Simulation study protocol

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	2026-03-03
Version	1.2

Simulation study protocol

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

Abstract

In the planned simulation study, the use of causal inference methods in the analysis of randomised controlled trials in the presence of intercurrent events and missing data under application of the ICH E9 addendum estimand framework will be investigated. The simulation study will comprise four scenario classes, covering 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 as intercurrent event with an estimand aiming at the principal stratum of compliers, and 4) scenarios with a safety endpoint in a chronic disease setting, treatment discontinuation as intercurrent event and a while-on-treatment strategy. Corresponding causal inference analysis methods will include variations of rank-preserving structural failure time models, inverse probability weighting, g-computation, g-estimation approaches and instrumental variable methods. Conventional comparator methods such as mixed models for repeated measures and proportional hazards Cox models will be included. The data will be simulated using parametric data-generating models informed by real example studies, in which the outcome as well as intercurrent events and potential missing data mechanisms depend on baseline and time-dependent covariates and may be mutually dependent. Different choices of parameter values allow for simulations that either meet or violate assumptions of the analysis models in order to explore their characteristics and robustness. Operating characteristics will be assessed in terms of type I error rate, power, bias, variance, mean squared error and confidence interval coverage and width.

Chapter 1

Introduction

The simulation study protocol includes the prespecification of the simulation scenarios and inference methods to be included in the simulation study, including considerations on assumed data generating mechanisms and according numerical values, the scope and specification of causal inference and comparator inference methods, as well as planned metrics for the assessment of operating characteristics of the studied methods.

The simulation study will focus 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, one or more potential intercurrent events and different mechanisms resulting in missing data.

The simulation will comprise 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, and 4) scenarios with a safety endpoint in a chronic disease setting, treatment discontinuation as intercurrent event and a while-on-treatment strategy. The data generating models and analysis methods are informed by example studies identified in the literature review. An overview on the planned scenarios, estimands and methods is given in Table 1.1

1.1 General considerations for the simulation study

The simulation study will focus on randomised trials comparing an experimental treatment to a control treatment, with the aim to test the null hypothesis of no treatment effect and provide a point estimate and a confidence interval for the treatment effect. Within each simulated data set, data of one patient is sampled independently from data of the other patients. The data generating models are parametrised via a set of scenario specific parameters (such as regression coefficients in a true regression model). Further parameters pertain to trial design characteristics such as sample size and recruitment rates. Scenarios are defined through data generating models from which individual patient data will be sampled. The choice of parameters and scenario configurations are guided by findings from the literature review and the review of European Medicines Agency scientific advices and European public assessment reports. The data-generating mechanisms are constructed within a structural causal framework, represented through structural causal models (SCMs) and corresponding directed acyclic graphs (DAGs), ensuring that the temporal ordering and causal dependencies between variables are respected.

To assess the operating characteristics of the studied inference methods in dependence of underlying

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

data generating mechanisms and trial design options, the simulations will be performed over a set of parameter values for the same scenario type. In addition, to assess the dependence on choices within the analysis methods, inferential methods may be applied with different specifications, e.g. including and excluding certain covariates in analysis models.

The following considerations apply to all scenarios:

Variation of parameter values: For each scenario class, a core scenario is defined via a specific set of parameter values in the respective data generating model. Variations in parameter values (such as the effect of covariates or the marginal probability for a given event) are prespecified and these deviations are designed to resemble different mechanistic assumptions on the data generation or to resemble meeting/violation of certain assumptions of different analysis methods. In the simulation, we will explore the core scenarios and mainly further scenarios that deviate from the core scenario in one parameter at a time.

Type I error rate: In all scenarios, simulations under the null hypothesis will be included. Hypothesis tests will be performed at the one-sided 2.5% significance level and confidence intervals will be calculated as two-sided 95% confidence intervals.

Sample size: The sample size is chosen such that the power provided by a usual sample size calculation for the according analysis problem is approximately 80%, given the assumed true effect size under the alternative. For scenarios under the null hypothesis, effect sizes as for the simulations under the alternative are assumed for sample size planning, however the subsequent simulation is performed under the null.

Randomisation: In all simulations we will consider simple randomisation with a 1:1 allocation ratio.

Number of simulation runs: In general, simulations will be based on 10,000 iterations. In case that this number of simulation results in unfeasibly large computation times for certain scenarios or analysis methods, a reduced number of iterations, e.g. 2000, may be considered. Notwithstanding the trade-off between computation time and number of simulation runs, the number of repetitions and the number of bootstrap samples will be chosen to be large enough to ensure sufficient precision of the estimated operating characteristics to be informative.

Operating characteristics: All scenarios will be evaluated in terms of common operating characteristics of the analysis methods, including type I error rate, power, bias, confidence interval coverage and width, see section 6.1.

This protocol is organised as follows: Chapter 1 contains general considerations for the overall simulation setup and aim of this project. The scenario classes are presented by their setting and type of intercurrent event together with the proposed strategy for handling intercurrent events. Chapters 2, 3, 4, and 5 go through each of the four scenarios in greater detail. They are structured in similar ways, providing specifications of the data generating mechanisms followed by specifications of analysis methods and software packages to be used. In Chapter 6, the planned metrics for the assessment of simulation results are described and the approach towards Case studies is outlined, which will be used to showcase the studied causal inference methods for specific examples. Chapters 7 provides details regarding the software implementation and programming of the simulation study.

Chapter 2

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

In this set of scenarios, we consider 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).

Estimand: The considered estimand has the following relevant attributes:

- 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.

Analysis methods: The following causal inference methods will 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 the comparator method we will use 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

As general trial design, we assume event driven studies with a (maximal) recruitment phase of 2 years and a maximal overall study duration of 7 years. The number of events e_{stop} required to stop the study is 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 will be $\alpha = 0.05$, the aimed for power will be $1 - \beta = 0.8$. For scenarios under the alternative hypothesis, the log-hazard ratio assumed for planning $\log(HR_{assumed})$ will be chosen to be equal to the simulated treatment effect $\beta_{death,4}$, such that the actual power will be in the range of 80%. For scenarios under the null hypothesis, the sample size will be calculated assuming $\log(HR_{assumed}) = \log(0.5)$ and $\log(HR_{assumed}) = \log(0.75)$ such that a scenario with a small sample size and a scenario with a larger sample size are included.

The recruitment rate will be chosen as e_{stop} per year, such that with 2 years recruitment 50% of patients will experience an event within the study. Of note, the median survival times in the simulation are chosen such that typically the required number of events is not reached before the end of the recruitment period but is achieved within the remaining five years of follow-up. The actual number of patients in a single simulation run will be sampled from a Poisson distribution with rate $2 * e_{stop}$. Note that with this approach the average study duration is in the order of magnitude of the median survival time added half the recruitment interval length.

Starting dates for patients in terms of calender time are sampled from a uniform distribution over the planned recruitment interval.

In case the required number of events is not reached within the maximal study duration, the simulated study will be stopped after the maximal study duration. This is mainly intended to avoid unrealistically large durations, while the simulation parameters are in general chosen such the maximal duration is not reached in the vast majority of runs.

Treatment switching is assumed to occur at the time of progression with a certain probability. The probability to switch is assumed to depend on observed covariates only, as it is an active decision and we assume the variables that inform this decision are collected. Switching is assumed to be allowed only from the control arm to the treatment arm. (Such that formally, the probability to switch is always 0 for patients in the treatment group.) Note that crossing over from control to experimental treatment is the only intercurrent event included here. In particular, no additional medication, such as start of a new anti-cancer therapy, is considered.

