Stockholms
universitet

The Classic Cox or the Counterfactual Potential Outcome? A Simulation Study of Treatment Effects in Survival Analysis

Ingunn Lilja Bergsdóttir

Masteruppsats i matematisk statistik
Master Thesis in Mathematical Statistics

Masteruppsats 2026:3
Matematisk statistik
Februari 2026

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Mathematical Statistics
Stockholm University
Master Thesis **2026:3**
http://www.math.su.se

The Classic Cox or the Counterfactual Potential Outcome? A Simulation Study of Treatment Effects in Survival Analysis

Ingunn Lilja Bergsdóttir*

February 2026

Abstract

This thesis explores and compares the Cox regression model, a classical epidemiological aaproach, and the modeling of potential outcomes using causal survival analysis (counterfactual approach). Both frameworks are used to estimate treatment effects in survival time analysis.

We present a common theoretical foundation for survival analysis and detail the methods and assumptions of the classical Cox model and the counterfactual framework. We compare and analyze the strengths and weaknesses of each procedure through a simulated observational study.

A core difference between the two is the assumption of proportional hazards: the Cox regression model requires it, whereas counterfactual models do not.

We apply the comparison to the Cox regression model, a discrete-time inverse probability-weighted pooled logistic regression model for a marginal treatment effect, and a standardized conditional pooled logistic regression outcome model. We draw results based on the estimated treatment effects, given a conditional or marginal estimate target, and compare to a hybrid between the two approaches, an inverse probability-weighted Cox model.

Finally, a visual comparison is drawn based on an estimated treatment effect derived from the chosen model parameters, regardless of the approach. Results show that each approach outperforms the other in its own setup for estimating a conditional or a marginal treatment effect.

*Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden. E-mail: ingunnlilja26@gmail.com. Supervisor: Johannes Heiny.

Contents

Preface

Acknowledgements

First and foremost I thank my supervisors Johannes Heiny, Pär Karlsson and Marie Linder for all their guidance and support. My friend Heiðrún Dís for help with coding and running the large simulations, and my friend Elísabet for thoughtful feedback and endless encouragement throughout this project. My parents for their support and my siblings for almost always answering my calls. Lastly I want to thank Ting-Yuan for always being there for me.

Use of AI Tools

AI tools were used for spelling and grammar checks, debugging in R/LaTeX and visualization.

Code availability

The simulation loops and main functions created for this work are available at https://github.com/ingunnlilja/degree_project_ilb_2026. The repository also contains the R markdown file used to create this document and all code for plots and summarized results.

Code is written in R and ran on R version 4.1.1.

1 Introduction

In statistics for epidemiology, a distinction is made between causal and classical inference in the analysis of survival time for treatments, where both aim to draw causal conclusions about the treatment effects. A popular approach to causal inference uses a counterfactual setting, asking: what if for all individuals we could both observe the factual outcome and the potential or counterfactual outcome? If there is a difference between the marginal or population-wise average of two potential outcomes, a causal effect of treatment is assumed to be present. This would then ideally be estimated as a function of time to show how the treatment effect evolves over time.

Classical inference compares observed survival times under study-specific treatment, which is then compared to a selected control treatment. The Centre for Pharmacoepidemiology at Karolinska Institutet applies classical inference when studying survival time. This is in practice done by an observational study for survival analysis using Cox proportional hazards regression, yielding a single fixed hazard ratio. A resulting hazard ratio of 1 is interpreted as indicating no difference between the treated and control groups. A hazard ratio lower than one means the rate of events is slower in the treated group, and a hazard ratio greater than 1 indicates a faster rate of events in the treated group. Based on this, in this work we will focus on observational studies, where treatment is observed but not assigned.

