tends to slightly overestimate vaccine efficacy, which is in line with results from D1. Finally, scenario

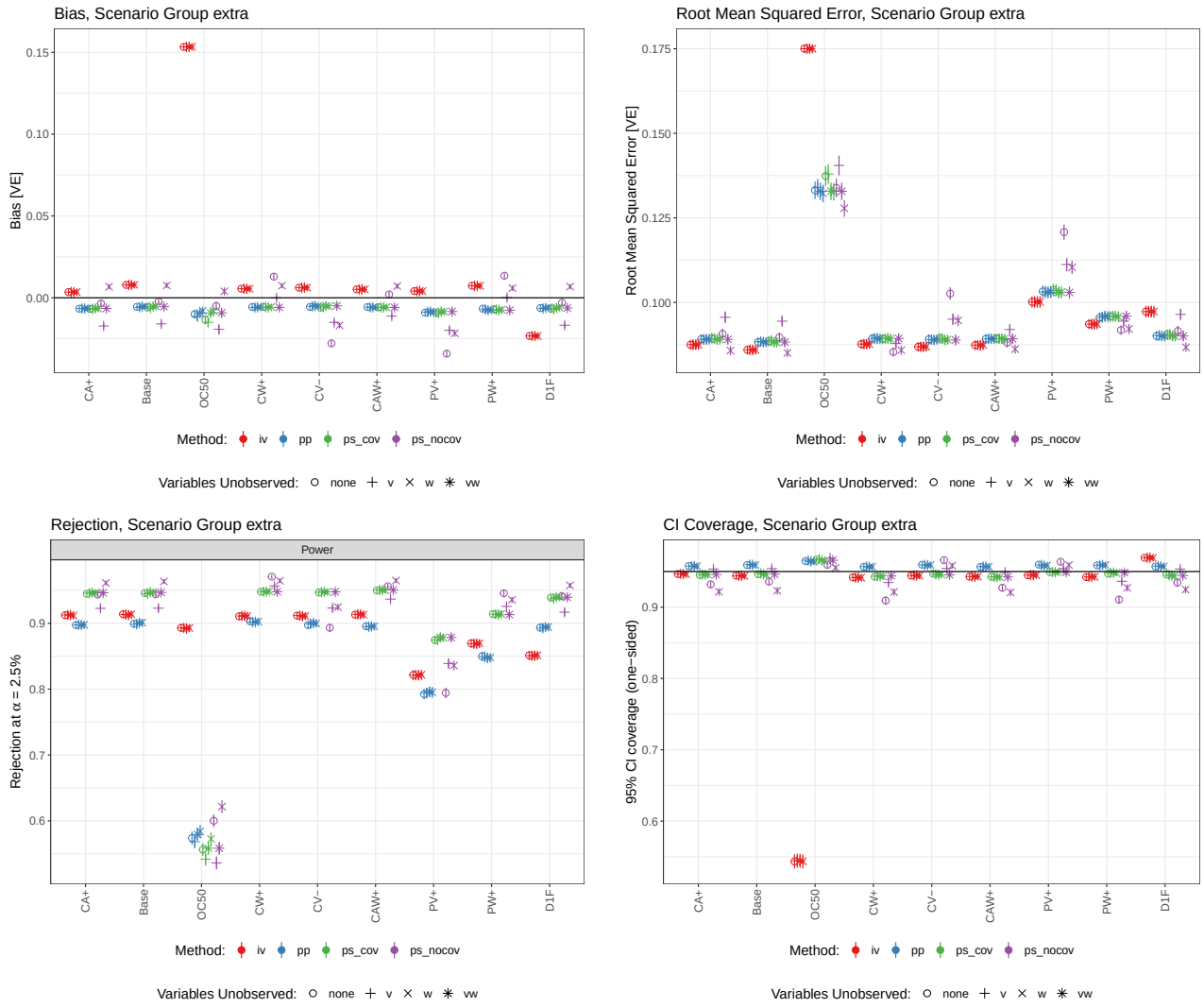


Figure 4.7: Operating characteristics for the additional stress-test scenarios. Bias (top left), RMSE (top right), rejection probability including power and type I error (bottom left), and 95% CI coverage (bottom right).

Table 4.8: Parameter configurations for scenario group extra.

Scenario	Incidence	β_{A2} (VE Scale)	**C**	δ_1	γ_A	γ_V	γ_W	γ_{AW}	β_V	β_W	β_{AW}	p_V	p_W	VE_{PS}
Base	1000	0.7	0.95	T	-0.36	0.5	-0.8	-0.3	0.41	0.18	-0.22	0.3	0.3	0.719
CA+	1000	0.7	0.95	T	**0.36**	0.5	-0.8	-0.3	0.41	0.18	-0.22	0.3	0.3	0.719
CAW+	1000	0.7	0.95	T	-0.36	0.5	-0.8	**0.3**	0.41	0.18	-0.22	0.3	0.3	0.719
CV-	1000	0.7	0.95	T	-0.36	**-0.5**	-0.8	-0.3	0.41	0.18	-0.22	0.3	0.3	**0.718**
CW+	1000	0.7	0.95	T	-0.36	0.5	**0.8**	-0.3	0.41	0.18	-0.22	0.3	0.3	**0.720**
D1F	1000	0.7	0.95	**F**	-0.36	0.5	-0.8	-0.3	0.41	0.18	-0.22	0.3	0.3	0.719
OC50	1000	0.7	**0.5**	T	-0.36	0.5	-0.8	-0.3	0.41	0.18	-0.22	0.3	0.3	**0.712**
PV+	1000	0.7	0.95	T	-0.36	0.5	-0.8	-0.3	**-0.41**	0.18	-0.22	0.3	0.3	0.719
PW+	1000	0.7	0.95	T	-0.36	0.5	-0.8	-0.3	0.41	**-0.18**	-0.22	0.3	0.3	**0.714**

`CA+` does not appear to affect performance even though identification assumptions (no Vaccine compliers) are violated.

RMSE: Increasing non-compliance to extreme levels of 50% increases RMSE substantially for all methods and most for the instrumental variable approach. Scenarios `PV+`, `PW+`, and `D1F` also appear to increase RMSE compared to the base scenario, if to a lesser extent.

Rejection: Power is around or above 90% for most scenarios in line with D1 (target is 80% not taking into account effect modification). Extreme non-compliance `OC50` to about 60% for most methods beside the instrumental variable approach, reflecting the large positive bias observed for this method. Further, power in scenarios `PV+`, `PW+`, and `D1F` appears noticeably lower compared to the base scenario from D1.

Coverage: Except for the setting with extreme non-compliance coverage is within the range observed in scenario groups above. In scenario `OC50` coverage of the instrumental variable approach becomes extremely liberal providing levels of only about 50% compared to the nominal 95%, whereas the other methods appear to provide coverage larger than the nominal level.

Limitations of Vaccine Trial Simulations Standard error estimation: A notable limitation of the present implementation concerns the variance estimation for the IV and principal-score methods. As described in subsection 4.2, both methods use `emmeans` for inference on the second-stage / outcome model, which by default extracts the model-based variance–covariance matrix of the fitted model rather than a robust (sandwich) estimator (as originally assumed during development). For the principal score and instrumental variable approaches this means that uncertainty arising from the first-stage or principal score estimation is not propagated into the standard error of principal stratum estimates. The protocol intended robust variance–covariance estimation with delta-method inference.

Re-running the simulation with corrected variance estimators was not feasible within the timeframe of the report. As a consequence, empirical confidence interval coverage and rejection probabilities for these two methods should be interpreted with this caveat in mind. A limited test comparing model based with sandwich estimators across scenarios in category A2 (based on 2000 replications), indicates that robust estimates have little impact on the performance (esp. coverage) for the instrumental variable approaches, while principal score methods saw substantial improvement bringing simulated coverage closer to nominal coverage. However, it appears that power advantages observed for principal score methods using model based standard errors may no longer apply to the same extent, either. Preliminary results, however, have the principal score methods still slightly below desired levels in some scenarios and more liberal than all other methods in respective settings. Bias is not influenced by changing the variance estimate. Power and Type I error rate are driven by both bias and the variance estimate. The estimates of power and Type I error rate for the effected methods are to be interpreted with caution, but the impact of the variance estimation is not easily separated from the impact of bias.

Our simulation results nevertheless show, the IV procedure performs broadly well in terms of bias, rejection and coverage related operating characteristics across the scenarios considered. The principal-score procedure, by contrast, shows substantial deviations in the majority of investigated settings. We tentatively attribute part of this behaviour to the variance estimation issue described above; a definitive assessment would require re-running

the affected analyses with appropriate variance estimators (e.g. sandwich or bootstrap).

Finite-sample bias on the VE scale: The modest negative bias common to all four methods (see e.g. Benchmark Scenarios subsubsection 4.4.1) can be explained by a finite-sample bias of the transformation $\mathrm{VE} = 1 - \exp(\hat{\beta})$ (as all models estimate a quantity on the log scale). Because the back-transform is convex, an asymptotically unbiased log-scale estimator maps to an upward-biased ratio and hence a downward-biased VE, with $\mathbb{E}[\widehat{\mathrm{VE}}] - \mathrm{VE} \approx -(1-\mathrm{VE})\, v/2$, where v is the variance of the log-effect estimator and is governed primarily by the number of events (Greenland, 2000; Lyles et al., 2012). This accounts for the bias being negative, largest near the null and decreasing in magnitude as VE increases (through the $(1 - \mathrm{VE})$ factor), and largely insensitive to baseline incidence, since sample sizes were calibrated to approximately constant power with approximately constant event counts. The bias could be removed on the reporting scale by a mean-bias correction (Lyles et al., 2012) we did not apply such a correction here. Due to the low baseline incidence, the bias is of similar magnitude for the different outcome-model link functions (logistic for the principal score estimators, and Poisson for the per-protocol and instrumental variable estimators).

However, this implies that the slightly smaller absolute bias of the instrumental variable approach could reflect a fortuitous cancellation of biases acting in opposite directions rather than improved estimation. Similarly, observed improvements of bias in other scenarios need to be interpreted with caution. Finally, non-collapsibility of models in combination with covariate adjustment, weights- and two-stage estimation could introduce additional bias with respect to the marginal estimand.

Model specification: On inspection of results we noticed that the outcome, score, and instrument model specifications included only main effects of covariates, not interaction terms. The principal-score model is fit in treated participants only, where the treatment-by-W interaction γ_{AW} is absorbed into the W main effect; it is therefore correctly specified even under treatment-dependent compliance interactions. The outcome models and the IV second stage, however, include treatment alongside the markers and omit the treatment-by-marker interaction β_{AW}, so they are misspecified in settings with effect modification (the b_{AW}=-0.22 scenarios in C1 and D1). This is reflected in the performance differences between principal-score methods in these scenarios, supporting our conclusion that these methods are particularly sensitive to outcome-model misspecification. Considering the subpar performance of principal-score methods even where the models were correctly specified (e.g. B1), our conclusion that these methods should be treated with caution remains supported.

Confidence intervals: In contrast to other Scenarios, simulations in the preventive vaccine efficacy setting returned one-sided 95% and 97.5% confidence intervals due to an oversight when coding the one-sided super-superiority tests. Above we present results for the 95% confidence intervals. Consequently, coverage needs to be interpreted with caution. While reported figures are faithful with respect to control statements of one-sided coverage and importantly Type I error control (see also below), the same conclusions may not fully apply to coverage of associated two-sided confidence intervals especially in settings with negatively biased estimates.

Type I error control: We considered the null hypothesis of at least 30% principal stratum vaccine efficacy. However, scenarios intended to represent settings under the null hypothesis (with hypothetical full-dose efficacy $\beta_{A2} = 0.3$) resulted in true principal stratum vaccine efficacies above or below the target value of 30%. Consequently, interpretation of rejection rates in terms of Type I error in respective scenarios is not entirely appropriate and at best indicative. Scenarios with beneficial treatment effect modification removes principal stratum vaccine efficacy even further away from null levels. To this end, evaluation of one-sided confidence interval coverage is considered more appropriate as it corresponds to the Type I error rate of a super-superiority test under a null hypotheses corresponding to the true principal stratum vaccine efficacy in each scenario.

4.5 Case study

As a case study a scenario where prognostic heterogeneity through V and predictive heterogeneity through W were present was chosen (First scenario from scenario category D). The dataset was simulated under the alternative with a VE of 70% after the second dose and a VE in the principal stratum of 71.8%. The simulated study included 7639 participants in each arm. The observed number of infection events was 54, 44 of which occured in the control arm and 10 in the treatment arm. There were 320 non-compliers in the control arm and 552 non-compliers in the treatment arm. (See Table 4.9 for descriptive statistics.)

The analysis methods described in subsection 4.2 were applied to the simulated dataset. All calculations were done by calculating auxiliary variables and application of standard linear and generalized linear regression functions available in R. The `emmeans` package was then used to derive marginal estimates from the models and to test the super-superiority null hypothesis of a VE of less than 30%. Appendix C includes exemplary R code to fit the according analysis models.

For the principal score weighting, weights were estimated by fitting a logistic regression model with compliance as the dependent variable and V and W as independent variables in only the treated participants and then estimating the weights used in the outcomes model according to the formula given in subsection 4.2.

For the IV regression, T was calculated as $T_i = \mathbb{1}(A_i = 1, C_i = 1)$. The first stage was a linear regression of T on the randomised allocation A (the instrument) and the covariates V and W, yielding fitted values $\widehat{T}$. The second stage was a Poisson regression of the infection indicator on $\widehat{T}$ together with V and W. The estimate based on the per-protocol set was calcualted as a Poisson regression of the occurrence of an infection on V and W in only the compliant participants.