A simulated data set will include, for each patient, the randomised treatment assignment, baseline and time dependent covariates reflecting the disease status measured at regular intervals, the time to progression, a time dependent indicator of the switching status (which is always 0 for subjects in the treatment arm), the time to death or last follow-up and the corresponding event indicator. The detailed data generating mechanisms are described in the following paragraphs. Subscripts i indicate the i-th patient, n denotes the overall sample size.

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.

A continuous baseline covariate X_i is sampled from $N(0, 1)$.

A time dependent covariate $W_i(t),\ t \geq 0$ is modelled as step function with jumps at $t = 0, 1, 2, \ldots, k$ years. For the simulation, k is chosen larger than the maximal trial duration, $k = 8$ will suffice for this purpose. Denote the values over the respective time intervals $W^*_{ij} = W_i(t) : t \in [j, j + 1)$. The data are sampled as $\boldsymbol{W}^*_i = (W^*_{i0}, \ldots, W^*_{ik}) \sim \boldsymbol{N}_k(\boldsymbol{\mu}_W, \boldsymbol{\Sigma}_W)$.

A further time dependent covariate, which will be assumed to be unobserved in the analysis, $L_i(t), t \geq 0$ modelled as step function with jumps at $t = 0, 1, 2, \ldots, k$ years, analogous as described for $W_i(t)$ based on values over the respective time intervals $L^*_{ij} = L_i(t) : t \in [j, j + 1)$. The data are sampled as $\boldsymbol{L}^*_i = (L^*_{i0}, \ldots, L^*_{ik}) \sim \boldsymbol{N}_k(\boldsymbol{\mu}_L, \boldsymbol{\Sigma}_L)$.

The considered parameter values for the mean vectors $\boldsymbol{\mu}_W$ and $\boldsymbol{\mu}_L$ as well as the covariance matrices $\boldsymbol{\Sigma}_W$ and $\boldsymbol{\Sigma}_L$ are shown in Table 2.1. In all scenarios, the covariance matrices are such that the variances (diagonal values) equal 1. In the core scenario, the mean vector $\boldsymbol{\mu}_W$ is identical for both treatment groups, whereas in further scenarios it may take different values for the two treatment groups, see Table 2.1. Similarly, the mean vector $\boldsymbol{\mu}_L$ is identical for both treatment groups in the core scenario and may take different values for the two treatment groups in further scenarios.

Table 2.1: Parameter values for multivariate normal distributions of time-dependent covariates.

Parameter	**Value in core scenario**	**Comment**	**Further values**	**Comment**
$\boldsymbol{\mu}_W$	(0,0,...,0)	Stable disease state	(1,0.5,0,-1,-1,...,-1)	Worsening disease state (capped at bottom), either in both groups or only in control group
$\boldsymbol{\Sigma}_W$	Exchangeable(0.5)	Equal correlation between all time-points	Toeplitz(0.9,0.8,0.7,...,0,...,0)	Decreasing correlation with time, capped at zero correlation
$\boldsymbol{\mu}_L$	(0,0,...,0)	Stable disease state	(1,0.5,0,-1,-1,...,-1)	Worsening disease state (capped at bottom), either in both groups or only in control group
$\boldsymbol{\Sigma}_L$	Exchangeable(0.5)	Equal correlation between all time-points	Toeplitz(0.9,0.8,0.7,...,0,...,0)	Decreasing correlation with time, capped at zero correlation

A time dependent switching indicator $D_i(t) \in 0, 1$ is defined, with $D_i = 0$ indicating that the patient has not switched treatment until time t and $D_i = 1$ indicating that switching has occurred at or before time t.

The hazard function for time to progression will be defined as $\eta_{prog,i}(t) = \exp(\beta_{prog,0} + X_i * \beta_{prog,1} + W_i(t) * \beta_{prog,2}(t) + L_i(t) * \beta_{prog,3} + A_i * \beta_{prog,4})$

In the control group, the probability to switch at the time of progression is defined via the corresponding log-Odds as $logOdds_{switch,i} = \beta_{switch,0} + X_i * \beta_{switch,1} + W_i(t) * \beta_{switch,2} + I(W_i(t) > \omega) * \beta_{switch,3} + L_i(t) * \beta_{switch,4}$.

The term $I(W_i(t) > \omega)$ is an indicator for $W_i(t)$ being larger than a threshold ω. In the simulation, $\omega = 0$ will be considered, such that a reasonable fraction of patients will meet the threshold at the time of progression. (The marginal distribution of $W_i(t)$ is $N(0, 1)$ for all t in the core scenario, however higher values of W_i will reduce the progression rate so that the fraction will be below 50%.) In the core scenario, $\beta_{switch,3} = 0$. In an additional scenario aimed to explore data close to the violation of the positivity assumption for the IPW method, $\beta_{switch,3}$ will be set to $\log(0.9/0.1)/\sqrt{2/\pi} - \log(1.5) = 2.348$ (see Table 2.3). Furthermore, in the core scenario $\beta_{switch,4} = 0$ such that switching only depends on observed covariates. In an additional scenario $\beta_{switch,4} = \log(1.5)$ will be considered to explore the effect of unobserved confounders in the switching model. 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 $\beta_{switch,0} = 0$ and $\beta_{switch,1} = \log(1.5)$, for an average patient with $W_i(t) > 0$, the probability to switch will be 90%.

The hazard for death is defined similar to the hazard for progression, albeit taking into account that the actual treatment is changed after switching. The hazard is defined as: $\eta_{death,i}(t) = \exp(\beta_{death,0} + X_i * \beta_{death,1} + W_i(t) * \beta_{death,2} + L_i(t) * \beta_{death,3} + A_i * \beta_{death,4} + D_i(t) * \beta_{death,5})$.

The time to progression is sampled via the inverse cumulative distribution method as follows: The cumulative distribution function for the time to progression is $F_{progr,i}(t) = 1 - \exp\left(-\int_0^t \eta_{prog,i}(s)ds\right)$. A random variable $u_{prog,i} \sim Uniform(0, 1)$ is sampled. The time to progression is $t_{prog,i}$ such that $F_{progr,i}(t_{prog,i}) = u_{prog,i}$. Switching is then sampled from a Bernoulli distribution with probability corresponding to $logOdds_{switch,i}$. Subsequently, the switching indicator $D_i(t)$ is defined for all time points $t > 0$. Finally, the time to death is sampled via the inverse cumulative distribution method, using $F_{death,i}(t) = 1 - \exp\left(-\int_0^t \eta_{death,i}(s)ds\right)$, sampling a random variable $u_{death,i} \sim Uniform(0, 1)$ and finding t_{death} such that $F_{death,i}(t_{death,i}) = u_{death,i}$. Event times will be censored at the individual time that corresponds to the calendar time where the required number of events occurs (administrative censoring).

The considered values for the parameters of the models for the hazard of progression, for switching and for the hazard of death are listed in Tables 2.2, 2.3 and 2.4.

As an additional censoring mechanism, leading to missing time at risk, a random censoring time $t_{rc,i}$ depending on covariates is defined through the hazard function $\eta_{cens,i} = \exp(\beta_{cens,0} + X_i * \beta_{cens,1} + W_i(t) * \beta_{cens,2} + L_i(t) * \beta_{cens,3})$.