A common complication of classical and causal inference is censoring of observations. Censoring means that the observation time ends before the event is observed. This can, for example, occur by the end of the study or by so called competing events. A competing event is something that could happen before the event of interest. Both classical and counterfactual approaches aim to find ways to adjust for the censoring as if it did not exist. A related concept and another common complication is selection bias, most commonly referring to a skewed estimate by non-random selection of participants at the start or over time, possibly caused by censoring, into the study.

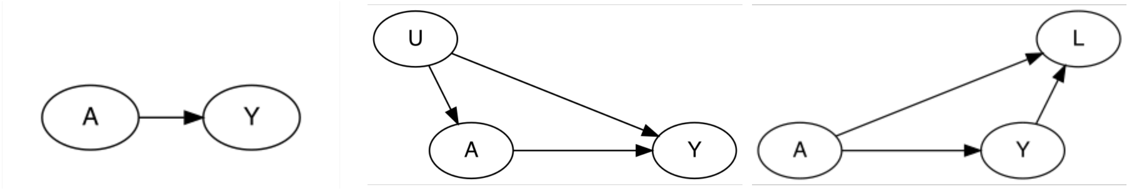
In practice, it is unclear whether the difference between the two approaches is significant enough to be of importance and whether the resulting estimates would not be approximately the same. The counterfactual approach favors marginal estimates but in many cases, a conditional one might be necessary, thus making the approaches closer to each other. Both approaches have their strengths and weaknesses, and the choice of approach may depend more on the research question itself.

1.1 Research Question

The motivation of this thesis is the following question. How are the estimated relationships between drug exposure and time to event impacted when causal inference using a counterfactual approach is applied, compared to standard Cox regression? Both methods are based on the same foundation of survival analysis; therefore, before discussing the specifics, the shared aspects of the two approaches are addressed. We will demonstrate the differences between the frameworks through simulations across various scenarios. In a simulation, the true conditions are known, allowing for the assessment of biases in estimates and confidence intervals.

1.2 Directed Acyclic Graphs

In causal inference, structural relationships are often described through directed acyclic graphs, but they can also be applied in classical inference and will be used in this thesis for both approaches. Directed acyclic graphs, or DAGs, are a straightforward way to visualize assumed relationships between variables and will therefore be used throughout this work to illustrate referenced relationships. The arrows or lack of them on a graph indicate the existence or direction of the assumed relationship, and as the name implies, the graph is acyclic, meaning that no circles are formed. We will write the most commonly referred to variable names as follows. Treatment is denoted by A, defined as $A \in \{0, 1\}$: treatment indicators (1 = treated, 0 = untreated). The outcome of interest is denoted by $Y \in \{0, 1\}$: outcome indicators (1 = event, 0 = no event). For a covariate referred to as a confounder that affects if people get treatment A aswell as the outcome Y we denote it by U, and for a variable called collider where treatment and outcome affect it but not vice versa we indicate it by L. These relationships are depicted in the following DAGs.



(a) Randomized **(b)** Confounding **(c)** Collider

Figure 1: DAGs of possible setups of a study

DAG (a): A hypothetical perfect study for survival analysis. Only A affects Y, we assume a causal relationship of treatment on outcome Y if Y differs between treated and untreated.

DAG (b): variable U, for example age, has an affect on both treatment A and outcome Y. Assumptions cannot be made about the relationship between treatment A and outcome Y without adjusting for U, adjusting for a variable is then depicted by a box around it on the DAG. Otherwise a measured effect of A on Y can be the effect of U on both. In a randomized study where participants are randomly assigned to a treatment group the goal is to eliminate confounding by design, treated and untreated participants are then exchangeable without affecting the outcome. Confounding is used in epidemiology but in causal inference the phrasing is rather to have exchangeability, or not, given the existence of a confounding variable.

DAG (c): Treatment A and outcome Y both have an affect on variable L, for example income. The treated group can, on average, have a higher income, and successful treatment can also lead to increased income for participants.