The results of the individual analyses are shown in Table 4.10. All methods over-estimated the true principal stratum VE of 71.8% but all CIs covered the true value. All methods clearly rejected the null-hypothesis. See Table 4.10 for the individual estimates.

Table 4.9: Case study: Descriptive statistics by treatment group. Values are absolute (relative) frequencies.

Variable	Control (n=7639)	Treatment (n=7639)
V	2271 (29.7%)	2264 (29.6%)
W	2273 (29.8%)	2314 (30.3%)
Non-Compliance	320 (4.2%)	552 (7.2%)
Infection	44 (0.576%)	10 (0.131%)

Table 4.10: Case study: Estimated vaccine efficace VE, one-sided 95% lower confidence limit and p-value for the super-superiority null hypothesis.

Method	VE	lower CI	p
Principal score weighting with covariates in outcomes model	77.9%	61.7%	0.00029
Principal score weighting without covariates in outcomes model	78.1%	62.0%	0.00026
Instrumental variable regression	79.8%	62.3%	0.00051
Per-protocol analysis	77.8%	59.4%	0.00087

4.6 Discussion

The per-protocol estimator provides robust operational performance across all investigated scenarios. Different specifications of the data-generating model with respect to the causal structure appears to have little to no impact on bias, RMSE, power, and coverage. It provides moderate negative bias, to the extent of $\sim 1\%$ point underestimation of true principal stratum vaccine efficacy in settings around 30% VE improving with higher VE levels. RMSE is broadly in line with comparator methods. Power is close to targeted values. Importantly, one-sided coverage probabilities provide close to but at least nominal coverage across all investigated scenarios (including Scenario Group extra), indicating also good performance in terms of Type I error control.

Performance of the instrumental variable approach is broadly robust across scenarios, with notable deviations only in settings with extreme levels of non-compliance (`OC50`) and interestingly if single-dose hypothetical efficacy is 0. The instrumental variable approach achieves modest bias reduction compared to the per-protocol estimator with a tendency for slight overestimation if vaccine efficacy is larger than 50%. It provides moderately better power than per protocol estimation with coverage closer to nominal levels with slight undercoverage in scenarios with larger vaccine efficacy.

Performance of principal score methods using observed covariates in both the score and outcome models shows a consistent performance profile across scenarios. Bias is close to the bias observed for the per protocol estimator, with little impact of covariate observation status or data generating model. Similar results are observed for RMSE. Power is substantially larger than with per protocol or instrumental variable methods to the extent of up to around 10% points in some settings with a target power of 80%. Coverage probabilities, however, are substantially below nominal levels, indicating severe Type I error inflation and unreliable interval estimates. Preliminary results indicate that using robust estimates of the standard error may alleviate these issues to some extent. However, it appears that power advantages observed for principal score methods using model based standard errors may no longer apply to the same extent.

Principal score methods that use covariates only in the score model show the largest heterogeneity in terms of performance across scenarios. In cases where active covariates are observed and therefore only accounted in the score model bias, coverage and in consequence Type I error are severely affected. In scenarios where only V or W are active issues observed with other principal score approaches are consistently exacerbated and the extent depends on whether one or both models are misspecified and use different explanatory variables due to observation status. Some improvements in terms of bias are observed in scenario group D1. The exact pattern of misspecification that drives these deviations is not entirely clear and would require further investigation designing specific scenarios. In general, however, this method introduces the largest bias and provides the worst control of one-sided coverage probability (and hence Type I error) across the majority of investigated scenarios, with only a few settings where bias is effectively removed and even fewer where bias reduction is noticeably better compared to the instrumental variable approach. Further, improvements in bias are only observed where all covariates are active and observed. Considering that in practice the true data generation model is unknown, potential improvements cannot be guaranteed without strong untestable assumptions while deviations can result in substantial exacerbation of bias.

Finally, the simulation results need to be interpreted within the scope of the data-generating mechanism and estimand considered here. Across the scenarios studied, bias on the VE scale remained modest for all methods (largest absolute values around 3%), with the per-protocol and adjusted principal-score (`ps_cov`) estimators performing near-identically and the instrumental variable estimator showing the smallest absolute bias once VE exceeded zero. These patterns, however, were observed under high compliance ($\approx 95\%$) and the specific marker and effect-modification structures imposed; they need not carry over to other settings, as illustrated by the breakdown of the instrumental variable estimator under heavy non-compliance (OC50). As for several of the methods in this study, more than one defensible implementation exists without clear guidance on the preferred choice. The principal-score weights used here follow the odds form of Jo and Stuart (2009); whether an alternative such as ratio weighting would be more efficient remains open. Likewise, fully doubly-robust estimators that combine weighting with an outcome regression, of which the adjusted variant (`ps_cov`) is a partial form, could offer further robustness or efficiency. Whether they would improve on the per-protocol or instrumental variable approaches warrants further investigation.

5 Scenario class 4: Chronic disease with safety endpoint, while on treatment strategy

In the fourth scenario we consider trials in chronic diseases where a safety endpoint needs to be taken into account (Unkel et al., 2019). An estimand with a while-on-treatment strategy is considered to assess the safety impact of the medication. A relevant example is the study by (Singh et al., 2006), where erythropoietin was studied as therapy for anaemia and major cardiac adverse event (MACE) was a relevant safety endpoint.

In the simulation we assume that the effect of the medication persists for a certain time after discontinuation. In the estimation of the while-on-treatment estimand we include a buffer time window after discontinuation in which adverse events are still counted.

We further assume that some patients may discontinue the study immediately after treatment discontinuation, such that they are not monitored in the buffer window and potential adverse events in the buffer window may be obscured. We use inverse probability of censoring weighting to account for study withdrawals that preclude the observation of events in the buffer window.

The aim of the analysis in clinical terms, is to assess whether the risk for an adverse event is decreased while the patient is under the effect of the experimental treatment compared to the control treatment.

Estimand
The considered estimand has the following relevant attributes:

- Endpoint: Time till the defined adverse event (e.g. MACE)
- Treatments: Experimental treatment versus control, administered in regular intervals
- Summary measure: Hazard ratio
- Population: Anaemic patients eligible for the considered treatments
- Intercurrent event strategy: While-on-treatment strategy to account for treatment discontinuation, estimating the effect of treatment on the endpoint while treatment is received plus a predefined buffer time window.

Assumptions
The IPW approach requires a correctly specified propensity model with no unobserved confounders. We will study violations of this assumption by including an unobserved covariate with increasing effect on the propensity to leave the study after treatment discontinuation.

The comparator Cox model without weights assumes that premature study discontinuation is independent of the time to adverse event. The simulation will include scenarios where this assumption is met as well as scenarios where it is violated.

For all analysis approaches, the applied buffer window length may impact the analysis result. We will explore deviations of the analysis-specific buffer length from the true time window of pertained effect in both directions, including scenarios with too short, correctly specified and too long analysis windows.

Missing data handling: Data that is missing in the buffer window, which follows after treatment discontinuation, will be accounted for by inverse probability weighting of observed patients.

5.1 Data generation

As general study design we assume a 1:1 randomised trial to compare an experimental treatment for anaemia to a control treatment. We assume that this study includes the time to occurrence of a major cardiac adverse events (MACE) as a safety endpoint, see e.g. (Singh et al., 2006).

We denote, for individual i, $T_{mace,i}$, the time to the MACE event, potentially censored, and $e_{mace,i}$, the indicator function of a MACE event occurring at $T_{mace,i}$. In the simulation, the time to discontinuation $T_{disc,i}$ is also

considered, potentially censored, with the indicator function $e_{disc,i}$. After discontinuation, a patient has a probability $P(withdraw_i)$ to withdraw from the study

The assigned treatment is noted $A_i \in \{0, 1\}$, with 0 representing the reference treatment and 1 representing the experimental treatment, sampled from a Bernoulli distribution, with probability 0.5. As outlined per the protocol, each individual has observed covariates X_i and Z_i and a confounding covariate L_i, all sampled from independent standard normal distributions.

Given the causal dependencies between the variables, the covariates and treatment allocation are first sampled, then used to sample the time to discontinuation and the probability to withdraw, and eventually all variables are used to sample the time to MACE.

First, the time to discontinuation is simulated according to a constant baseline hazard model, characterizing $\eta_{disc,i}$, for individual i:

$$\begin{aligned}\eta_{disc,i} &= \exp(\beta_{disc,0} + (0.5 - A_i) \times \beta_{disc,trt} + X_i \times \beta_{disc,X} + Z_i \times \beta_{disc,Z} + L_i \times \beta_{disc,L}) \\ \beta_{tot,disc,i} &= \beta_{disc,0} + (0.5 - A_i) \times \beta_{disc,trt} + X_i \times \beta_{disc,X} + Z_i \times \beta_{disc,Z} + L_i \times \beta_{disc,L}\end{aligned}$$

with $\beta_{disc,0}$ the baseline hazard of discontinuation, of approximately 50% after one year, $\beta_{disc,trt}$ the treatment effect on discontinuation, $\beta_{disc,X}, \beta_{disc,Z}, \beta_{disc,L}$ the covariate effects on the discontinuation, and $\beta_{tot,disc,i}$ the total discontinuation hazard.

To simulate a time to discontinuation $T_{disc,i}$, we sample u_i from a uniform distribution and report the following:

$$T_{disc,i} = \frac{-\log(u_i)}{\exp(\beta_{tot,disc})}$$

Next, we simulate the time to a MACE event according to a constant baseline hazard model:

$$\begin{aligned}\eta_{mace,i} &= \exp(\beta_{mace,0} + (0.5 - A_i) \times \beta_{mace,trt,tv}(t) + X_i \times \beta_{mace,X} + Z_i \times \beta_{mace,Z} + L_i \times \beta_{mace,L}) \\ \beta_{fixed,mace,i} &= \beta_{mace,0} + X_i \times \beta_{mace,X} + Z_i \times \beta_{mace,Z} + L_i \times \beta_{mace,L} \\ \beta_{tot,trt,mace,i}(t) &= (0.5 - A_i) \times \beta_{mace,trt,tv}(t) \\ \beta_{tot,mace,i}(t) &= \beta_{fixed,mace,i} + \beta_{tot,trt,mace,i}(t)\end{aligned}$$

with $\beta_{mace,0}$ the baseline hazard of MACE event, at approximately 10% after one year, $\beta_{mace,trt,tv}(t)$ the time-varying treatment effect, $\beta_{mace,X}, \beta_{mace,Z}, \beta_{mace,L}$ the covariate effects on the MACE events, and $\beta_{tot,mace,i}$ the total MACE hazard.

As outlined by the protocol, the treatment effect changes depending on the discontinuation status. In fact, we define:

$$\begin{aligned}\beta_{tot,mace,i}(t) &= \beta_{fixed,mace,i} + (0.5 - A_i) \times \beta_{mace,trt} && = \beta_{tot,mace,wot,i} && \text{if } t \leq T_{disc,i} \\ &= \beta_{fixed,mace,i} + (0.5 - A_i) \times \beta_{mace,trt,buffer} && = \beta_{tot,mace,buffer,i} && \text{if } T_{disc,i} < t \leq T_{disc,i} + buffer \\ &= \beta_{fixed,mace,i} + (0.5 - A_i) \times \beta_{mace,trt,after} && = \beta_{tot,mace,after,i} && \text{if } t > T_{disc,i} + buffer\end{aligned}$$

with $\beta_{tot,mace,wot,i}$ the total hazard of MACE while on treatment, $\beta_{tot,mace,buffer,i}$ the total hazard of MACE after discontinuation and within the buffer period $buffer$, and $\beta_{tot,mace,after,i}$ the the total hazard of MACE after discontinuation and the buffer window.

To simulate a time to MACE $T_{mace,i}$ with a time-varying treatment effect, we sample u'_i from a uniform distribution and report the following, by noting $t_{0,i} = T_{disc,i}$, $t_{1,i} = T_{disc,i} + buffer$, $h_{0,i} = \exp(\beta_{tot,mace,wot,i})$, $h_{1,i} = \exp(\beta_{tot,mace,buffer,i})$, $h_{2,i} = \exp(\beta_{tot,mace,after,i})$, and $E_i = -\log(u'_i)$:

$$
\begin{aligned}
T_{mace,i} &= \frac{E_i}{h_{0,i}} && \text{if } E_i \leq t_0 h_{0,i} \\
&= t_0 + \frac{E_i - t_0 h_{0,i}}{h_{1,i}} && \text{if } t_0 h_{0,i} < E_i \leq t_0 h_{0,i} + (t_1 - t_0) h_{1,i} \\
&= t_1 + \frac{E_i - t_0 h_{0,i} - (t_1 - t_0) h_{1,i}}{h_{2,i}} && \text{if } E_i > t_0 h_{0,i} + (t_1 - t_0) h_{1,i}
\end{aligned}
$$

Finally, we simulate the withdrawal status after discontinuation, from a logistic model:

$$logit(P(withdraw_i)) = \beta_{wd,0} + X_i \times \beta_{wd,X} + Z_i \times \beta_{wd,Z} + L_i \times \beta_{wd,L}$$

with $\beta_{wd,0}$ the baseline probability of withdrawal, at approximately 50% and $\beta_{wd,X}, \beta_{wd,Z}, \beta_{wd,L}$ the covariate effects on the withdrawal probability.

Censoring

After data generation, if the time to MACE is shorted than the time to treatment discontinuation, there is no censoring. However, MACE events are censored if they happen after a discontinuation event and if the patient withdrew from the study. Additionally, all events are censored administratively after one year.