The event time for patient i is censored at $t_{rc,i}$ if $t_{rc,i} < t_{death,i}$ and $t_{rc,i}$ is smaller than the administrative censoring time of patient i. We will consider scenarios without random censoring, with random censoring being independent of other data generating process, with random censoring being dependent on baseline covariates only and with random censoring being dependent on baseline and time-dependent covariates. To generate these scenarios, the considered values for $\boldsymbol{\beta}_{cens} = (\beta_{cens,0}, \ldots, \beta_{cens,3})$ are $\boldsymbol{\beta}_{cens} = \mathbf{0}$ (core scenario), $\boldsymbol{\beta}_{cens} = (\log(-\log(1 - 0.025)), 0, 0, 0)$, such that censoring is independent of other data processes and on average 2.5% of patients would exhibit random censoring within one year unless death or administrative censoring occurs earlier, $\boldsymbol{\beta}_{cens} = (\log(-\log(1 - 0.025)), \log(0.5), 0, 0)$, such that censoring depends on the baseline covariate X, and $\boldsymbol{\beta}_{cens} = (\log(-\log(1-0.025)), \log(0.5), \log(0.5), 0)$ as well as $\boldsymbol{\beta}_{cens} = (\log(-\log(1-0.025)), \log(0.5), \log(0.5), \log(0.5))$, such that censoring is in the order of 2.5% per year and its hazard function depends on baseline and either observed time-dependent covariates or both observed and unobserved time-dependent covariates to a similar extent as the hazard for the primary outcome.

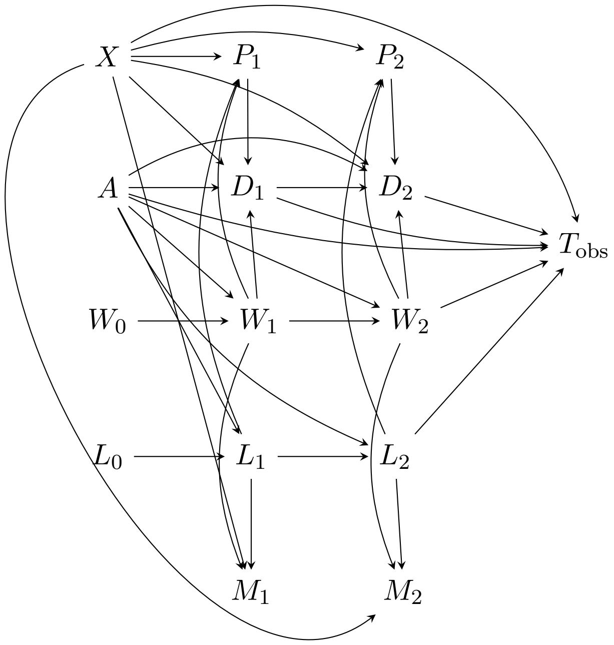

Figure 2.1: Directed acyclic graph (DAG) illustrating the assumed causal structure of the simulated longitudinal data, with baseline covariates (X), randomly assigned treatment (A), measured time varying covariates (W_j), unmeasured time varying covariates (L_j), indicator of censoring other than administrative 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}).

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

True treatment effect

The true treatment effect is determined through the data generating model in terms of a hazard ratio conditional on time-dependent covariates. The summary measure defined in the estimand is 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 $\int_0^\infty Y_0(t)Y_1(t)/(Y_0(t) + Y_1(t)\exp(\beta))(p\lambda_1(t) - (1-p)\exp(\beta)\lambda_0(t))dt = 0$, 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.

In the simulation we consider as true treatment effect the expected value of this summary measure when calculated from an (ideally infinitely) large population under the hypothetical scenario that no treatment switching is possible. The value will be determined by calculating the estimate from an independently simulated large data set with the required number of events set to 100,000 (and the recruitment rate set accordingly to 100,000 per year).

Table 2.2: Parameter values for hazard of progression model

Parameter	Value in core scenario	Comment	Further values	Comment
$\beta_{prog,0}$	log(log(2)/0.5))	Such that the marginal median time to progression under control is around 0.5 years (covariates have mean 0)	log(log(2)/1)	Slower progression
$\beta_{prog,1}$	log(0.5)	We assume larger covariate values are better and the effect is in the same order as the treatment effect.		
$\beta_{prog,2}$	log(0.5)		0	No effect of time-dependent covariate
$\beta_{prog,3}$	0	No unobserved confounders	log(0.5)	Unobserved confounder
$\beta_{prog,4}$	log(0.5)	Hazard ratio treatment vs. control of 0.5, conditional on covariates	0, log(0.75)	No or weaker treatment effect.

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 $\hat{U}_i(\psi)$ are independent of the randomised treatment group, conditional on the covariates X and W at baseline.

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 will be explored by including scenarios where the treatment effect after switching is weaker than the initial treatment effect.

The conditional hazard ratio HR_{RPSFTM} is then estimated by a Cox model using the $\hat{U}_i(\psi)$ for

Table 2.3: Parameter values for switching model

Parameter	Value in core scenario	Comment	Further values	Comment
$\beta_{switch,0}$	log(0.5/(1-0.5))	Such that the marginal probability to switch is around 0.5.	log(0.9/(1-0.1))	High probability to switch, may cause positivity problem
$\beta_{switch,1}$	log(1.5)	We assume larger covariate values are better, and more healthy patients are more likely to be switched (more tolerant to side effects).		
$\beta_{switch,2}$	log(1.5)	We assume larger covariate values are better, and more healthy patients are more likely to be switched (more tolerant to side effects).	0	No effect of time-dependent covariate
$\beta_{switch,3}$	0		$\log\left(\frac{0.9}{0.1}\right)\sqrt{\frac{\pi}{2}} - \log(1.5)$	High probability to switch, may cause positivity problem
$\beta_{switch,4}$	0		log(1.5)	Effect of unobserved confounder

Table 2.4: Parameter values for hazard of death model

Parameter	Value in core scenario	Comment	Further values	Comment
$\beta_{death,0}$	log(log(2)/4))	Such that the marginal median time to death is around 4 years.	log(log(2)/2))	Higher marginal event rate
$\beta_{death,1}$	log(0.5)	We assume larger covariate values are better and the effect is in the same order as the treatment effect.		
$\beta_{death,2}$	log(0.5)		0	No effect of time-dependent covariate
$\beta_{death,3}$	0	No unobserved confounders	log(0.5)	Unobserved confounders
$\beta_{death,4}$	log(0.5)	Hazard ratio treatment vs. control of 0.5, conditional on covariates	0, log(0.75)	No treatment effect (null hpothesis) or weaker treatment effect
$\beta_{death,5}$	log(0.5)	Such that after switching, patients have the full treatment effect	0, log(0.75)	No or weaker treatment effect. $\beta_{death,5} \leq \beta_{death,4}$ in all scenarios

the control group and observed event times for the treatment group. The Cox model will include 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). We will include RPSFTM with and without recensoring. The RPSFTM method will be implemented using the rpsftm function from the trtswitch package in R. Confidence intervals for the hazard ratio will be calculated using the package's default method via the p-value of the ITT logrank test (White et al., 1999). That means, the standard error of the switching-adjusted hazard ratio is estimated as SE such that $|\widehat{logHR}_{ITT}| + z_{1-p/2} * SE = 0$ is met, where p is the two sided p-value for the logrank test of the ITT analysis. The confidence interval HR_{RPSFTM} is then calculated as $\exp(\widehat{logHR}_{RPSFTM} \pm SE * z_{1-\alpha/2})$.

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. 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 may have an impact on the calculation of the counterfactual survival times and this will be explored in scenarios that include unobserved confounder in the data generating mechanisms for progression, switching and time to death.

A Cox model is then fitted using counterfactual survival times for switchers. Inference is based on the estimated parameters and model based standard errors of this Cox model.

The TSE will be 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) will include treatment group and X and W at baseline as covariates in order to estimate the conditional hazard ratio. The TSE will be fitted, both, without recensoring and with recensoring for administrative censoring.

2.2.3 Inverse probability of censoring weighting (IPCW)

The aim of the IPCW is to construct a pseudo population, where 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 will be explored in scenarios that include unobserved confounders.