2 Methods

Most of us have at some point taken or been recommended to take the drug ibuprofen to relieve pain or reduce swelling. It can be sold over the counter or in larger doses prescribed to patients by a physician. We set up a simulated observational study to investigate how it performs in reducing swelling, but first, hypothetically, we can try it ourselves on day 1 of the injury. If swelling decreases more rapidly than we anticipated without the drug, our perception would be that the drug has a positive effect on reducing swelling. Our intuition is then based on the causal counterfactual estimate of the drug's effect; we compare the observed outcome with the drug to the hypothetical outcome we would have observed without the drug. If we observe that the swelling only went down faster if we took the drug at a specific time after the swelling started, we might think that the drug only works if taken in that time window. Now our estimate is conditional; we believe the drug has a treatment effect given the time condition. We are moving closer to a classical conditional estimate of the treatment effect. Before we could have generalized and thought that since the drug works for us, it could be safe to recommend to others. Now we have a conditional estimate, and we do not know how it affects other people without knowing when other people take it. Do people in general stick to taking the drug in the right time frame? Our estimate of a positive treatment effect, given a condition that may be rare for people to follow, can mean that the marginal or population-wise treatment effect in the causal sense is close to none.

Building on this example, we can simulate a simple observational study with 50,000 individuals who, we assume with a probability of 0.5, take ibuprofen and, independently, with the same probability of 0.5, are divided into those who take other (helpful) measures or not. If we observe that people who take ibuprofen report shorter times until swelling is gone, we could conclude that, if everyone had been treated, the average time until swelling is reduced would be lower than if no one had been. From this, we conclude that ibuprofen has a treatment effect on reducing swelling times in our small population sample.

Our conclusion relies on the assumption that other factors influencing healing time are evenly distributed between the treated and untreated groups. This is a significant assumption. For instance, we might assume that someone likely to obtain ibuprofen is also likely to have used other measures to speed up healing. To keep things simple, we assume people either use other helpful measures or do not use them at all. Those who do use such measures are more likely to also take ibuprofen.

This is now an example of the study depicted in DAG (b) earlier. Ibuprofen use (A) and healing time (Y) are both affected by other measures (U).

In Section 2.3.3, we will describe in detail how we simulate treatment assignment, and in Section 2.4.5, we go over the simulation of survival times. In Section 2.4.6 we use the described methodology in the first two sections to describe the simulation of the observational ibuprofen study. For now, we continue to analyze how we estimate the treatment effect from the simulated data using both classical and causal approaches, and compare the two approaches.

2.1 Categorical Estimation of Treatment Effect

Many of us are used to seeing contingency tables that show total counts over a period and then work with probabilities $p = P(Y = 1)$, odds $O = P(Y = 1|A = 1)/P(Y = 0|A = 1)$, and odds ratios,

$$OR = \frac{P(Y = 1|A = 1)/P(Y = 0|A = 1)}{P(Y = 1|A = 0)/P(Y = 0|A = 0)}.$$

To begin, we can view the study as a binary question: who, out of the 50,000 reported cases, experienced reduced swelling within one week of injury, ignoring any potential correlation with other measures. Then we have a binary yes-no setup for a contingency table that allows us to calculate these metrics from. Our study reported the following results.

Table 1: Contingency table for events by treatment group

	By outcome		
Treatment	Event	No event	Total
No ibuprofen	8156	10609	18765
Ibuprofen	20249	10986	31235

From the table, we obtain the odds ratio of 2.4, indicating that the odds of swelling decreasing within a week are 140% higher with ibuprofen.