Sample size

The sample size is chosen to provide approximately 80% power under the alternative. The number of events e_{mace} is determined by Schoenfeld's formula (Schoenfeld, 1981) as:

$$e_{mace} = 4 \times \left(\frac{z_{1-\alpha/2} + z_{1-\beta}}{\log(HRassumed)} \right)^2$$

with z_γ is the γ quantile of the standard normal distribution. For all simulations, the nominal two-sided significance level is $\alpha = 0.05$, the aimed for power is 1 - $\beta = 0.8$. For scenarios under the alternative hypothesis, the log-hazard ratio assumed for planning log(HRassumed) is chosen to be equal to the simulated treatment effect $\beta_{mace,trt}$, such that the actual power is in the range of 80%. For scenarios under the null hypothesis, log(HRassumed) = log(1.5) and log(HRassumed) = log(2) is considered such that a scenario with large sample size and a scenario with smaller sample size are respectively included.

The number of subjects to be included is calculated as:

$$n = \frac{e_{mace}}{1 - \exp(-\exp(\beta_{mace,0}))}$$

where the denominator corresponds approximately to the marginal proportion of patients with an event within one year.

To summarise the data generating mechanisms, Figure 5.1 visualises a simplified version of the assumed structure and dependencies among the variables.

5.1.1 True treatment effect

The true treatment effect is defined in terms of the expected value of the estimate of a Cox model with covariates, in a setting without study withdrawals. We use a Cox model of the time to MACE event, including treatment allocation and observed covariates at baseline (X and Z). This value is determined by calculating the estimate from an independently simulated data set with 1 000 000 patients.

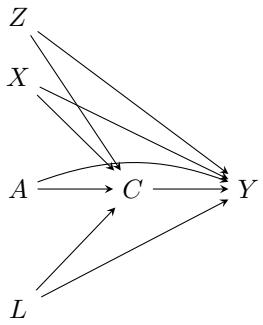


Figure 5.1: Directed acyclic graph (DAG) illustrating the assumed causal structure linking the randomly assigned treatment (A), measured covariates (X, Z), unmeasured covariates (L), the indicator for censoring due to withdrawal (C) and the outcome (Y).

5.1.2 Simulation scenarios

The simulation scenarios are defined as a combination of the parameters values used to simulate the data, as outlined in Table 5.1. The simulation scenarios are based on the core parameters defined in the table, and by changing one parameter value at a time, with the exception of some parameters which are changed together, such as parameters related to the covariate effects of X and Z, the while-one-treatment effect $\beta_{mace,trt}$ and the treatment effect within the buffer window $\beta_{mace,trt,buffer}$. Also, for the type 1 error assessment, scenarios are assessed with two sample sizes. Three type 1 error assessment were added, with the addition of a confounding covariate effect on respectively the discontinuation, withdrawal and MACE event. The description of the simulation scenarios can be found in Table 5.2.

5.2 Analysis methods

Time to event will be defined as time from treatment initiation to occurrence of the adverse event. Follow-up times will be censored at the end of a predefined buffer period after treatment discontinuation or after withdrawal, or at the end of follow-up for the patient, whichever comes first.

A Cox proportional hazard model is used to estimate the treatment effect, with or without the X and Z covariates.

A model with inverse probability weighting (IPW) is also used to estimate the treatment effect, with or without the X and Z covariates. The goal here is to weigh patients by the inverse probability of not being censored until the observed time. Since withdrawal is defined conditionally on discontinuation, time-varying weights are considered. Before discontinuation, the probability of being missing is 0, and, if the patient i has discontinued, the probability of being missing is $P(withdraw_i)$. Therefore, the following time-varying weights can be defined as follows for individual i at time t_{ij}:

$$IPW(t_{ij}) = \begin{cases} 1 & \text{if } t_{ij} < T_{disc,i} \\ \frac{1}{1-P(withdraw_i)} & \text{if } t_{ij} \geq T_{disc,i} \end{cases} \tag{6}$$

For the implementation, a logistic model will be fit on patients who have discontinued, to estimate the probability of withdrawal, separately per treatment group, with X and Z as covariates. Then, the logistic model can be used to calculate individual IPW weights.

To summarise, we use time-varying weights for the IPW model. If a patient has a MACE event without any discontinuation, they are assigned a weight of one. If a patient experiences a discontinuation, they are assigned a weight of 1 until treatment discontinuation. Then, if they have not withdrawn from the study, they are assigned the inverse probability of not withdrawing, only during the buffer window.

Table 5.1: Parameter values for the data generating model in the safety/while-on-treatment scenarios.

Parameter	Value in core scenario	Comment	Further values	Comment
MACE				
$\beta_{mace,0}$	log(-log(1-0.1))	10% marginal one-year probability for MACE	log(-log(1-0.3))	Higher event rate
$\beta_{mace,trt}$	log(1.5)	Moderate treatment effect	0, log(2)	No (null hypothesis), or stronger treatment effect
$\beta_{mace,trt,buffer}$	log(1.5)	Treatment effect persists in buffer window	0, log(1.25)	No or reduced treatment effect in buffer window
$\beta_{mace,trt,after}$	0			
$\beta_{mace,X}$	log(2)		0	Combinations of presence and absence of effects of X,Z, and L will be explored
$\beta_{mace,Z}$	log(2)		0	
$\beta_{mace,L}$	0	No unobserved confounders	log(2)	
Discontinuation				
$\beta_{disc,0}$	log(-log(1-0.5))	50% marginal one-year discontinuation probability	log(-log(1-0.2))	Smaller discontinuation probability
$\beta_{disc,trt}$	log(2)			
$\beta_{disc,X}$	log(2)		0	Combinations of presence and absence of effects of X,Z, and L will be explored
$\beta_{disc,Z}$	log(2)		0	
$\beta_{disc,L}$	0	No unobserved confounders	log(2)	
Withdrawal				
$\beta_{wd,0}$	log(0.5/(1-0.5))	Approx. 50% withdrawal	log(0.75/(1-0.75))	Higher withdrawal rate
$\beta_{wd,X}$	log(2)		0	Combinations of presence and absence of effects of X,Z, and L will be explored
$\beta_{wd,Z}$	log(2)		0	
$\beta_{wd,L}$	0		log(2)	
Buffer window				
True window [years]	1/12			
Assumed window [years]	1/12	Match true window	0, 2/12	Assumed window shorter or longer than true window

Table 5.2: Description of the simulation scenarios, TE: Treatment Effect, CE: Covariate Effect, BH: Baseline Hazard, Disc: Discontinuation, Withd: Withdrawal, Null: type 1 error assessment

Scenario	Change from Core Scenario
Baseline	No change (core scenario)
MACE - Higher BH	$\beta_{mace,0} = \log(-\log(1-0.3))$
MACE - No CE	$\beta_{mace,X} = \beta_{mace,Z} = 0$
MACE - Confounding CE	$\beta_{mace,L} = \log(2)$
Disc - Lower BH	$\beta_{disc,0} = \log(-\log(1-0.2))$
Disc - No CE	$\beta_{disc,X} = \beta_{disc,Z} = 0$
Disc - Confounding CE	$\beta_{disc,L} = \log(2)$
Withd - Higher BH	$\beta_{withd,0} = \log(0.75/(1-0.75))$
Withd - No CE	$\beta_{withd,X} = \beta_{withd,Z} = 0$
Withd - Confounding CE	$\beta_{withd,L} = \log(2)$
Assumed window - Lower	0
Assumed window - Higher	2/12 year
TE - Higher	$\beta_{mace,trt} = \log(2)$
TE - No TE buffer	$\beta_{mace,trt,buffer} = 0$
TE - Lower TE buffer	$\beta_{mace,trt,buffer} = \log(1.25)$
Null - High Sample Size	$\beta_{mace,trt} = \beta_{mace,trt,buffer} = 0$ and $HR_{assumed} = 1.5$
Null - Low Sample Size	$\beta_{mace,trt} = \beta_{mace,trt,buffer} = 0$ and $HR_{assumed} = 2$
Null - Confounding CE MACE	$\beta_{mace,trt} = \beta_{mace,trt,buffer} = 0$ and $\beta_{mace,L} = \log(2)$
Null - Confounding CE Disc	$\beta_{mace,trt} = \beta_{mace,trt,buffer} = 0$ and $\beta_{disc,L} = \log(2)$
Null - Confounding CE Withd	$\beta_{mace,trt} = \beta_{mace,trt,buffer} = 0$ and $\beta_{withd,L} = \log(2)$

The IPW model is explored with and without covariates. Covariates are always used for the IPW weights. However, the IPW model without covariates refers to a model without covariates on the time to MACE.

Inference from weighted Cox models will be based on robust standard error.

5.3 Deviations from the study protocol

In the data generation, there was a deviation in the definition of the true treatment effect, where it is now defined by simulating and estimating with covariate effects, to align with the estimation in the data analysis.

In the data analysis, there is one deviation from the protocol is the formula for the IPW weights. For individual i, observed at time j, the probability of not being censored was characterised as $P(T_{disc,i} > t_{mace,j}) \times (1 - P(withdraw_i))$. However, this didn't account for the fact that withdrawal is defined conditionally on the discontinuation status. Therefore, if we consider discontinuation status a time-varying covariate, we get that the probability of not being censored at time t is 1 if discontinuation has not happened at time t, and only depends on the probability of withdrawal otherwise.

5.4 Results

In the following tables, the naming of the specific data generation scenarios are presented. For the parameter values for the core scenario, see the second column of Table 5.2.

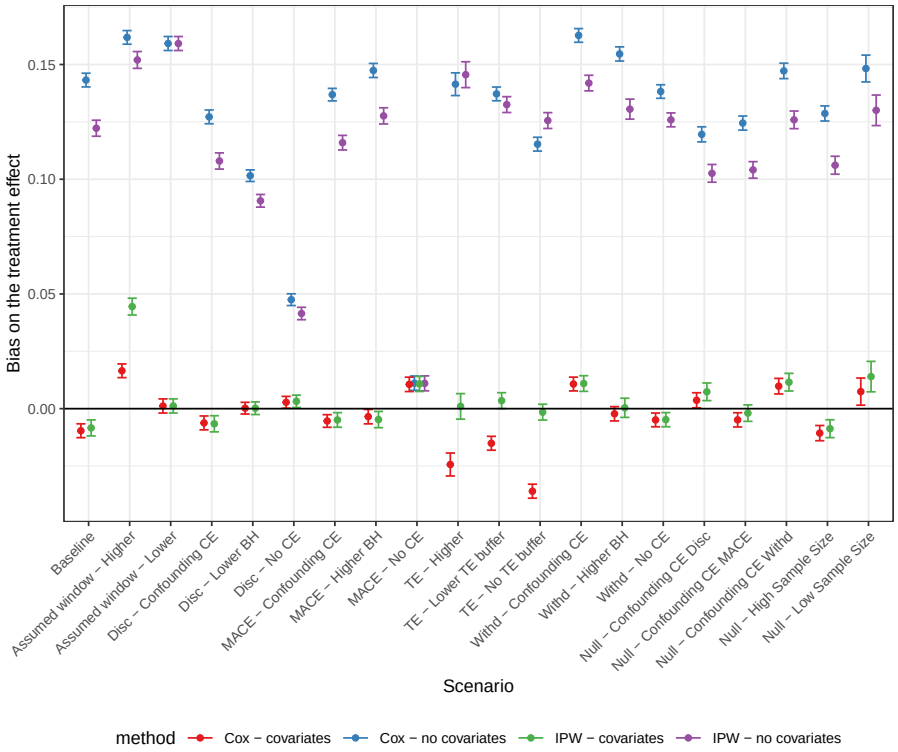
Figures 5.2–5.5 summarise the performance of the investigated methods across the scenarios for the Cox model and IPW model. The results are evaluated in terms of bias of the estimated treatment effect and Hazard Ratio, Root Mean Squared Error (RMSE), Confidence Interval (CI) coverage and width, type 1 error and power. The table containing all values is provided in the supplementary file *mace_results_table.xlsx*.