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 will be estimated using data from the control group patients who experienced a disease progression. The model will include as predictors the baseline covariate X and the time-dependent covariate W with the value observed at the time of progression. The function glm in R will be used for this step. For patients who did not switch, the model based probability to switch, $\hat{\pi}_{sw,i}$ will be calculated, as well as according inverse probability weights $w_i = 1/\hat{\pi}_{sw,i}$. To estimate the treatment effect, a Cox model for the primary endpoint will be fit that includes X

and W at baseline as covariates and furthermore includes time dependent weights: All patients in the treatment group will receive a weight of 1. For patients in the control group, weights will be 1 until the time of progression. After progression, patients who did not switch will receive the weight $w_i = 1/\hat{\pi}_{sw,i}$, as defined above. These weights remain constant across further time, as switching is no longer possible post progression in the considered scenarios. For patients who switched, survival times will be considered as censored at the time of progression. The function coxph from the R package survival will applied for this step. Time dependent weights will be 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 will 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 will be applied by using the function gformula_survival in the R package gfoRmula. For this purpose, the data set will be restructured into a long format with time intervals of 30 days length. Disease progression, switching and death events occuring within a time interval will be included in this data set by binary variables set to 0 if the respective event has not happened up to the start of the interval and set to 1 if the respective event occurs within the interval or has occurred previously. Baseline covariate information will be carried along all intervals. Time-dependent covariates will be included with the value at the beginning of the interval. Similarly, the actual treatment will be coded by 0 (control) or 1 (experimental treatment) according to the treatment state at the beginning of the interval.

The g-computation will be applied separately for both treatment groups. The g-computation via gformula_survival will include a model of the form $W \sim lag1\_W + X + time$ to model the time varying covariate W to depend on its value in the previous time interval (using the lag1 prefix defined in the gfoRmula package), models $y \sim X + W + time$ and $PD \sim X + W + time$ to model the event indicator y and the disease progression indicator PD in a given interval to depend on X and the value of W in the same interval. Time (from baseline) is included to allow for linear time trends in the covariate processes. In the g-computation for the control group data, a further model $trt \sim X + W + time$ will be specified to model the actual treatment status (i.e. switching) to depend on X and W in the same interval, with the restrictions that $trt = 1$ if trt at the previous time point was 1 and that $trt = 0$ if $PD = 0$ (no progression up to the current time point). These models are estimated from all time intervals pooled, assuming that the same mechanism applies at all time intervals.

The key assumption of g-computation is that the used models correspond to the true data generating mechanisms. In the core scenario, this assumption is met. In scenarios with non-linear time trends of time-dependent covariates and in scenarios with unobserved confounders the impact of violations of this assumption will be explored.

The number of time intervals for which to sample trajectories in the subsequent g-computation will be set to the maximum number of intervals among the patients included in the data set. The number of trajectories to simulate (parameter nsimul of gformula_survival) will be set to number of patients in the actual data set in the respective randomised treatment group.

The conditional hazard ratio will be estimated by fitting a Cox model, including treatment group and X and W at baseline as covariates, to the data set that is comprised of the g-computation-simulated trajectories. For this purpose, data from simulated trajectories is brought to the same format, with one line per patient, as the original data set.

Bootstrap will be applied to the whole procedure (fitting the models, g-computation under the assumption of no switching and calculation of the hazard ratio). The number of bootstrap samples will be set to 500. If this value turns out to be unfeasible in terms of computation time (as the g-computation itself is already relatively time consuming) the number of bootstrap samples may be reduced to 200. Bootstrap 95% confidence intervals for the hazard ratio will be calculated using the percentile method. The null hypothesis of no treatment effect will be considered rejected if the confidence interval does not contain the value 1. No formal p-value will be calculated for this method.

2.2.5 Cox proportional hazards model as comparator method

A Cox model will be fit for overall survival, in which event times will be censored at the time of switching. The model will include 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 include scnenarios with censoring, in addition to administrative censoring, that is either independent of covariates, dependent on the baseline covariate X or dependent on both baseline and time dependent covariates, see section 2.1.

The analysis models for all considered methods include 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 will therefore include an additional analysis approach that utilizes time-dependent inverse probability of censoring weights as follows:

A Cox model will be fit for the time to censoring (other than adminstrative), including treatment group and X as baseline covariates and $W(t)$ as time-dependent covariate. Death and administrative censoring will be treated as censoring events in this model. Based on this model, cumulative distribution function for time to censoring given the covariate history will be estimated for each patient. This will be achieved 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 is calculated. Finally, $1/\pi_i(t)$ is applied as time-dependent inverse probability weight in the outcome Cox models for the RPSFTM, TSE and IPCW methods described above. Note that for IPCW, the weights accountign 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.

For the g-computation, the application of inverse probability weights for random censoring is directly implemented in the gfoRmula package which we use for this method. Here, an additional indicator variable C for the random censoring event will be introduced and using the argument censor_model in the function gformula_survival, a model for censoring of the shape $C \sim X + W + time$ will be applied. (Note treatment is not included here because the g-compuatation is performed for each group separately).

Table 2.5: 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**
RPSFTM	Uses g-estimation based on log-rank test	Hazard ratio	trtswitch::rpsftm, recensor $\in \{\mathrm{TRUE}, \mathrm{FALSE}\}$
Two-stage estimation (TSE)	Cox model using counterfactual unswitched survival times	Hazard ratio	trtswitch::tsesimp
Inverse probability of censoring weighting (IPCW)	Weighted Cox model with unstabilised weights	Hazard ratio	trtswitch::ipcw
G-computation			gformula::gformula
Cox PH regression model	Comparator method	Hazard ratio	survival::coxph

Chapter 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 will 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 will be included in the analysis and the focus of the study will be on handling missing outcome data. The causal inference analysis method will be inverse probability weighting. Conventional comparator methods will be 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 will be simulated included a mechanism to generate missing data. In additional simulations, data will be 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. For simplicity, the time unit in the simulation is years and visits are assumed to occur every 1/12 year after treatment start for every patient.

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, an individual treatment response modifier $\alpha_{response,i} \in (0,1)$ 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 exponential 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 HbA1ch 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 -0.498 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

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

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, $\beta_{rescue,y} = \log(150)$ will be 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 continuosly 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

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

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 $\alpha_{response\_rescue,i} \sim Unif(0,1)$ 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.)

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 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 $\beta_{withdraw,resc} = \log(150)$ will be considered on an additional scenario to model close-to violation of positivity.

In the core scenario, data will be set to missing after the visit of the withdrawal decision, such that data is missing at random. In additional scenario, data will be set to missing including the visit of the withdrawal decision, such that data is missing not at random.

These core settings result in approximately 10% missing data at the final visit.

Sample size

For scenarios with different values for the treatment effect δ, the sample size will be 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 will include baseline HbA1c as covariate, the residual variance adjusted for baseline HbA1c will be calculated as $\sigma^2_{adj} = \sigma^2 * (1 - \rho^2)$. Then the sample size per group will be 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.

Figure 3.1 illustrates the causal structure for this scenario.

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, $\log(\frac{0.04}{1-0.04})$	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	Strong dependence of withdrawal on rescue

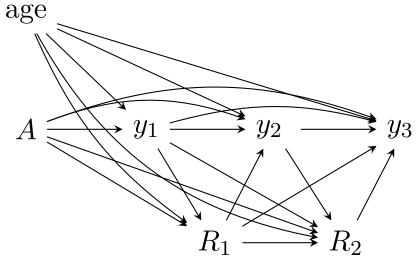

Figure 3.1: Directed acyclic graph (DAG) illustrating the assumed causal structure of the simulated longitudinal data, with the randomly assigned treatment (A), measured baseline covariate (age), indicator of rescue medication at visit j (R_j), and measurement of the outcome at visit j (y_j).