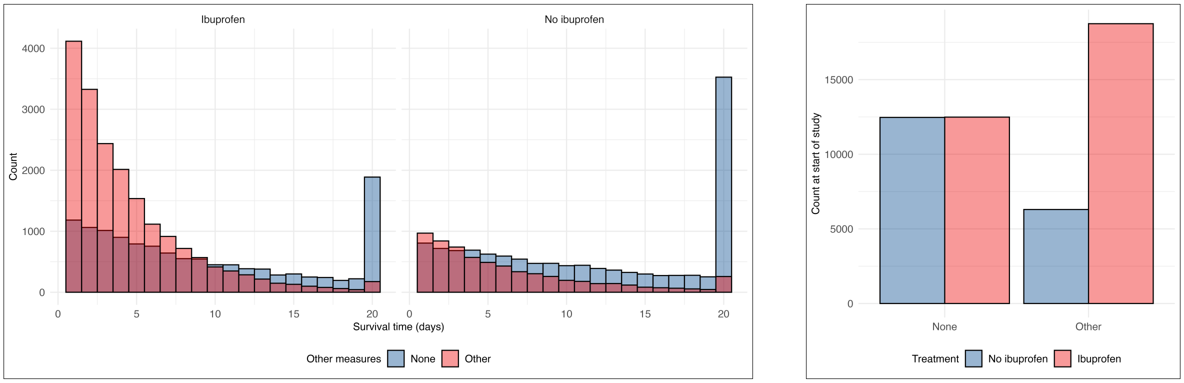
This could be a convenient viewpoint for a pharmacy to advertise ibuprofen as a great option that speeds recovery. We also knew that more factors can affect recovery time, and we have not checked how that is impacting our statement. Both from a classical and causal perspective, we are jumping to conclusions before adjusting for this bias.

Starting by splitting the sample into two groups based on whether they received other treatments or not, and finding the group-specific odds ratios as 1.71, 1.93 respectively. These odds ratios support the assumption of an effect from other treatments. Notably, we cannot collapse these two odds ratios into the marginal odds ratio from before; odds ratios are non-collapsible, as we cannot obtain the marginal odds ratio from a weighted average of the two conditional odds ratios. The two groups are split into 25,000 individuals each, and the weighted average of the two odds ratios is then 1.82, which is not equal to the marginal odds ratio from before of 2.4. Even if we did not have the bias of other measures affecting the taking of ibuprofen, the odds ratio is still, by design, non-collapsible.

To expand from the odds ratio of swelling decreasing within a week to the ratio of the rate of swelling healing per day, we need to move to survival time analysis.

2.2 Survival Time Analysis

Instead of the overall counts, we analyse the treatment effect by using each individual's reported time to event, also referred to as the survival time. Standard data exploration plots of the healing times per treatment group and other measures are a good starting point for us to get a feel for the data, regardless of the framework we are using. We continue with the simulated data, where other measures affect both taking ibuprofen and healing times, and extend this analysis to examine survival times over the first 20 days after injury, times over that are set to 20 and considered censored by the end of the study.



(a) Histograms of healing time per group

(b) Total counts per other measures

Figure 2: Healing time by ibuprofen and taking other measures

In Figure 2 we see that people who take ibuprofen report shorter healing times, and those who take other measures also report shorter healing times, which suggests that ibuprofen might be helpful in this situation. To identify the sole effect of ibuprofen, it is important to clearly define the research question and assumptions being made. For example, someone might ask their doctor how likely it is that the swelling will go down within 5 days without ibuprofen, or if it has not gone down by 10 days while using ibuprofen. A person taking ibuprofen and on day 4 of swelling may also wonder how probable it is for the swelling to heal overnight, since they are already on day 4. All three questions could be asked in terms of the overall population rather than for a specific individual.

The way the questions are answered and which ones we receive answers to differs between classical survival analysis and causal survival analysis. In this work we will refer to causal survival analysis as the counterfactual approach to draw a clear distinction since both approaches aim to draw causal conclusions about treatment effects. Our aim is to find out how and if the final answers will differ between the two approaches.

The three questions are referring to the standard survival analysis estimates of risk probabilities, survival probabilities and the hazard rate.

2.2.1 Survival, Risk and Hazard

To define these estimates we first set T as a random variable of the time to event and denote its density of probability mass function by f. At time $t \geq 0$, $f(t) = P(T = t)$.