Figure 5.2 shows the bias of all MACE scenarios. This graphs show that when not adjusting for covariates, the bias on the treatment effect is high (> 0.1) for both the Cox model and the IPW model in all scenarios (except without covariate effects on MACE parameters). For the IPW model with covariates, the bias is low in

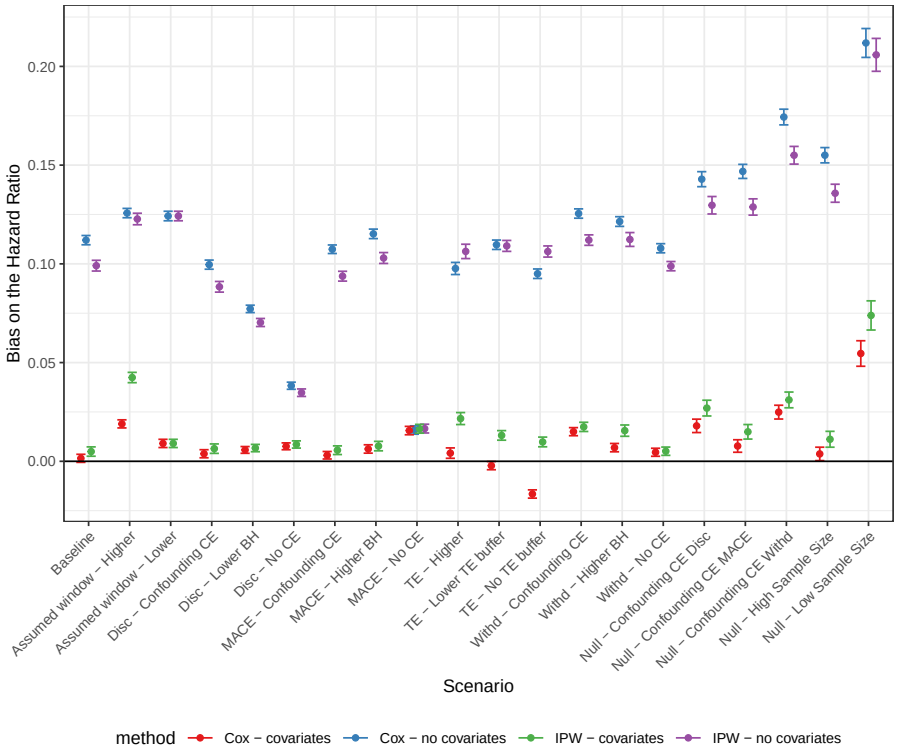
all scenarios (except with an assumed buffer window longer than the value used in the data generating process, underestimating the treatment effect). Finally, for the Cox model with covariates, the bias is low in all scenarios, except for three scenarios: TE - Higher ; TE - Lower TE buffer ; TE - No TE buffer, i.e. when the treatment effect changes in the buffer window (overestimating the treatment effect with a negative bias, since the estimated treatment effect is negative). Figure 5.3 shows the RMSE of the treatment effect in all MACE scenarios, and the results are consistent with the bias on the treatment effect.

In Figure 5.4, we see the mean width of the confidence interval on the Hazard Ratio, as well as the coverage rate of this confidence interval. We can see that the Cox model with covariates always has the lowest mean width of the CI, closely followed by the Cox model without covariates and the IPW model with covariates, then followed by the IPW model without covariates. We also see that for all models, their scenarios simulated under H_0 have a larger CI width than their scenarios simulated under H_1. For the coverage rate, we see that the Cox model and IPW model with covariates have a value very close to 95% in all scenarios, but the models without covariates have a consistently lower coverage rate.

In Figure 5.5, we see the type 1 error rate and the power of the models for all scenarios. First, we can see that for the models without covariates, the type 1 error is significantly lower than 2.5%. For the models with covariates, the type 1 error rate is controlled for all models and scenarios, except for the IPW model with low sample size. For the power we see that, for all scenarios, the models without covariates have a lower power than the models with covariates (except for the scenario with no covariates on the MACE parameters). For all scenarios, we see that the model with the highest power is consistently the Cox model with covariates, ranging from 0.74 to 0.87. We also see that with covariates, the power of the Cox model is higher than the IPW model in most scenarios, except with lower assumed window or without covariate effects on the discontinuation.



(a) Treatment effect



(b) Hazard Ratio

Figure 5.2: Bias

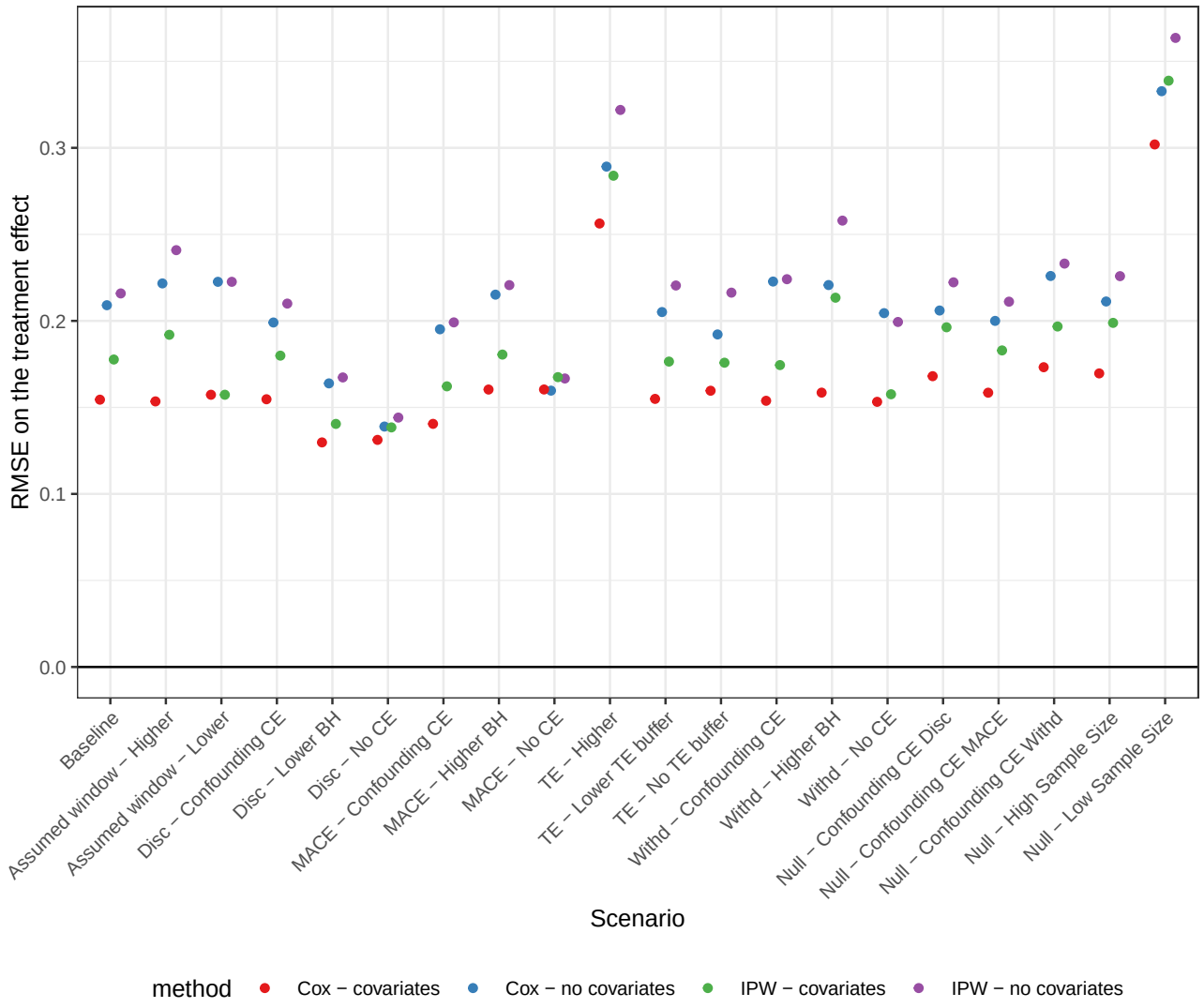
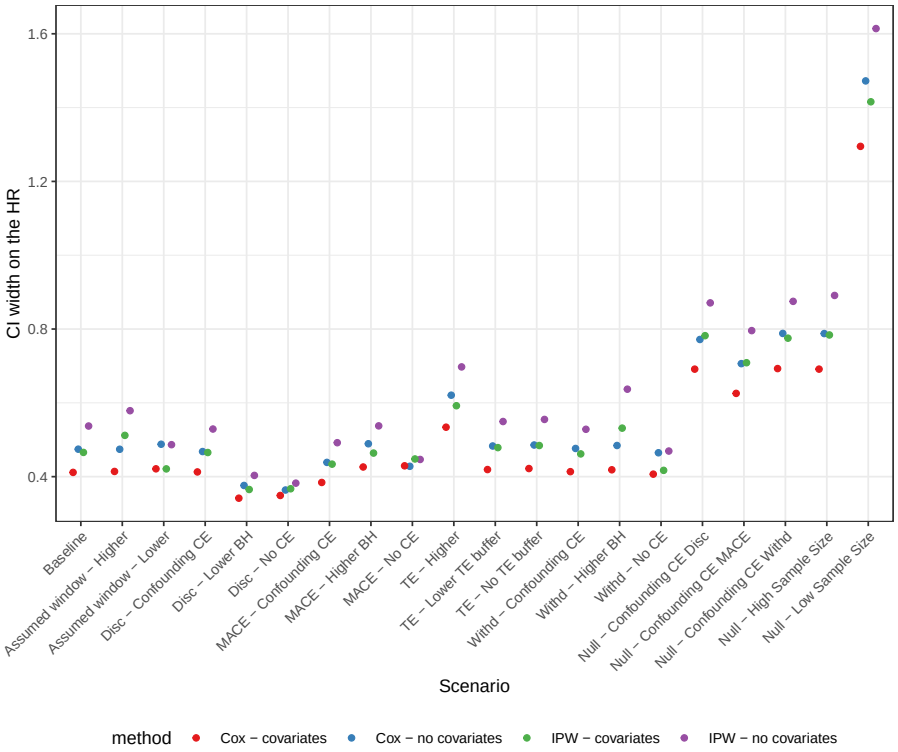
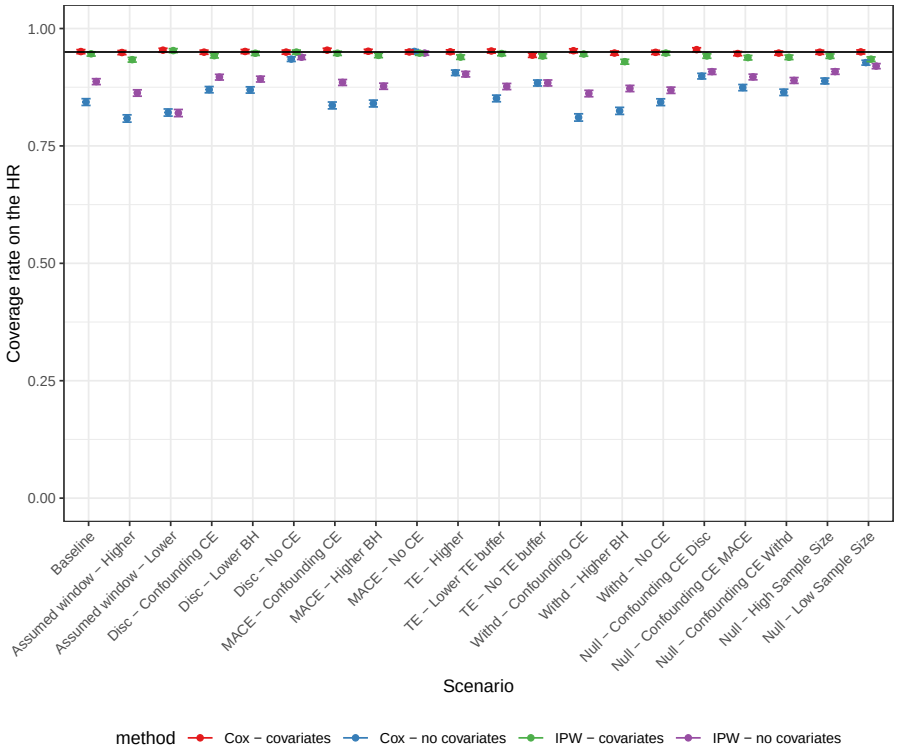