3.2 Analysis methods for the treatment policy strategy

In this section the analysis methods in case of the treatment policy strategy will be 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 are applied to an ANCOVA model that includes only patients with non-missing primary outcome data at the final visit. The ANCOVA model will include treatment group, baseline HbA1c and age at baseline as and HbA1c change from baseline as outcome. The weights are estimated using the ipwtm function in the ipw package in R. Inference will be based using robust standard errors.

The IPW approach assumes a correctly specified propensity model and that the missing data satisfies the missing at random assumption. The impact of violations of these assumptions will be investigated using the scenarios with unobserved confounders and the missing not at random mechanisms where withdrawal masks already the value at the current visit.

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 will utilise an unstructured covariance matrix. The R package mmrm will be used to fit these models (Bell and Rabe, 2020).

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.

3.2.3 Multiple imputation conditional on rescue medication

Multiple imputation is applied in which the imputation is performed conditional on the rescue medication status and observed covariate values in addition to past outcome values. The R package mice will be used. The imputation will include the variables HbAc1 and rescue medication status at each visit to be imputed if missing; observed HbA1c values and rescue status and age at baseline will be used as predictors. The restriction that rescue medication is taken continuously once started will be included using the passive imputation specification in mice. The imputation will be performed for each treatment group separately. A number of 10 multiple imputations will be used. For each imputed data set, an ANCOVA model for change from baseline in HbA1c will be fit including treatment group, age at baseline and HbA1c at baseline as explanatory variables. Results will be pooled via Rubin's rule using the function pool in the mice R package.

3.3 Analysis methods for the hypothetical strategy

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

Missing data after treatment discontinuation will be handled by applying multiple imputation prior to the actual analysis. The subsequent analysis will be performed for each imputed data set and the results will be pooled according to Rubin's rule.

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, 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. Inference will be based using robust standard errors.

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.
- 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 will be 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.

3.3.3 G-computation

Similar as discussed in section 2.2, the idea of g-computation is to model the full joint distribution of age at baseline, HbA1c trajectories and administration of rescue over time by fitting respective models for data at each visit conditional on data at the previous visits. The function gformula_continuous_eof in the R package gfoRmula will be used for this method (McGrath et al., 2020).

The models utilised in the G-computation will be models of the shape $y \sim lag1\_y + age + A + R$ and $R \sim y + age$. A restriction will be implemented such that the rescue indicator R_{ij} equals 1 if $R_{ij-1} = 1$ to take into account that rescue medication is taken continuously once started.

The fitted models 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 mean difference of the outcome between groups is estimated from the sampled data.

Bootstrap will be 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 will be set to 500. If this value turns out to be unfeasible in terms of computation time the number of bootstrap samples may be reduced to 200. Bootstrap 95% confidence intervals for the mean difference will be calculated using the percentile method. The null hypothesis of no treatment effect will be considered rejected if the confidence interval does not contain the value 0. No formal p-value will be calculated for this method.

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

For the MMRM, data after initiation of rescue medication will be 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 furher comparator method. As with MMRM, data after initiation of rescue medication will be set to missing and multiple imputation is applied to obtain outcome values under the assumption that rescue was not available.

The R package mice will be used. HbA1c at each visit will be imputed if missing, based on observed HbA1c values and age at baseline. The imputation will be performed for each treatment group separately. A number of 10 multiple imputations will be used. For each imputed data set, an ANCOVA model for change from baseline in HbA1c will be fit including treatment group, age at baseline and HbA1c at baseline as explanatory variables. Results will be pooled via Rubin's rule using the function pool in the mice R package.

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	
		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	(mediation package in R) glm and lm
G-computation	Modelling all nodes in the DAG	Difference in means	gformula
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

Chapter 4

Scenario class 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.

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

- Instrumental variable (IV) approaches (Bowden et al., 2021)
- Propensity or principal score weighting (Jo and Stuart, 2009)

As comparator method we will 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 (EMA, 2023). In line with (Horne et al., 2000), we will refer to this analysis as per-protocol analysis, acknowledging that in practice other deviations (e.g. use of protocol prohibited medication) - which lead to exclusion from the per-protocol set - may occur.

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:

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

- 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.
- 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 will therefore rely on assumptions implied by the specified data-generating model and the chosen parameter values. Compliance will be 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

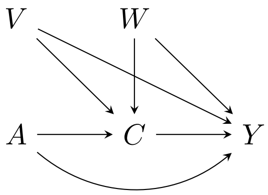

Figure 4.1: Assumed causal structure linking the randomly allocated treatment (A), confounders (V, W), compliance variable (C), and the binary outcome (Y).

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 parameterisation 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 $i = 1, \dots, n$ 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.

4.1.1 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

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

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.

4.1.2 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).

4.1.3 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 generate compliance $C_i(a)$ under potential treatment condition $a \in \{0, 1\}$ using a logistic model conditional on baseline covariates. To this end, we first define for each participant i and each potential treatment assignment $a \in \{0, 1\}$ compliance propensities $p_i(a)$, such that higher propensities imply a higher probability to comply. Specifically, $p_i(a)$ is defined by:

$$\text{logit}\{p_i(a)\} = \gamma_0 + \gamma_W W_i + \gamma_V V_i + \gamma_A a + \gamma_{AW} a W_i. \tag{4.1}$$

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

$$C_i(a) = \mathbb{1}\{U_i < p_i(a)\}, \qquad a \in \{0, 1\}. \tag{4.2}$$

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 are $C_i(0) \geq C_i(1)$ (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).

4.1.4 Infection hazard

The individual infection hazard $\lambda_i(t)$ is specified via a proportional hazards model driven by $\lambda_0(t)$. The baseline hazard $\lambda_0(t)$ will be modeled 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

$$\lambda_i(t) = \lambda_0(t) \exp\Big(\beta_V V_i + \beta_W W_i + A_i\, \theta_i(t)\Big), \tag{4.3}$$

$$\theta_i(t) = \theta^{(1)}(W_i) + \mathbb{1}(t \geq 14)\, C_i \Big\{\theta^{(2)}(W_i) - \theta^{(1)}(W_i)\Big\}, \tag{4.4}$$

$$\theta^{(d)}(W_i) = \beta_A^{(d)} + \beta_{AW}^{(d)}\, W_i, \qquad d \in \{1, 2\}. \tag{4.5}$$

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 will 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.

4.1.5 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.

To connect to the potential outcomes notation, we denote by $T_i(a)$ the infection time that would be observed under treatment assignment $A_i = a$, and define the corresponding binary endpoint as

$$Y_i(a) = \mathbb{1}(T_i(a) \leq \tau)\,, \qquad a \in \{0, 1\}.$$

In the simulated trial we observe $Y_i = Y_i(A_i)$.

True treatment effect

Define the principal stratum of *always-compliers* as

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

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

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

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 $t \leq 14$ and $\lambda_0(t) = \lambda_{\mathrm{post}}$ for $14 < t \leq \tau$, 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 standardization 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 $\mathrm{VE}_{\mathrm{AC}} = 1 - \mathrm{RR}_{\mathrm{AC}}$.

4.1.6 Simulation scenarios and parameter ranges

We will organize 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 initial, interpretable suggestions; where needed, they will be calibrated within the implemented data-generating models to achieve realistic target quantities such as overall compliance and cumulative incidence. Where informative, we may also expand the parameter ranges in Table 4.1 beyond the primary, realistic settings to construct "stress-test" scenarios that are expected to challenge assumptions and induce method failure.

- **Scenario A0: 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 A1: 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 will be included where V drives baseline risk and compliance heterogeneity and W drives protection and compliance heterogeneity. This scenario is intended as a representative stress-test and will not be implemented across the full combination of all parameter options.
- **Observed-data variants.** Robustness to limited covariate availability will be 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) will be 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 will be used to develop case studies for illustration, as described in section 6.2.

4.2 Analysis methods

Let S be the principal stratum of always compliers. The vaccine efficacy in this stratum is then defined as one minus the relative risk of infection attributable to vaccination:

$$VE_{\text{AC}} = 1 - \frac{P\left(Y(1) = 1 \mid S_{\text{AC}}\right)}{P\left(Y(0) = 1 \mid S_{\text{AC}}\right)}$$

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

For simplicity, we provide notation only for a scenario where marker V is observed, and W unobserved. Scenarios where we assume different settings of observed and unobserved confounders will be analysed in analogy.

4.2.1 Instrumental variable (IV)

A method utilising the instrumental variable approach will be applied, using treatment allocation as the instrument. This approach identifies principal stratum effects under the assumption of monotonicity, and in addition it does not rely on the assumption of no unmeasured confounding.

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.

Treatment allocation is used as the instrument since it satisfies the required assumptions; relevance, randomisation and the exclusion restriction (Bowden et al., 2021; Angrist et al., 1996; Jiang and Ding, 2021; Frangakis and Rubin, 2002).

The function ivreg in the R package ivreg will be applied to calculate the instrumental variable model using two stage estimation. The models will include Y as outcome, furthermore an endogenous treatment variable $T \in 0, 1$, with $T_i = 1$ if $A_i = 1$ and $C_i = 1$, and $T_i = 0$ otherwise, and A_i as instrumental variable. An additional model that further includes W as exogenous variable will also be fit. The vaccine efficacy and its variance and subsequent p-value and confidence intervals are calculated by application of the delta method to estimated means and their variance covariance matrix from the linear instrumental variable model.

4.2.2 Principal score weighting

This procedure is presented in (Jo and Stuart, 2009), and it approaches to approximate causal effects among likely compliers.

This approach works under the assumption of principal ignorability (i.e. $\beta_V = 0$ if V is unobserved) and monotonicity (i.e. no placebo compliers $\gamma_A + \gamma_{AW} W_i \geq 0$).

1. Using only the treated participants, fit the principal score model predicting compliance given the covariates by a logistic regression

$$\text{logit}(P(C_i = 1 \mid V_i)) = \beta_0 + \beta_1 V_i$$

2. Predict principal scores p_i for all participants.

3. Assign weights

$$w_i = \begin{cases} 1, & \text{if } A_i = 1, C_i = 1 \\ 0, & \text{if } A_i = 0, C_i = 0 \\ 0, & \text{if } A_i = 1, C_i = 0 \\ \frac{p_i}{1-p_i}, & \text{if } A_i = 0, C_i = 1 \end{cases}$$

4. The complier average causal effect (CACE) is estimated using these weights in a logistic regression model with outcome variable Y and explanatory variable A. The estimated infection probabilities under treatment and under control are calculated from the estimated regression coefficients of this model. Based thereon, an estimate for the vaccine efficacy is calculated. The variance and subsequent p-value and confidence intervals for this estimate are calculated by application of the delta method to the robust variance covariance matrix estimate of the logistic regression model.

It relies on the assumption of principal ignorability, that the stratum membership is independent of the potential outcomes conditional on the observed history.

4.2.3 Per-protocol analysis

In practice the primary analysis for vaccine efficacy is frequently based on the per-protocol set (Horne et al., 2000; Baden et al., 2021; Polack et al., 2020). Consequently, as comparator method, we will estimate vaccine efficacy restricted to participants who adhere with vaccination according to their observed compliance status using a binomial model (Nauta, 2020). Such an analyses, limited to ”observed compliers”, can be considered a “pragmatic” approach to approximates the biological effect, yet is subject to selection bias unless restrictive and implausible assumptions apply. Therefore, evaluating the potential bias of such an analysis under various scenarios of non-compliance represents a practically relevant objective.

Table 4.2: 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
Instrumental variable (IV)	Using allocated treatment as instrument	Vaccine efficacy	ivreg::ivreg
Principal score weighting	Regression method using weights from principal score model (Jo and Stuart, 2009)	Vaccine efficacy	stats::glm
Per-protocol analysis	Comparator method	Vaccine efficacy	stats::glm

4.3 Discussion

A central methodological challenge in preventive vaccine efficacy trials is the definition and estimation of estimands that aim to capture a “biological” effect of vaccination (Beckers et al., 2025) in the presence of post-randomisation events such as incomplete dosing. Principal stratum estimands provide a coherent causal target for the effect among participants who would comply with the full vaccination schedule under either treatment assignment, but require assumptions for identification and are sensitive to deviations from these assumptions (Gilbert et al., 2003; Beckers et al., 2025).

The primary aim of the simulation is to evaluate and compare analysis strategies for estimating the principal stratum estimand under controlled settings that vary the extent to which the required identification conditions hold. We will consider instrumental variable approaches (Bowden et al., 2021) and principal score methods (Jo and Stuart, 2009), with per-protocol analyses serving as reference comparators. Across scenarios, we will vary key features that affect identifiability and robustness, including: (i) settings with no confounding of compliance and infection risk beyond randomisation; (ii) settings where confounding is fully captured by observed baseline covariates; and (iii) settings where residual confounding is induced by unobserved baseline factors affecting both compliance and infection risk. We will additionally explore effect modification of vaccine efficacy by baseline covariates.

These scenario classes allow performance assessment under “favourable” conditions where method-specific assumptions are satisfied (e.g. adequate separation of compliance patterns by observed covariates) and under “unfavourable” conditions where assumptions are violated (e.g. unobserved confounding, weak separation leading to practical non-identifiability, or structural features that undermine point identification of the targeted principal stratum estimand). The resulting comparisons will inform which methods are most robust to plausible deviations from idealised conditions, and under what circumstances the targeted estimand may be estimable with acceptable bias and precision in preventive vaccine trials.

An additional issue in many vaccine trials is the occurrence of infections during the ramp-up period, i.e. before full vaccine-induced protection is achieved. Such early infections can complicate the interpretation of a biologic estimand tied to completion of a multi-dose schedule, because the outcome process may begin before the post-dose-2 effect is operative. While this protocol focuses on scenarios that avoid infections prior to dose 2 (by design in the data generating model), the ramp-up problem remains important in practice and motivates alternative estimand choices, including hypothetical estimands that conceptualise elimination of the ramp-up period (Michiels et al., 2022). We anticipate that the implications of ramp-up infections for estimand definition and analysis could be revisited in a separate case study.

Chapter 5

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

In the fourth scenario we will 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 will be 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 will assume that the effect of the medication persists for a certain time after discontinuation. In the estimation of the while-on-treatment estimand we will include a buffer time window after discontinuation in which adverse events are still counted. Different lengths of this buffer time window (allowing for differences between the true and the assumed buffer time) will be explored. Furthermore, the decision to discontinue treatment may depend on some information regarding the patient's disease state or risk for an adverse event, hence the discontinuation and the occurrence of a subsequent event may not be independent. To assess the impact of such a relation, we will study two different data generating mechanisms. In the first mechanism, patients with a larger risk for the adverse event will also have a larger propensity for treatment discontinuation. In the second one, the decision for treatment discontinuation and the adverse event process are independent. These mechanisms are implemented in the data generation by differential dependence of the hazard for the adverse event and the hazard for treatment discontinuation on covariates.

We will 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 will 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 increased while the patient is under the effect of the experimental treatment compared to the control treatment. We will consider the hazard ratio as a summary measure. This summary measure addresses the inherent risk at each time-point, which may be increased while on treatment. The interpretation of the hazard ratio depends on the plausibility of the proportional hazards assumption. This assumption may be relaxed by calculating time-dependent hazard ratios, e.g. for predefined intervals of time under treatment. In the simulation we will, however, focus on a single hazard ratio covering the full time under treatment. An alternative summary measure would be the difference in proportions. This summary measure would address the overall risk for a treatment related adverse event, albeit under the particular treatment regimen of the study, including the study specific decision process of when to stop treatment. The risk difference therefore may not allow for straight forward generalisation.