Figure 5.3: RMSE on the treatment effect

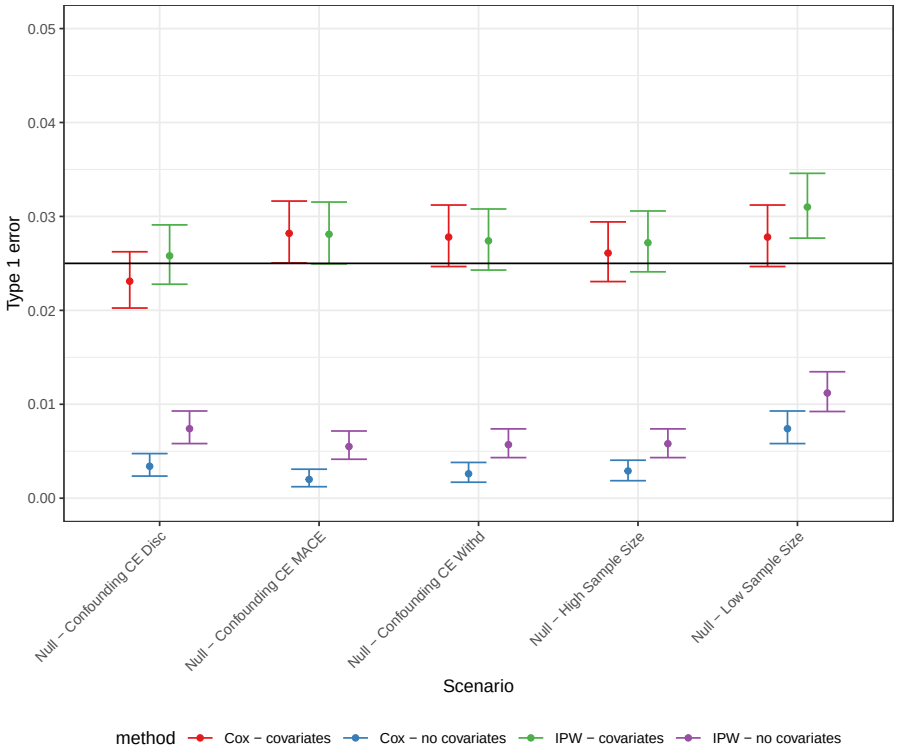


(a) CI width

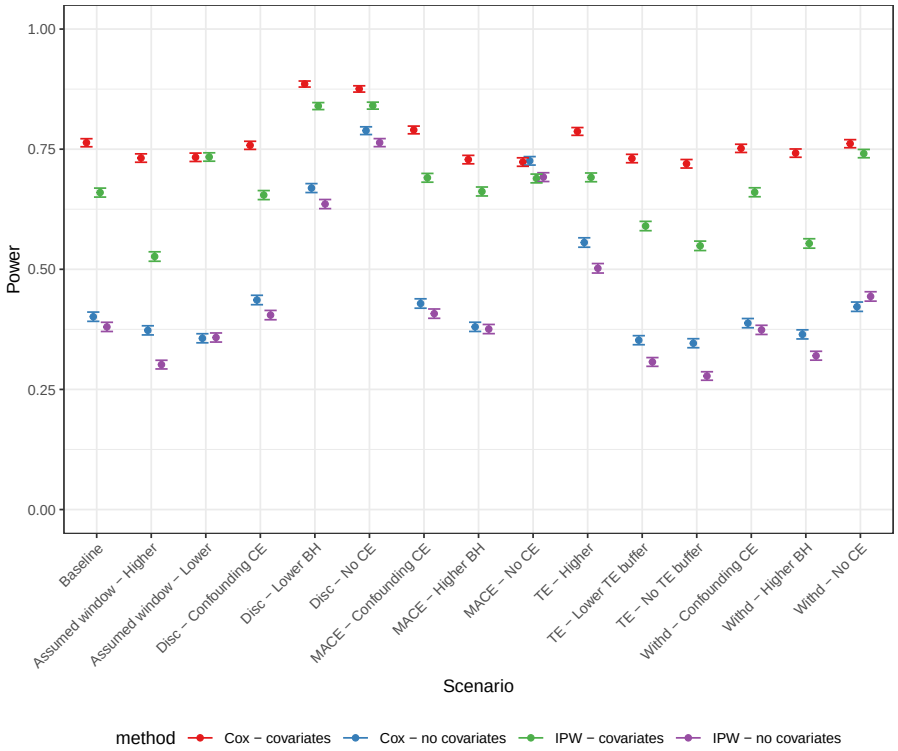


(b) CI coverage

Figure 5.4: Confidence Interval



(a) Type 1 error



(b) Power

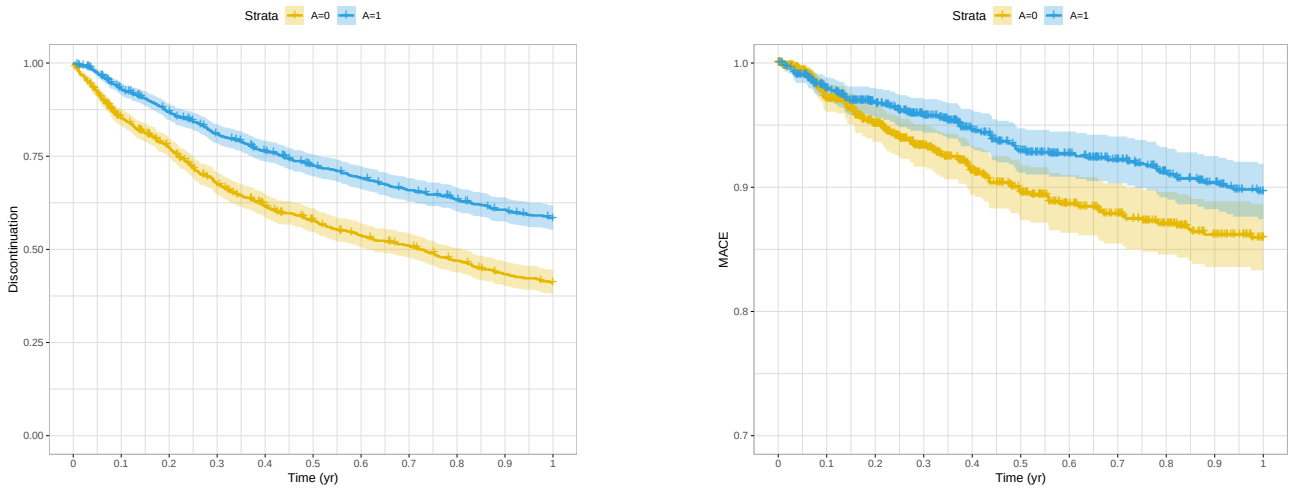
Figure 5.5: Type 1 error rate and power

5.5 Case study

For scenario 4, data simulated according to the core data generating model was used, analysed with the Cox and IPW models, with and without covariates.

In Figure 5.6, we see the Kaplan-Meier curve of a simulated replicate with a sample size of 1910 patients, with the description of the data in Table 5.3.

In Table 5.4, we see the results of the parameter estimation for the Cox and IPW models, with and without covariates. For a true treatment effect of -0.4, we see that without covariates, the Cox and IPW model have similar outputs in terms of estimation (-0.36) and uncertainty, yielding very similar p-values. With covariates, we see that the Cox model estimates a more extreme treatment effect than the IPW model, and with lower uncertainty, translating to a lower p-value.



(a) Discontinuation

(b) MACE

Figure 5.6: Kaplan-Meier plot

	Disc. = 0		Disc. = 1			
	MACE = 0	MACE = 1	MACE = 0		MACE = 1	
			Withd. = 0	Withd. = 1	Withd. = 0	Withd. = 1
A=0	345	96	228	312	2	-
A=1	483	77	161	206	0	-

Table 5.3: Table of events in the simulated case study

Model	**estimated treatment effect**	**SE**	**HR**	**HR - lower CI**	**HR - upper CI**	**p-value**	**true treatment effect**
Cox model							
without covariates	-0.355	0.152	0.701	0.519	0.945	0.019	-0.405
with covariates	-0.557	0.155	0.572	0.422	0.777	0.0003	-0.405
IPW model							
without covariates	-0.358	0.153	0.699	0.517	0.944	0.019	-0.405
with covariates	-0.504	0.159	0.604	0.442	0.825	0.0015	-0.405

Table 5.4: Table of estimated parameters with the Cox and IPW models with and without covariates

5.6 Discussion

As for the other scenario classes, simulation results should be interpreted within the context of the assumed data-generating mechanisms and estimand.

In the results section we saw that the models without covariates had higher bias, lower coverage, lower type 1 error and lower power than the models with covariates, except when data was simulated without covariate effects on the MACE probability. The significant reduction in the type 1 error can be explained by the fact that since the treatment effect is negative, a positive bias in treatment effect (underestimating the effect) would result in a lower probability of the estimated treatment effect to be different from 0.

When comparing the performance of the Cox model and IPW model with covariates, in most scenarios, the Cox model outperformed the IPW model in terms of power. However, we also saw that the IPW model was less biased than the Cox model in scenarios where the treatment effect was different while-on-treatment and during the buffer window after discontinuation. There are also some scenarios where the two models have similar power, namely when the assumed window is shorter. This is because, with an assumed window of 0, the Cox model and IPW models are equivalent (since the weights of the IPW model only change after discontinuation). The performance of the two models is also very close (in terms of bias and power) in the case where withdrawal is simulated without a covariate effect, which could be explained by the fact that in that case, the censoring becomes non-informative.

With a time-varying treatment effect, a bias is observed with the standard Cox model and not with the IPW model, even though the true treatment effect accounts for the discontinuation. In fact, the true treatment effect only accounts for discontinuation and not withdrawal (inducing a missing buffer window). In the case where the treatment effect is the same while on treatment and during the buffer window, maybe a low bias is observed because the model can still estimate the treatment effect while-on treatment. However, if the treatment effect changes in the buffer window, observations are needed in the buffer window to estimate the effect. However, buffer windows can be missing due to withdrawal, which could induce a bias in the estimation of the treatment effect. The IPW model corrects the estimation by giving more weight to the observations in the buffer window.

The models without covariates have larger bias than their counterpart with covariates. This could be due to the non-collapsibility of hazard ratios, since the true treatment effect was estimated from a Cox model with observed covariates, and it is thus defined as a conditional effect, rather than a marginal one. On the other hand, covariates also inform the probability of discontinuation and withdrawal, making the censoring informative, which could also cause the bias in the estimation of the treatment effect without covariates.

Taken together, these results indicate that, in the context of the use of a while on treatment strategy for a chronic disease with a safety endpoint, a standard Cox model was the most powerful approach in all scenarios, while controlling the type 1 error rate. However, this study also show that, the Cox model is more biased when the treatment effect changes after discontinuation. In this work, the IPW model was shown to be less powerful than the Cox model in most scenarios, most likely due to the higher estimated standard error of the treatment effect compared to the Cox model. It can also be explained by the fact that, while the IPW model correctly re-weights patients after discontinuation, the rate of MACE event is low (10% after one year), and the buffer window is short (1 month), so it is very unlikely for a MACE event to happen during the buffer window. Therefore, the performance of the IPW model could also be investigated with other trial settings, such as higher baseline hazard for MACE events, longer true buffer window, higher covariate effects on MACE, discontinuation and withdrawal (for more informative censoring).

6 Conclusions

In the first scenario class, we investigated causal inference methods for estimation and inference under a hypothetical estimand strategy to address the intercurrent event of treatment switching. The studied methods included inverse probability weigthing (IPW), the rank-preserving structural failure time model (RPSFTM), the simplified two stage estimation (TSE), a g-formula approach and, as comparator, a Cox model in which event times were censored at switching. Additionally, extensions of the methods to accommodate the analysis to random censoring by additional inverse probability of censoring weights were explored. The simulations across scenarios with different variations of the data generating mechanism showed that the studied causal inference methods are in principle applicable to account for treatment switching under a hypothetical estimand strategy. However, all studied methods were sensitive with respect to their inherent assumptions. Violations of assumptions frequently led to increased bias of estimates and deviations of type I error rates and confidence interval coverage from their nominal levels, which, depending on the combination of methods and scenarios, could lead to too liberal or too conservative procedures.

Among the studied methods, IPW, and the censoring comparator consistently had more power than the other methods, closely followed by TSE. RPSFTM was found to overall provide inferior power and to be particularly susceptible to violations of its key assumption of a common treatment effect. G-formula was observed to be slightly more biased than other methods. We perceive that challenges with g-formula in an actual application may be to accustom its implementation to a definite study setting, to properly prespecify its internal models to match the true (and unknown) data generating mechanism, and to keep a necessary balance between model complexity and a typically limited amount of data to support the estimation of the underlying models.

The causal inference methods, more so than the censoring comparator, were found to be affected by small sample sizes. It should thus also be kept in mind that these methods may either need some additional small sample adjustments, e.g. when estimating the robust variance for IPW, or should in general be applied to sufficiently large data sets.

For certain patterns of unobserved confounding, specific causal inference methods were found to be particularly more useful than the censoring comparator. In detail, there were three scenarios, in which the the comparator Cox model resulted in biased estimates and accompanying deviations from the nominal type I error rate while specific causal inference methods allowed for negligible bias and type I error control: "Progression - unobserved confounding with death", "switch - unobserved confounding with death" and "death - unobserved confounding with progr. and switch". The first case, unobserved confounding between progression and death affects the comparator method. However, for IPW and TSE, this setup does not correspond to unobserved confounding as their models are conditional on having observed progression. In particular IPW was unbiased in this case with sufficiently large sample size. In the second and third case, unobserved confounding of switching and death affected the comparator, IPW and TSE alike. However, in all three scenarios, the assumption of a constant treatment effect for RPSFTM was met and this method thus also had low or negligible bias. As discussed, though, the power of RPSFTM was in general low, and violation of its assumption, if it occurred, resulted in strong bias, such that this method may nonetheless not be the first choice for inference in a randomized trial.

In the second scenario we considered the exemplary setup of a diabetes trial with a continuous endpoint, where the intercurrent events consisted of initiation of rescue medication and treatment discontinuation and outcome data was potentially affected by missing values. The strategies for handling the intercurrent events were the treatment policy strategy and the hypothetical strategy, making this scenario consist of two different estimands. The three analysis models investigated under the treatment policy strategy, IPW, MMRM, and MI, performed quite equal, and were approximately unbiased in scenarios with a moderate amount of missingness. Under the hypothetical estimand two additional methods were investigated, de-mediation and g-computation. For the hypothetical estimand strategy, method performance depended more strongly on modelling assumptions and scenario characteristics than for the treatment policy strategy. Consequently, method choice was found to be more critical when targeting the hypothetical estimand.

In summary, the results of this simulation showed the similarity of estimator performance under favourable conditions, especially for the treatment policy estimand. For the hypothetical estimand, MMRM often had smaller bias and higher power, while IPW had better type I error control in the increased rescue probability

scenario. Therefore, IPW might be preferable when type I error control under increased rescue probability is prioritized, whereas MMRM is competitive for bias, RMSE, and power in many scenarios.

The third scenario investigated causal inference methods for a preventive vaccine efficacy trial with a binary infection endpoint, where the intercurrent event was non-completion of a two-dose regimen and a principal stratum strategy was used to target vaccine efficacy among always compliers. The studied methods were an instrumental variable (IV) approach using randomised allocation as the instrument and principal score weighting in two variants (with and without the baseline markers in the outcome model), with a per-protocol analysis restricted to observed compliers as a pragmatic comparator. The per-protocol estimator was notably robust across all data-generating configurations, with only a modest negative bias (around one percentage point of VE near the null, decreasing at higher efficacy), near-target power, and one-sided confidence interval coverage at or slightly above the nominal level. The IV approach was likewise broadly robust and achieved the smallest bias in the more complex scenarios, at the cost of slight overestimation at high efficacy. However, a clear breakdown of performance under extreme non-compliance, and a notable negative bias when single-dose efficacy was absent were noted. The principal score methods were more problematic. With the baseline markers included in both the score and outcome models, bias was comparable to the per-protocol estimator, but confidence interval coverage was consistently below nominal, indicating inflated type I error and unreliable interval estimates. The variant omitting the markers from the outcome model was highly sensitive to model misspecification and showed the largest bias and worst coverage in the majority of settings. Part of this behaviour is attributable to a variance-estimation issue (model-based rather than robust standard errors), and preliminary checks suggest that corrected variances would bring coverage closer to nominal while reducing the apparent power advantage.

In summary, the pragmatic per-protocol analysis and the IV approach emerged as the more dependable choices for this estimand across investigated settings. With the small sample bias due to back transformation of log-scale metrics a comparison of bias is challenging. This notwithstanding, instrumental variables showed some improvement compared to the pragmatic per-protocol approach with respect to bias. In a setting where all assumptions of the principal scoring method were met and all corresponding covariates were observed, it appeared to remove most of the bias. However, that was very unstable and resulted in poor performance once deviations from assumptions where introduced in the data generating model. Importantly deviations due to unobserved covariates in the data generating model had substantial impact which would be difficult to address in practice.

The principal score methods should be applied with caution given their sensitivity to outcome-model specification and the difficulty of verifying the underlying assumptions in practice. As in the other scenarios, correct estimation was achievable under appropriate assumptions, but these are largely untestable. Refinements such as robust or bootstrap variance estimation and doubly-robust extensions are promising directions for future work.

The fourth scenario aimed to investigate analysis methods for the analysis of a safety endpoint for a chronic disease with a while on treatment strategy. The goal was to account for a buffer window after treatment discontinuation to record extra events. While, in this context, a standard Cox proportional hazard model was the most powerful approach, the estimation of the treatment effect was biased in the case where the treatment effect changes after discontinuation. On the other hand, while the IPW model was less biased in that case (with correct specification of the withdrawal model), it was also less powerful than the standard Cox model in all scenarios, most likely due to a higher uncertainty in the estimation of the treatment effect. In summary, this fourth scenario highlighted that causal inference approaches such as inverse probability weighting could be used in lieu of 'standard' approaches in the case where there is reason to believe that the treatment effect would change after treatment discontinuation in order to obtain more reliable estimation of the treatment effect, if the assumptions made on the censoring mechanism are correct, in particular the duration of the buffer window.

The interpretation of these simulation findings requires careful consideration of (i) the estimand definition, (ii) the underlying data-generating mechanism, and (iii) the plausibility of identifying assumptions. As an overall finding across the considered scenario classes and estimand strategies, the simulation results showed that the studied causal inference methods are sensitive with respect to their specific model assumptions. Violations of these assumptions can lead to poor performance in terms of power and bias, incorrect inference, and misleading conclusions. In many scenarios, more simple covariate adjusted regression models resulted in similar or superior performance compared to the dedicated causal inference methods, which may limit the potential use of these methods. Furthermore, the methods are based on complex estimation procedures, which involve the estimation

of additional model parameters or weights compared to simpler comparator methods. The estimation of these additional parameters may lead to increased variance of estimates for the treatment effect, which may result in slightly reduced power and reduced robustness in small sample settings.

In conclusion, while correct estimation of different estimands in randomised trials via causal inference methods was found to be possible under appropriate assumptions, it will remain challenging to ensure that, mostly untestable, assumptions are met in an actual application. Future research may focus on developing sensitivity analysis to support the application of the studied methods.

References

Al Tawil, A., McGrath, S., Ristl, R., and Mansmann, U. (2024). Addressing treatment switching in the ALTA-1L trial with g-methods: Exploring the impact of model specification. *BMC Medical Research Methodology*, 24(1):314.

Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91(434):444–455.

Baden, L. R., El Sahly, H. M., Essink, B., Kotloff, K., Frey, S., Novak, R., Diemert, D., Spector, S. A., Rouphael, N., Creech, C. B., McGettigan, J., Khetan, S., Segall, N., Solis, J., Brosz, A., Fierro, C., Schwartz, H., Neuzil, K., Corey, L., Gilbert, P., Janes, H., Follmann, D., Marovich, M., Mascola, J., Polakowski, L., Ledgerwood, J., Graham, B. S., Bennett, H., Pajon, R., Knightly, C., Leav, B., Deng, W., Zhou, H., Han, S., Ivarsson, M., Miller, J., Zaks, T., and Group, C. S. (2021). Efficacy and safety of the mRNA-1273 SARS-CoV-2 vaccine. *The New England Journal of Medicine*, 384(5):403–416. Epub 2020-12-30.

Bang, H. and Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973.

Beckers, F., Karkada, N., Yang, Y., Scott, J., Huang, L., Klinglmüller, F., Fay, M. P., and ... (2025). Adopting the estimand framework in prophylactic vaccine trials. *Vaccine*, 64(127645):127645.

Bell, M. L. and Rabe, B. A. (2020). The mixed model for repeated measures for cluster randomized trials: a simulation study investigating bias and type I error with missing continuous data. *Trials*.

Bhattacharjee, A., Ali, M. R., Kloecker, D., Ling, S., Balasubramanian, G. V., Gillies, C., Davies, M., Khunti, K., and Zaccardi, F. (2025). Five-year trajectories of hba1c by age, sex, ethnicity and deprivation in adults with newly diagnosed type 2 diabetes: Observational study in england. *Diabetes, Obesity and Metabolism*, 27(5):2896–2900.

Bongaerts, B., Kuss, O., Bonnet, F., Chen, H., Cooper, A., Fenici, P., Gomes, M. B., Hammar, N., Ji, L., Khunti, K., et al. (2023). Hba1c trajectories over 3 years in people with type 2 diabetes starting second-line glucose-lowering therapy: The prospective global discover study. *Diabetes, Obesity and Metabolism*, 25(7):1890–1899.

Bowden, J., Bornkamp, B., Glimm, E., and Bretz, F. (2021). Connecting instrumental variable methods for causal inference to the estimand framework. *Statistics in Medicine*, 40(25):5605–5627.

Camidge, D. R., Kim, H. R., Ahn, M.-J., Yang, J. C. H., Han, J.-Y., Hochmair, M. J., Lee, K. H., and ... (2021). Brigatinib versus crizotinib in alk inhibitor-naive advanced alk-positive nsclc: Final results of phase 3 alta-1l trial. *Journal of Thoracic Oncology: Official Publication of the International Association for the Study of Lung Cancer*, 16(12):2091–2108.

Demetri, G. D., Garrett, C. R., Schöffski, P., Shah, M. H., Verweij, J., Leyvraz, S., Hurwitz, H. I., and ... (2012). Complete longitudinal analyses of the randomized, placebo-controlled, phase iii trial of sunitinib in patients with gastrointestinal stromal tumor following imatinib failure. *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research*, 18(11):3170–3179.

European Medicines Agency (EMA) (2023). Guideline on clinical evaluation of vaccines. Scientific guideline EMEA/CHMP/VWP/164653/05 Rev. 1, European Medicines Agency (EMA). Adopted by CHMP 16 January 2023; date for coming into operation: 1 August 2023.

Fay, M. P., Follmann, D., Swihart, B. J., and Dang, L. E. (2026). Vaccine efficacy estimands implied by common estimators used in individual randomized field trials.

Frangakis, C. E. and Rubin, D. B. (2002). Principal stratification in causal inference. *Biometrics*, 58(1):21–29.

Greenland, S. (2000). Small-sample bias and corrections for conditional maximum-likelihood odds-ratio estimators. *Biostatistics*, 1(1):113–122.

Horne, A., Lachenbruch, P. A., and Goldenthal, K. L. (2000). Intent-to-treat analysis and preventive vaccine efficacy. *Vaccine*, 19(2):319–326.

Jiang, Z. and Ding, P. (2021). Identification of causal effects within principal strata using auxiliary variables. *Statistical Science: A Review Journal of the Institute of Mathematical Statistics*, 36(4).

Jo, B. and Stuart, E. A. (2009). On the use of propensity scores in principal causal effect estimation. *Statistics in Medicine*, 28(23):2857–2875.

Lasch, F. and Guizzaro, L. (2022). Estimators for handling covid-19-related intercurrent events with a hypothetical strategy. *Pharmaceutical Statistics*, 21(6):1258–1280.

Lasch, F., Guizzaro, L., Pétavy, F., and Gallo, C. (2023). 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. *Statistics in Biopharmaceutical Research*, 15(2):386–399.

Latimer, N. R. and Abrams, K. R. (2014). NICE DSU technical support document 16: Adjusting survival time estimates in the presence of treatment switching. Technical report, National Institute for Health and Care Excellence (NICE), London.

Latimer, N. R., Abrams, K. R., Lambert, P. C., Crowther, M. J., Wailoo, A. J., Morden, J. P., Akehurst, R. L., and Campbell, M. J. (2017). Adjusting for treatment switching in randomised controlled trials – a simulation study and a simplified two-stage method. *Statistical Methods in Medical Research*, 26(2):724–751.

Latimer, N. R., Bell, H., Abrams, K. R., Amonkar, M. M., and Casey, M. (2016). Adjusting for treatment switching in the metric study shows further improved overall survival with trametinib compared with chemotherapy. *Cancer Medicine*, 5(5):806–815.

Latimer, N. R., White, I. R., Tilling, K., and Siebert, U. (2020). Improved two-stage estimation to adjust for treatment switching in randomised trials: G-estimation to address time-dependent confounding. *Statistical Methods in Medical Research*, 29(10):2900–2918.

Loh, W. W., Moerkerke, B., Loeys, T., Poppe, L., Crombez, G., and Vansteelandt, S. (2020). Estimation of controlled direct effects in longitudinal mediation analyses with latent variables in randomized studies. *Multivariate Behavioral Research*, 55(5):763–785.

Lyles, R. H., Guo, Y., and Greenland, S. (2012). Reducing bias and mean squared error associated with regression-based odds ratio estimators. *Journal of Statistical Planning and Inference*, 142(12):3235–3241.

McGrath, S., Lin, V., Zhang, Z., Petito, L. C., Logan, R. W., Hernán, M. A., and Young, J. G. (2020). gfoRmula: An R package for estimating the effects of sustained treatment strategies via the parametric g-formula. *Patterns*, 1(3).

Michiels, H., Vandebosch, A., and Vansteelandt, S. (2022). Estimation and interpretation of vaccine efficacy in covid-19 randomized clinical trials. *medRxiv*.

Mullahy, J. (1997). Instrumental-variable estimation of count data models: Applications to models of cigarette smoking behavior. *Review of Economics and Statistics*, 79(4):586–593.

Nauta, J. (2020). *Statistics in Clinical and Observational Vaccine Studies*. Springer Series in Pharmaceutical Statistics. Springer International Publishing, Cham. eBook ISBN: 978-3-030-37693-2; Hardcover published 15 March 2020.

Nicolaisen, S. K., le Cessie, S., Thomsen, R. W., Witte, D. R., Dekkers, O. M., Sørensen, H. T., and Pedersen, L. (2024). Longitudinal hba1c patterns before the first treatment of diabetes in routine clinical practice: A latent class trajectory analysis. *Diabetes research and clinical practice*, 212:111722.

Olarte Parra, C., Daniel, R. M., Wright, D., and Bartlett, J. W. (2025). Estimating Hypothetical Estimands with Causal Inference and Missing Data Estimators in a Diabetes Trial Case Study. *Biometrics*, 81(1).

Polack, F. P., Thomas, S. J., Kitchin, N., Absalon, J., Gurtman, A., Lockhart, S., Perez, J. L., Pérez Marc, G., Moreira, E. D., Zerbini, C., Bailey, R., Swanson, K. A., Roychoudhury, S., Koury, K., Li, P., Kalina, W. V., Cooper, D., Frenck, Robert W., J., Hammitt, L. L., Türeci, Ö., Nell, H., Schaefer, A., Ünal, S., Tresnan, D. B., Mather, S., Dormitzer, P. R., Şahin, U., Jansen, K. U., Gruber, W. C., and Group, C. C. T. (2020). Safety and efficacy of the BNT162b2 mrna Covid-19 vaccine. *The New England Journal of Medicine*, 383(27):2603–2615. Epub 2020-12-10.

Ristl, R., Jespersen, L., Klingmüller, F., Fellinger, T., Posch, M., König, F., Centanni, M., Friede, T., Benda, N., Geroldinger, A., Urach, S., Hooker, A., Hohberg, M., Huang, Z., Schneider, L., Schulz, M., Starke, P., and Karlsson, M. (2026). Use of Causal Inference Methods Commonly Used in Non-interventional Studies to Estimate Treatment Effects in Clinical Trials: Simulation Study Protocol. Simulation Study Protocol ROC36, Tender SC03 to FWC EMA/2020/46/TDA/L3.02, European Medicines Agency. CONFIRMS Consortium (CONsortium For Innovation in Regulatory Medical Statistics). Version 1.2. Available via the HMA-EMA Catalogues of real-world data sources and studies.

Robins, J. (1986). A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical modelling*, 7(9-12):1393–1512.

Robins, J. M. and Finkelstein, D. M. (2000). Correcting for noncompliance and dependent censoring in an aids clinical trial with inverse probability of censoring weighted (ipcw) log-rank tests. *Biometrics*, 56(3):779–788.

Robins, J. M. and Tsiatis, A. A. (1991). Correcting for Non-Compliance in Randomized Trials Using Rank Preserving Structural Failure Time Models. *Communications in Statistics: Theory and Methods*, 20(8):2609–2631.