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.

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.

Analysis methods
A cause specific proportional hazards model will be used with inverse probability weighting to account for missing data after premature study discontinuation. See section 5.2. For comparison an unweighted the cause specific model will fit.

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.

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 as safety endpoint the time to occurrence of a major cardiac adverse events (MACE), see e.g. Singh et al. (2006). The aim is to investigate whether the new treatment is superior in terms of this safety outcome compared to control while patients are on treatment.

The maximal follow-up for a patient in the study is one year and event times greater than one year are censored.

The sample size will be chosen to provide approximately 80% power under the alternative. The number of events e_{mace} is determined by Schoenfeld's formula Schoenfeld (1981) as $e_{mace} = ((z_{1-\alpha/2} + z_{1-\beta})/\log(HR_{assumed}))^2 * 4$. Here z_γ is the γ quantile of the standard normal distribution.

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

As before, patients are indexed $i = 1, \ldots, n$. 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 probability 0.5, corresponding to an allocation ratio of 1:1.

Two continuous baseline covariates X_i, Z_i and a further covariate L_i, which is considered as unobserved covariate in the analysis, are sampled from independent $N(0, 1)$ distributions.

A constant hazard for treatment discontinuation is modelled to depend on the covariates as

$$\eta_{disc,i} = \exp(\beta_{disc,0} + (0.5 - A_i) * \beta_{disc,trt} + X_i * \beta_{disc,X} + Z_i * \beta_{disc,Z} + L_i * \beta_{disc,L})$$

In the core scenario, $\beta_{disc,0} = \log(-\log(1 - 0.5))$, such that the marginal probability for discontinuation within the one year follow-up period is approximately 50%. (Note the mean is zero for all covariates.)

To allow for well detectable effects of confounding in the simulation, covariate effects are chosen to be rather pronounced, with a value of $\log(2)$ for $\beta_{disc,trt}$, $\beta_{disc,X}$, $\beta_{disc,Z}$ and $\beta_{disc,L}$ in scenarios where the respective covariate has an effect and parameter value of 0 otherwise.

The hazard for MACE is modelled similarly as

$$\eta_{mace,i}(t) = \exp(\beta_{mace,0} + (1 - A_i) * \beta_{mace,trt}(t) + X_i * \beta_{mace,X} + Z_i * \beta_{mace,Z} + L_i * \beta_{mace,L})$$

Note that $\beta_{mace,trt}(t)$ is time dependent and may take different values before treatment discontinuation, within the buffer window and after the buffer window, as indicated in Table 5.1.

In the core scenario, $\beta_{mace,0} = \log(-\log(1-0.1))$ to provide a marginal probability of approximately 10% for a MACE event within one year under experimental treatment, which resembles the one-year event rate reported for patients treated with epoetin alfa in Singh et al. (2006).

The effect of treatment prior to treatment discontinuation is chosen as $\beta_{mace,trt} = \log(1.5)$, which is a slightly stronger effect than seen for the comparison between high and low target hemoglobin in Singh et al. (2006).

After discontinuation, the effect of treatment on the hazard for MACE is modified. Within a buffer window of 1/12 year, $\beta_{mace,0}$ remains at $\log(1.5)$ in the core scenarios and is reduced to $\log(1.25)$ in a further scenario. After the buffer window, $\beta_{mace,0}$ is set to 0. Note that the treatment effect is here modelled such that the experimental treatment does not increase the risk of MACE compared to baseline while the control treatment leads to an increased risk for MACE.

Regression coefficients for remaining covariates will be set to either $\log(2)$ or 0, such that their magnitude of effects (if included) is comparable for the hazard of MACE and the hazard of discontinuation.

At the time of treatment discontinuation, the patient may withdraw from the study with a certain probability modelled as

$$logOdds_{withdraw,i} = \beta_{wd,0} + X_i * \beta_{wd,X} + Z_i * \beta_{wd,Z} + L_i * \beta_{wd,L}$$

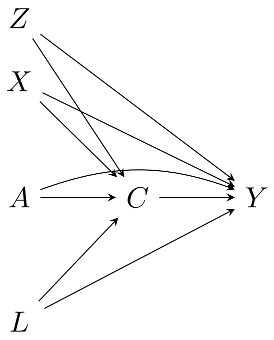

Figure 5.1: Assumed causal structure linking the randomly assigned treatment (A), measured covariates (X, Z), unmeasured covariates (L), the indicator for censoring due to withdrawal (C) and the outcome (Y).

A value $\beta_{wd,0} = \log(0.5/(1-0.5)) = 0$ is considered to result in a probability to withdraw, given treatment was discontinued, in the order of magnitude of 50%. (Though to some extent above 50% as conditional on discontinuing treatment, covariate mean values are greater than 0.) $\beta_{wd,X}$, $\beta_{wd,Z}$ and $\beta_{wd,L}$ are set to log(2) or 0, depending on whether the respective covariates are assumed to have an effect on the decision to withdraw. The decision to withdraw in reaction to treatment discontinuation is considered not to depend on the treatment group.

The considered parameter values are listed in Table 5.1.

All event times are censored after one year due to the considered maximal follow-up time.

For a summary on the dependencies in the data generating mechanism see Figure 5.1.

True treatment effect

Similar to the oncology scenarios, the true treatment effect is defined in terms of the expected value of the estimate of a Cox model. To calculate this value an independently simulated data set with the required number of events set to 100,000 will be generated, in which the event data is observed completely. (That means, there will be no missing data in the buffer window.) Time-to-event data after the buffer window will be considered as censored. From this data set, the hazard ratio will be calculated from Cox model for MACE with treatment group as only covariate, as well as from a Cox model with additional covariates X and Z. These will serve as true values for the respective marginal and conditional hazard ratio.

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 at the end of follow-up for the patient, whichever comes first. The analysis described below will be performed with different buffer periods, including buffer lengths of 0, 1 and 2 months.

Cause specific proportional hazards model
A Cox model will be fit for the time to event, with treatment group as single covariate. In addition, a Cox model with additional covariates X and Z will be calculated.

Inverse probability weighting (IPW)

For every patient who discontinued treatment, the probability to withdraw from the study at this

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
$\beta_{mace,trt}$ before discontinuation	log(1.5)	Moderate treatment effect	0, log(2)	No (null hypothesis), or stronger treatment effect
$\beta_{mace,trt}$ within buffer post discontinuation	log(1.5)	Treatment effect persists in buffer window	0, log(1.25)	No or reduced treatment effect in buffer window
$\beta_{mace,trt}$ post buffer	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/(1-0.25))	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

occasion, $P(withdraw_i)$, is estimated using a logistic regression model with covariates X and Z, fit separately per treatment group using data from all patients who discontinued treatment. Note that in the simulation scenarios, withdrawal can only occur at the time of treatment discontinuation such that a logistic regression (as opposed to a time-to event model) is appropriate.

Time dependent inverse probability weights are calculated as $w_i(t) = 1$ for time points t prior to treatment discontinuation and $w_i(t) = 1/(1 - P_{withdraw,i})$ for time points within the assumed buffer window after treatment discontinuation. (Time points after the buffer window may formally be assigned a weight of 0, however these time-points are not included in the analysis as all event times are artificially censored after the buffer.)

These weights are used in the calculation of weighted Cox models for the time to MACE. A model with treatment group as single covariate as well as a model with X and Z as additional covariates will be considered.