Rousselet, G. A., Pernet, C. R., and Wilcox, R. R. (2021). The percentile bootstrap: A primer with step-by-step instructions in r. *Advances in Methods and Practices in Psychological Science*, 4(1):2515245920911881.

Schoenfeld, D. (1981). The asymptotic properties of nonparametric tests for comparing survival distributions. *Biometrika*, 68(1):316–319.

Siddiqui, O. (2011). Mmrm versus mi in dealing with missing data—a comparison based on 25 nda data sets. *Journal of biopharmaceutical statistics*, 21(3):423–436.

Singh, A. K., Szczech, L., Tang, K. L., Barnhart, H., Sapp, S., Wolfson, M., Reddan, D., and Investigators, C. (2006). Correction of Anemia with Epoetin Alfa in Chronic Kidney Disease. *The New England Journal of Medicine*, 355(20):2085–2095.

Unkel, S., Amiri, M., Benda, N., Beyersmann, J., Knoerzer, D., Kupas, K., Langer, F., Leverkus, F., Loos, A., Ose, C., et al. (2019). On Estimands and the Analysis of Adverse Events in the Presence of Varying Follow-up Times within the Benefit Assessment of Therapies. *Pharmaceutical Statistics*, 18(2):166–183.

White, I. R., Babiker, A. G., Walker, S., and Darbyshire, J. H. (1999). Randomization-based Methods for Correcting for Treatment Changes: Examples from the Concorde Trial. *Statistics in Medicine*, 18(19):2617–2634.

Appendices

A Case study for treatment switching - example code

Read case study data

The Rdata file trt_switch_case_studies.Rdata includes the data set for both case treatment switching case studies as well as the function analyse_oncology_gformula_returndata(), which contains the implementation of the g-formula approach.

```
load("trt_switch_case_studies.Rdata")
#select case study
data<-data_case_study_1
#or uncomment
#data<-data_case_study_2
```

Inverse probability weighting (IPW)

```
library(survival)
#Set of control patients with progression
set<-data$trt == 0 & data$prog_ev == 1
#Fit logistic regression to estimate probability of
#switching at the time of progression, depending on covariates X and W
#the time of progression (secondary baselin)
wmod <- glm(switch ~ X_2BL + W_2BL, family = binomial, data = data, subset = set)
pred <- predict(wmod, type = "response")
#Allocate weights
pred_a <- ifelse(data$switch[set] == 1, pred, 1 - pred)
data$w <- 1
data$w[set] <- 1 / pred_a

#Bring data to a start-stop layout. Date for patients with progression
#is split in two lines, weights are applied to the second line
sdat <- tmerge(data1 = data, data2 = data, id = id, tstop = event_time)
sdat <- tmerge(data1 = sdat, data2 = data, id = id, death_event = event(event_time, ev))
sdat <- tmerge(data1 = sdat, data2 = data, id = id, PD = tdc(prog_time))
sdat$w[sdat$PD == 0 | sdat$trt == 1] <- 1
#Switchers are censored at the time of switching
remo<-sdat$trt == 0 & sdat$PD == 1 & sdat$switch == 1
dat_a<-sdat[!remo,]
#Fit a Cox model with IPW. Use robust variance for inference.
mod_ipw<-coxph(Surv(time=tstart,time2=tstop,event=death_event)~trt+X_0+W_0,
    data=dat_a,weights=w,robust=TRUE,id=id)
summary(mod_ipw)
```

Rank-preserving structural failure time model (RPSFTM)

```
#For RPSFTM the data needs to contain an information regarding
#the relative time in treatment. This is added here:
prep_data_RPSFTM_fun <- function(data) {
  data$time_on_trt <- 0
  data$time_on_trt[data$trt == 1] <- data$event_time[data$trt == 1]
  set_sw <- data$trt == 0 & data$switch == 1
  data$time_on_trt[set_sw] <- data$event_time[set_sw] - data$prog_time[set_sw]
  data$time_on_trt_relative <- data$time_on_trt / data$event_time
  data$max_FU <- data$calendar_end_of_study - data$calendar_start_time
  set_admin_cens <- data$max_FU < data$event_time
  data$max_FU[set_admin_cens] <- data$event_time[set_admin_cens]
  data
}
data_rpsftm <- prep_data_RPSFTM_fun(data)

#The actual analysis is performed using the rpsftm function
#in the trtswitch package
library(trtswitch)
```

```
RPS <- rpsftm(
  data = data_rpsftm, id = "id", time = "event_time", event = "ev", treat = "trt",
  base_cov = c("X_0", "W_0"),
  rx = "time_on_trt_relative",
  psi_test = "phreg",
  alpha = 0.05,
  censor_time = "max_FU",
  autoswitch = TRUE,
  recensor = FALSE,
  gridsearch = FALSE,
  root_finding = "bisection",
  boot = TRUE,
  n_boot = 1000,
  nthreads = 1,
  seed = 851
)
data.frame(HR = RPS$hr,SElogHR = sd(log(RPS$hr_boots),na.rm=TRUE),low = RPS$hr_CI[1],up =
    RPS$hr_CI[2],p = RPS$pvalue)
```

Simplified two-stage estimation (TSE)

```
#For TSE the data needs to contain an information regarding
#the relative time in treatment, similar to RPSFTM.
data_rpsftm <- prep_data_RPSFTM_fun(data)
#The actual analysis is performed using the tsesimp function
#in the trtswitch package. The data set needs to contain covariate
#values at secondary baseline, which is already included in the case
#study data.
TSE <- tsesimp(
  data = data_rpsftm,
  id = "id",
  time = "event_time",
  event = "ev",
  treat = "trt",
  censor_time = "max_FU",
  pd = "prog_ev",
  pd_time = "prog_time",
  swtrt = "switch",
  swtrt_time = "prog_time",
  base_cov = c("X_0", "W_0"),
  base2_cov = c("X_2BL", "W_2BL"),
  aft_dist = "weibull",
  alpha = 0.05,
  recensor = FALSE,
  swtrt_control_only = TRUE,
  offset = 0,
  boot = TRUE,
  n_boot = 1000,
  nthreads = 1,
  seed = 761
)
data.frame(HR = TSE$hr,SElogHR = sd(log(TSE$hr_boots),na.rm=TRUE),low = TSE$hr_CI[1],up =
    TSE$hr_CI[2],p = TSE$pvalue)
```

Non-parametric g-formula

The g-formula approach requires relatively complex code to perform all steps as described in Section XXX. Furthermore, the code is less general but my need adaptations for specific data structures or assumptions regarding the associations between observed variables. It is therefore not displayed in full length here, but exemplary analysis by the function analyse_oncology_gformula_returndata() is shown here. The code of this function is included in the Rdata file trt_switch_case_studies.Rdata together with the case study data sets. Note that this function, as opposed to the function applied in the simulation, returns a two-sided p-value.

```
set.seed(741)
#Note that this function may take substantial computation time when used with
#B=1000 bootstrap samples.
gform<-analyse_oncology_gformula_returndata(B=1000,reps=10,n_ev_cutoff_no_bootstrap=2000,
```

```
    return_data=TRUE)(condition=list(k=10,max_duration=7),data)
as.data.frame(gform[1:5])
```

Censoring event times after switching

```
data_cens<-data
set_cens <- data$trt == 0 & data$switch == 1
data_cens$event_time[set_cens] <- data$prog_time[set_cens]
data_cens$ev[set_cens] <- 0
mod_cens <- coxph(Surv(time = event_time, event = ev) ~ trt + X_0 + W_0,
    data = data_cens)
summary(mod_cens)
```

B Tables from diabetes scenario

Tables B.1, B.2, B.3, B.4, B.5, and B.6 summarise the characteristics of the simulated datasets and the performance of the investigated estimators across scenarios.

Table B.1: For each scenario the number of patients, number of patients in each treatment group, proportion of patients who received rescue medication, proportion of patients with missing data are shown, and the true treatment effect under the treatment policy strategy.

Scenario	n	n trt	n ctrl	Rescue in %	Missingness in %	True tp effect
CS	98	48.8822	49.1178	0.1009418	0.0494398	-0.4416098
NC	130	64.9802	65.0198	0.1056477	0.0497938	-0.4391157
RTE	388	193.9519	194.0481	0.1084278	0.0531062	-0.2147860
MR	98	49.0555	48.9445	0.3432561	0.0523306	-0.3204601
PV	98	49.0696	48.9304	0.0611449	0.0494153	-0.4601917
ROT	98	48.9506	49.0494	0.1010051	0.0488847	-0.4789763
NM	98	49.0130	48.9870	0.1061459	0.0000000	-0.4410794
HM	98	49.0247	48.9753	0.0806806	0.1311051	-0.4417916
CS_null	98	48.9804	49.0196	0.1181735	0.0579541	0.0000000
NC_null	130	65.0511	64.9489	0.1245331	0.0594462	0.0000000
RTE_null	388	194.2919	193.7081	0.1183598	0.0581515	0.0000000
MR_null	98	49.0301	48.9699	0.3890929	0.0595857	0.0000000
PV_null	98	48.9808	49.0192	0.0833061	0.0589704	0.0000000
ROT_null	98	48.9926	49.0074	0.1176602	0.0582469	0.0000000
NM_null	98	49.0309	48.9691	0.1246224	0.0000000	0.0000000
HM_null	98	49.0121	48.9879	0.0910092	0.1696622	0.0000000

Table B.2: For each scenario and for each method the bias is presented. Bias is defined as the average difference between the estimated treatment effect at the final visit and the true treatment effect at the final visit across simulations.

Scenario	ipwtp	mmrmtp	mitp	ipwhyp	dmhyp	gcomhyp	mmrmhyp	mihyp
CS	0.0006502	-0.0016925	0.0004807	0.0047533	0.0169440	0.0111538	-0.0027947	0.0116192
NC	-0.0008941	-0.0008885	-0.0012211	-0.0052149	0.0210029	-0.0053637	-0.0053040	-0.0038430
RTE	0.0011141	0.0001451	0.0012313	0.0038295	0.0223919	0.0084660	0.0001640	0.0064983
MR	0.0032206	0.0009799	0.0013502	0.0145732	0.1248019	0.0277584	-0.0081529	0.0378065
PV	-0.0016425	-0.0031288	-0.0016856	0.0158128	0.0066908	0.0284722	-0.0039473	0.0285220
ROT	0.0023803	0.0008065	0.0034767	0.0050526	0.0010487	0.0182732	-0.0029841	0.0124535
NM	-0.0017520	-0.0017520	-0.0017520	0.0028942	0.0141991	0.0079547	-0.0037271	0.0083724
HM	0.0279325	0.0147959	0.0355904	0.0412961	0.0653089	0.0554190	0.0233721	0.0650841
CS_null	-0.0032367	-0.0030869	-0.0031406	-0.0025971	-0.0018671	-0.0029073	-0.0023090	-0.0028334
NC_null	-0.0007668	-0.0008456	-0.0008103	0.0000997	-0.0003729	-0.0000057	0.0001700	0.0008865
RTE_null	0.0001563	0.0000560	0.0000478	-0.0004121	0.0000287	-0.0004598	-0.0001833	-0.0006469
MR_null	0.0006187	0.0009227	0.0005143	0.0009779	0.0002063	0.0016744	0.0010304	0.0026139
PV_null	-0.0012818	-0.0014263	-0.0014815	-0.0014481	-0.0014616	-0.0017344	-0.0015612	-0.0013153
ROT_null	0.0030588	0.0028168	0.0032357	0.0037160	0.0022363	0.0030622	0.0027775	0.0038298
NM_null	0.0018154	0.0018154	0.0018154	0.0022200	0.0026920	0.0020908	0.0026413	0.0023485
HM_null	-0.0023036	-0.0026282	-0.0029003	-0.0021912	-0.0013864	-0.0026107	-0.0027508	-0.0023443

Table B.3: For each scenario and each method the root mean squared error of the estimates are shown. Root mean squared error is defined as the square root of the mean squared error, where mean squared error is the average of the squared differences between the estimated treatment effect at the final visit and the true treatment effect at the final visit across simulations.

Scenario	ipwtp	mmrmtp	mitp	ipwhyp	dmhyp	gcomhyp	mmrmhyp	mihyp
CS	0.1876746	0.1867252	0.1875379	0.1988223	0.2010368	0.2010251	0.1964578	0.2006226
NC	0.1840961	0.1841798	0.1841757	0.1954471	0.1967988	0.1946760	0.1961591	0.1975928
RTE	0.0927408	0.0922710	0.0926689	0.0986366	0.0995629	0.1001530	0.0967920	0.0994371
MR	0.1889020	0.1880609	0.1887357	0.2387242	0.2406884	0.2345766	0.2301783	0.2467305
PV	0.1857066	0.1847242	0.1857207	0.1920387	0.2095927	0.1949215	0.1908673	0.1948671
ROT	0.1861721	0.1854089	0.1860522	0.1979874	0.1981909	0.1905421	0.1950226	0.1990834
NM	0.1855557	0.1855557	0.1855557	0.1959526	0.1974727	0.1985274	0.1944962	0.1967922
HM	0.1892125	0.1867890	0.1905767	0.2008334	0.2098892	0.2055599	0.1953288	0.2104219
CS_null	0.1841909	0.1837779	0.1843015	0.1963476	0.1953881	0.1967223	0.1931410	0.1984686
NC_null	0.1817718	0.1820279	0.1817499	0.1940933	0.1923399	0.1932031	0.1950058	0.1968291
RTE_null	0.0919131	0.0913392	0.0917821	0.0982097	0.0953776	0.0988292	0.0965914	0.0991133
MR_null	0.1853933	0.1843628	0.1853240	0.2445464	0.2018097	0.2381298	0.2364746	0.2544165
PV_null	0.1823168	0.1812118	0.1824539	0.1913195	0.2056231	0.1907829	0.1910792	0.1931071
ROT_null	0.1835469	0.1823894	0.1832452	0.1966027	0.1959879	0.1871662	0.1935434	0.1982457
NM_null	0.1789837	0.1789837	0.1789837	0.1923704	0.1907758	0.1939772	0.1904131	0.1933431
HM_null	0.1873740	0.1863341	0.1873972	0.1979329	0.1983898	0.1994605	0.1953284	0.2035925

Table B.4: For each scenario and each method the confidence interval coverage is shown. Confidence interval coverage is defined as the proportion of simulations in which the 95% confidence interval for the treatment effect at the final visit contained the true treatment effect at the final visit.

Scenario	ipwtp	mmrmtp	mitp	ipwhyp	dmhyp	gcomhyp	mmrmhyp	mihyp
CS	0.9498	0.9507	0.9505	0.9498	0.9337	0.9351	0.9457	0.9426
NC	0.9527	0.9531	0.9536	0.9518	0.9362	0.9424	0.9503	0.9481
RTE	0.9459	0.9463	0.9464	0.9467	0.9290	0.9332	0.9441	0.9436
MR	0.9484	0.9482	0.9478	0.9486	0.8736	0.9344	0.9416	0.9285
PV	0.9510	0.9508	0.9516	0.9517	0.9402	0.9347	0.9507	0.9454
ROT	0.9538	0.9526	0.9544	0.9522	0.9361	0.9383	0.9504	0.9495
NM	0.9489	0.9494	0.9494	0.9483	0.9354	0.9360	0.9461	0.9461
HM	0.9515	0.9514	0.9465	0.9476	0.9249	0.9304	0.9493	0.9331
CS_null	0.9497	0.9484	0.9498	0.9504	0.9381	0.9365	0.9493	0.9479
NC_null	0.9544	0.9540	0.9548	0.9528	0.9356	0.9391	0.9517	0.9483
RTE_null	0.9465	0.9465	0.9467	0.9453	0.9407	0.9368	0.9454	0.9451
MR_null	0.9498	0.9472	0.9488	0.9469	0.9231	0.9355	0.9360	0.9228
PV_null	0.9463	0.9464	0.9467	0.9488	0.9387	0.9360	0.9458	0.9460
ROT_null	0.9507	0.9504	0.9504	0.9499	0.9326	0.9386	0.9472	0.9448
NM_null	0.9472	0.9481	0.9481	0.9473	0.9332	0.9361	0.9458	0.9438
HM_null	0.9492	0.9470	0.9492	0.9521	0.9381	0.9351	0.9465	0.9390

Table B.5: For each scenario and each method the type 1 error is shown. Type 1 error is defined as the proportion of simulations in which the null hypothesis of no treatment effect was rejected at a significance level of 0.025, among simulations where the true treatment effect was zero.

Scenario	ipwtp	mmrmtp	mitp	ipwhyp	dmhyp	gcomhyp	mmrmhyp	mihyp
CS_null	0.0244	0.0240	0.0235	0.0253	0.0302	0.0317	0.0257	0.0260
NC_null	0.0217	0.0218	0.0212	0.0232	0.0308	0.0294	0.0233	0.0257
RTE_null	0.0254	0.0242	0.0256	0.0256	0.0286	0.0287	0.0254	0.0256
MR_null	0.0253	0.0262	0.0260	0.0268	0.0384	0.0329	0.0320	0.0390
PV_null	0.0273	0.0281	0.0266	0.0254	0.0326	0.0323	0.0281	0.0265
ROT_null	0.0243	0.0244	0.0246	0.0253	0.0340	0.0313	0.0265	0.0285
NM_null	0.0250	0.0247	0.0247	0.0261	0.0331	0.0315	0.0273	0.0275
HM_null	0.0245	0.0253	0.0239	0.0237	0.0285	0.0326	0.0256	0.0296

Table B.6: For each scenario and each method the power is shown. Power is defined as the proportion of simulations in which the null hypothesis of no treatment effect was rejected at a significance level of 0.025, among simulations where the true treatment effect was not zero.

Scenario	ipwtp	mmrmtp	mitp	ipwhyp	dmhyp	gcomhyp	mmrmhyp	mihyp
CS	0.6373	0.6500	0.6393	0.6791	0.6899	0.6530	0.7068	0.6697
NC	0.6445	0.6480	0.6481	0.7051	0.7003	0.7013	0.7047	0.6973
RTE	0.6421	0.6510	0.6446	0.6977	0.6644	0.6630	0.7291	0.6887
MR	0.3861	0.3953	0.3921	0.5087	0.4976	0.5027	0.6060	0.4932
PV	0.6831	0.6927	0.6871	0.6858	0.6662	0.6527	0.7307	0.6710
ROT	0.7101	0.7199	0.7103	0.6802	0.7220	0.6811	0.7127	0.6647
NM	0.6605	0.6591	0.6589	0.6984	0.7168	0.6757	0.7201	0.6913
HM	0.5774	0.6138	0.5645	0.6094	0.5934	0.5768	0.6635	0.5739

C Case study for vaccine scenario - example code

Principal Score Weighting

```
library(emmeans)

# prepare data
dat1 <- dat |>
  within({
    C <- factor(C, levels=c("0", "1"))
    trt <- factor(trt, levels=c("1", "0"))
  })

# propensity score model
mod_ps <- glm(C ~ V + W, subset = (trt==1), data=dat1, family=binomial())
# get odds from predicted values on the link (log-odds) scale
odds <- predict(mod_ps, newdata = dat1, type="link") |>
  exp()

# calculate weights
dat1 <- dat1 |>
  dplyr::mutate(
    weight = dplyr::case_when(
      ((trt==1) & (C=="1")) ~ 1,
      ((trt==0) & (C=="0")) ~ 0,
      ((trt==1) & (C=="0")) ~ 0,
      ((trt==0) & (C=="1")) ~ odds
    )
  )


# calculate outcomes model, weighted by score
# version 1, with V and W in the outcomes model
outcome_mod_1 <- glm(evt ~ trt + V + W, weights=weight, family=binomial(), data = dat1)

res <- emmeans(outcome_mod_1, ~ trt) |>
  regrid("log") |>
  contrast(method="pairwise", type="response") |>
  summary(null=log(1-0.3), side="<", infer=TRUE, level=0.95)

# calculate VE as 1-risk ratio
1-res$ratio
1-res$asymp.UCL
res$p.value

# calculate outcomes model, weighted by score
# version 2, without V and W in the outcomes model
outcome_mod_2 <- glm(evt ~ trt, weights=weight, family=binomial(), data = dat1)

res <- emmeans(outcome_mod_2, ~ trt) |>
  regrid("log") |>
  contrast(method="pairwise", type="response") |>
  summary(null=log(1-0.3), side="<", infer=TRUE, level=0.95)

# calculate VE as 1-risk ratio
1-res$ratio
1-res$asymp.UCL
res$p.value
```

Instrumental Variable Regression

```
library(emmeans)

# prepare data
dat1  <- dat |>
  within({
    # Indicator 1 if complier in treatment group
    T <- factor(ifelse(((trt==1) & (C==1)), 1L, 0L), levels=c("1", "0"))
```

```
    # recode 0/1 coded variables as factors
    V <- factor(V)
    W <- factor(W)
    trt <- factor(trt)
    # recode F/T coded variable to 0/1
    evt <- as.integer(evt)
  })

# estimate stage 1 and save to dat1
x_stage1 <- model.matrix(~ trt + V + W, data=dat1)
y_stage1 <- model.matrix(~T + V + W, data=dat1)
lm_stage1 <- lm.fit(x=x_stage1, y=y_stage1)
dat1[, c("stage1_T", "stage1_V", "stage1_W", "stage1_1")] <-
  lm_stage1$fitted.values[, c("T0", "V1", "W1", "(Intercept)")]

# estimate stage 2
stage2 <- glm(
  evt ~ stage1_1 + stage1_T + stage1_V + stage1_W - 1,
  data = dat1,
  family = poisson(link="log")
)

res <- emmeans(stage2, ~ stage1_T, at=list(stage1_T=c(0,1))) |>
  pairs(type="response") |>
  summary(null=log(1-0.3), side="<", infer=TRUE, level=0.95)

# calculate VE as 1-risk ratio
1-res$ratio
1-res$asymp.UCL
res$p.value
```

Per-Protocol Analysis

```
library(emmeans)

dat1 <- dat |>
  within({
    # recode as factor
    trt <- factor(trt, levels = c("1", "0"))
  })

mod_ve <- dat1 |>
  subset(C==1) |>
  glm(evt ~ trt + V + W, data = _, family=poisson(link="log"))

res <- emmeans(mod_ve, ~ trt) |>
  pairs(type="response") |>
  summary(null=log(1-0.3), side="<", infer=TRUE, level=0.95)

# calculate VE as 1-risk ratio
1-res$ratio
1-res$asymp.UCL
res$p.value
```

D Case study for while-on treatment strategy - example code

While-on-treatment strategy - Cox and IPW model

```
library(survival)

### select patients who discontinue and don't withdraw
disc_not_withdraw <- data$event_disc==1 & data$withdraw==0



### censor MACE events happening after discontinuation + buffer window
```

```
data[disc_not_withdraw,"event_mace"] <-
    ifelse(data[disc_not_withdraw,"t_mace"]-data[disc_not_withdraw,"t_disc"]<buffer,
                                            data[disc_not_withdraw,"event_mace"],
                                            0)

data[disc_not_withdraw,"t_mace"] <-
    ifelse(data[disc_not_withdraw,"t_mace"]-data[disc_not_withdraw,"t_disc"]<buffer,
                                          data[disc_not_withdraw,"t_mace"],
                                          pmin(1,data[disc_not_withdraw,"t_disc"]+buffer))

### Cox model ###
fit_cox_cov <- coxph(Surv(t_mace, event_mace) ~ A+X+Z, data = data, id = ID)




### IPW model ###

### convert dataset into long format with t_disc as a breakpoint

## select patients who discontinue and don't withdraw
disc_not_withdraw <- data$event_disc==1 & data$withdraw==0
data$disc=0
data2 <- data[disc_not_withdraw,]

## first line per ID -> t_start = 0 t_stop = t_mace (if t_mace<t_disc) or t_disc or 1
data$t_mace_start <- 0
data$t_mace_stop <-
    ifelse(data$event_disc==1,data$t_disc,ifelse(data$event_mace==1,data$t_mace,1))
data$event_mace <- ifelse(data$event_disc==1,0,data$event_mace)

## for patients who discontinued -> second line per ID -> t_start = t_disc t_stop = t_mace or
    t_disc + buffer
data2$t_mace_start <- data2$t_disc
data2$t_mace_stop <- data2$t_mace
data2$disc <- 1

## check if buffer window is of length 0 -> if yes return the original dataset
if (mean(data2$t_mace_start==data2$t_mace_stop)==1) {
  data <- data
} else {
  data <- (rbind(data,data2)|>arrange(ID,t_mace_start))
}



### select patients who discontinued and fit logistic model on withdrawal seprately
### per treatment arm
dat_withdraw <- data |> filter(event_disc==1,!duplicated(ID))
fit_withdraw_0 <- glm(withdraw~X+Z,data=dat_withdraw|>filter(A==0),family = binomial(link =
    "logit"))
fit_withdraw_1 <- glm(withdraw~X+Z,data=dat_withdraw|>filter(A==1),family = binomial(link =
    "logit"))

### predict the probability of not withdrawing from the study after discontinuation
p0 <- predict(fit_withdraw_0, newdata = data, type = "response")
p1 <- predict(fit_withdraw_1, newdata = data, type = "response")
lp <- c()
for (i in 1:dim(data)[1]) {
  if (data$disc[i]==1) {
    if (data[i,"A"]==0) {
      lp <- c(lp,1-p0[i])
    }else {
      lp <- c(lp,1-p1[i])
    }

  } else {
    lp <- c(lp,1)
```

```
  }
}

### compute IPW weights
data$IPW <- 1/lp

### estimate IPW model
(coxph(Surv(t_mace_start,t_mace_stop,event_mace)~A+X+Z,data=data,id=ID,weights=IPW,robust=T))
```

E Attached files

Numeric simulation results for the treatment switching scenarios, corresponding to the presented figures in Section 2 are included in the files

- "trt_switch_small_n_H1_results.xlsx"
- "trt_switch_small_n_H0_results.xlsx"
- "trt_switch_large_n_H1_results.xlsx"
- "trt_switch_large_n_H0_results.xlsx"
- "trt_switch_extra_cens_H0_results.xlsx"
- "trt_switch_extra_cens_H1_results.xlsx"

Data for the case studies for treatment switching and the R function analyse_oncology_gformula_returndata() to fit the g-formula is included in the Rdata file "trt_switch_case_studies.Rdata".

Numeric simulation results for the vaccine scenarios corresponding to the results presented in Section 4 are provided in the attached files

- "results_vaccine_scenarios_A1.xlsx"
- "results_vaccine_scenarios_A2.xlsx"
- "results_vaccine_scenarios_B1.xlsx"
- "results_vaccine_scenarios_C1.xlsx"
- "results_vaccine_scenarios_D1.xlsx"
- "results_vaccine_scenarios_extra.xlsx"

simulation results for the while-on-treatment scenarios with MACE endpoint corresponding to the results presented in Section 5 are provided in the attached file "mace_results_table.xlsx".