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

Chapter 6

Presentation of simulation results and case studies

6.1 Metrics for assessment and presentation of results

To evaluate the performance of the different estimators, we will assess several standard simulation metrics that collectively describe both accuracy and reliability. Bias quantifies the systematic deviation of the estimator's average value from the true parameter, indicating whether a method tends to over- or underestimate the effect. Variance measures the dispersion of estimates across simulation replicates, reflecting precision. The mean squared error (MSE) combines both components, capturing overall estimation accuracy through MSE = Bias^2 + Variance.

We will also evaluate the width of the 95% confidence intervals (CIs) as an indicator of estimator precision, and the empirical coverage probability, to assess the validity of the estimated uncertainty. Finally, to evaluate inferential performance, we will compute the empirical type I error rate, defined as the proportion of simulations incorrectly rejecting the null hypothesis when it is true, and the power, defined as the proportion correctly rejecting the null when a true effect exists. Together, these metrics provide a comprehensive assessment of each method's bias–variance trade-off, uncertainty calibration, and ability to detect true treatment effects under varying data-generating scenarios.

The results will for each scenario be presented graphically, to visualise the impact of systematic variations of design elements. A condensed overview with averaged metrics across different settings will also be considered.

Results for case studies will be presented in terms of descriptive statistics for the analysed data set. For each applied analysis method, the estimated summary measure together with a 95%confidence interval will be reported as well as the result of a hypothesis test for the null hypothesis of no treatment effect. Results will be tabulated and may be supported by descriptive graphics such as Kaplan-Meier plots.

6.2 Case studies

For each scenario class, one simulated data set will be utilised for a case study to illustrate the application of the investigated analysis, the presentation of analysis results and to allow for an exemplary comparison of results of different methods.

For Scenario class 1 (treatment switching), the data for the case study will be simulated according to the mechanisms described in section 2.1. A setting in which average trajectories for the time-dependent covariate W are different between treatment groups will be chosen. The case study will include both administrative censoring and additional random censoring.

For Scenario class 2 (rescue medication), data according to the core data generating model in section 3.1 will be used, i.e. the data will include both rescue medication as intercurrent event and will in addition be affected by missing outcome data. The case study analysis will include, both, the treatment policy estimand as well as the hypothetical estimand strategy.

For Scenario class 3 (vaccine trial), data generated as described in section 4.1 will be used. Missing data may be included in the case study data set. As case study scenarios we will consider settings with a subgroup of subjects at higher risk for infection (scenario subtype B in subsection 4.1.6) or where the propensity for adverse reaction and the propensity for larger immunogenicity are correlated (scenario subtype C in subsection 4.1.6).

For Scenario class 4 (safety), data generated according to section 5.1 will be used, with a reduced hazard rate in the buffer window post treatment discontinuation.

The aim of the case studies is to illustrate relevant use cases. Therefore, detailed settings of the case study scenarios may be adapted based on findings on the simulation study. Additional case study scenarios may be included depending on learnings from the simulation study.

Chapter 7

Software and programming

For the core simulation functionality, the SimDesign package will be used, additional simulation software that is to be implemented is expected to include:

1. one or more modules to generate pseudo-random numbers according to scenarios with different assumptions on causal relationship and distribution of the variables;
2. several modules implementing algorithms to fit and evaluate different causal inference methods and estimators;
3. an output and presentation module that implements tabulation and plotting of simulation results.

Such a modular simulation framework has the attractive property as individual components can be implemented by different partners, permits extensive use of existing libraries and promises high reusability for potential follow-up investigations. Seamless interaction between modules is guaranteed by comprehensive specification and documentation of module interfaces. This is directly facilitated by package SimDesign, which implements dispatch, execution, and result collation. To this end, it requires specification of the simulation design (parameter scenario), data generating function (with design as input), analysis function (with generated data as input), and summary function (with analysis as input).

Compartmentalization of output and presentation into a separate module permits that corresponding software development can be deferred to a later stage in the project. For testing purposes and preliminary communication of early results between partners and EMA a rough prototype can suffice. This will free up resources in the initial stages to focus on implementation of the core functionality and permit flexible adaptation of the output to meet publication and presentation needs.

Quality Control
To ensure that software code will be intuitive to read and debug, comprehensive naming and coding conventions will be agreed between involved partners. In addition, complete interface specifications and common object and data-type models will be defined at the design stage. Software code will be extensively documented. For all high-level functions manual pages will be written (facilitated by packages roxygen and devtools). Usage of the overall simulation package will be described in a vignette.

In order to produce code that is flexible and extensible a functional programming approach - that prioritises mapping over looping - will be used. Such an approach (e.g. relying on packages provided within the tidyverse) facilitates the development of computationally efficient code that can be easily scaled on the parallel computing infrastructure available within the consortium.

Table 7.1: Software packages to be used in the simulation.

Scenario	Package for data generation	Strategy	Method	Package used for estimation	Bootstrap needed
Treatment switching in oncology setting	mvtnorm, miniPCH	Hypothetical	RPSFT	rpsftm	No
			Two stage estimation	survival, trtswitch	No
			IPCW	ipw	No
			g-formula	gfoRmula	Yes
			Cox-model	survival	
Rescue medication in diabetes trial	mvtnorm	Treatment policy and hypothetical	IPW	ipw	No
			de-mediation	glm and lm	No
			g-computation	gfoRmula	Yes
			mmrm	mmrm	No
Preventive vaccine efficacy setting	mvtnorm	Principal stratum	IV-regression	ivreg	No
Safety study with time to event endpoint	mvtnorm, miniPCH	While on treatment	Cox model and inverse probability weighting	survival and ipw	No

To ensure timely detection and correction of implementation errors, a comprehensive unit testing framework will be implemented (e.g. using R package testthat). Test cases will be prospectively planned and implemented independently from corresponding software modules. At each development iteration, results from data generating processes and analysis methods will be automatically checked against predefined test cases with known outcomes. In addition, outputs from data generating procedures and analysis results will be routinely checked visually and using summary statistics.

Software packages used
Table 7.1 lists the software packages used for the data generating mechanisms and estimation for each scenario.

If proposed packages have too slow performance to feasibly include them in the simulation study, give invalid results due to implementation error or fail for a relevant proportion of simulation replications, the following steps will be undertaken:

1. Search for another R package implementing the same estimation method.

2. Investigate the feasibility of implementing the same estimation method with standard packages.

3. Leave out time-consuming estimations like bootstrap CIs, or calculations where errors or invalid results occur and drop the methods from performance metrics where. those calculations are needed

4. Entirely drop the estimation method from the evaluation.

Any such changes will be documented in the final study report.

Simulation study
Several measures to ensure safe and reliable execution of simulation studies will be implemented. Errors and warning conditions (e.g. to detect convergence failures) will be implemented and tracked. In addition, failsafe conditions (e.g. to terminate execution in case of overly long runtimes) will be implemented. Filename conventions will be specified to avoid accidental overwriting of existing files. To minimise the impact of network errors, power outages or other hardware failures. Intermediate results will be saved to the hard disk and functions implemented that permit continuation of computation once the hardware failure has been resolved.

Reproducibility of simulations will be ensured by thorough tracking of seeds, software versions and hardware configurations as implemented in the SimDesign package. Simulation study results will be checked against results from previous related studies previously conducted by members of the Consortium. Results from novel scenarios will be checked for plausibility visually and using summary statistics. Simulation study reports will be reviewed by internal reviewers preferably from a partner in the Consortium not involved in the implementation and execution of the study.

Bibliography

(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.

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.

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.

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

Gilbert, P. B., Bosch, R. J., and Hudgens, M. G. (2003). Sensitivity analysis for the assessment of causal vaccine effects on viral load in hiv vaccine trials. *Biometrics*, 59(3):531–541.

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.

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*.

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.

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.

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

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.