The estimates are then defined mathematically in the following manner. The risk of event before time t is the cumulative distribution function of the event up to time t,

$$F(t) = P(T \leq t). \tag{1}$$

The survival probabilities of the event occurring after time $t \geq 0$ are,

$$S(t) = P(T > t) = 1 - F(t). \tag{2}$$

And the hazard rate, the rate of events per time interval, is defined as,

$$h(t) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} P(t \leqslant T < t + \Delta t \mid T \geq t). \tag{3}$$

Simplified to a function of the probability density function of an event at time t, and the survival function we have the relation

$$h(t) = \frac{f(t)}{1 - F(t)} = \frac{f(t)}{S(t)}. \tag{4}$$

In discrete time, the hazard rate is defined as the conditional probability of event occurring in the next instant after t, given that it has not occurred yet. In Section 1.2 we denoted the outcome by Y and treatment by A, therefore in discrete-time where we denote time points by $t = 0, 1, 2, ...$ and specify an outcome at each one as Y_t the hazard rate can be written as,

$$h(t) = P\left[Y_{t+1} = 1 \mid Y_t = 0, A\right]. \tag{5}$$

The cumulative hazard function is then, $H(t) = \int_0^t h(s)ds$ and a link to the survival function is given by, $S(t) = \exp\left\{-H(t)\right\}$. Based on this connection, the survival probabilities are often estimated using a counting process approach, where the number of events per unit time and the number at risk per unit time are used to approximate the metrics. Most notably, the nonparametric Kaplan-Meier estimator for survival probabilities is based on this counting process approach (David W. Hosmer, 2011).

2.2.1.1 Kaplan-Meier Estimator of Survival The most common classical estimator of the survival function is the nonparametric Kaplan-Meier estimator. Its goal is to estimate the survival function $S(t)$ as if censoring did not exist. The estimator only assumes that censoring is independent of all measured and unmeasured covariates, ignoring this requirement is common practice though as pointed out in (Odd Aalen, 2008) but does lead to a biased estimator if a variable is affecting the censoring mechanism.

The Kaplan-Meier estimator, also referred to as the product-limit estimator is defined as follows (Odd Aalen, 2008).

Since survival time is represented by the random variable T, we let T_i represent the survival time of time ordered individual i, on a fixed interval $[0, t_{end}]$ and $T_1 < ... < T_n$ are ordered observed survival times of n individuals in the interval. Number of events, defined as d_t if tied event times, are then counted by the counting process, $N(t)$, and each jump of N corresponds to an event time T_i. We can define the number of individuals at risk at each time t by the total number of individuals n as $R(t) = n - N(t)$. As the number of events grows, the risk set shrinks. Assuming all individuals start at the same time and experience an event, the number of individuals at risk starts at $R(0) = n$ and decreases to zero at the time point T_n, where $R(T_n) = 0$. For the interval $[0, t_{end}]$ we can break it down into equal distance ordered time points and denote each time point by k, $0 = t_0 < t_1 < ... < t_k = t_{end}$. Each interval between time points t_k, $k + 1$ has an associated number of events d_k and risk set $R(T_i)_k$.

The Kaplan-Meier estimator is defined as the conditional probability that an event occurs after time point t_k, given that it has not occurred before time t_{k-1}. Using the multiplication rule for conditional probabilities, the Kaplan-Meier estimator is calculated as the product of the conditional probabilities of the survival function at each time point t_k where the last interval K ends at time t_{end},

$$S(t) = \prod_{k \leq K} S\left(t_k \mid t_{k-1}\right). \tag{6}$$

Which in the counting process notation of number of events at time t, d_t for the individuals at risk in the risk set $R(T_i)$ becomes the Kaplan-Meier estimator defined as,

$$\widehat{S}(t) = \prod_{T_i \leq t} \left\{1 - \frac{d_t}{R\left(T_i\right)}\right\}. \tag{7}$$

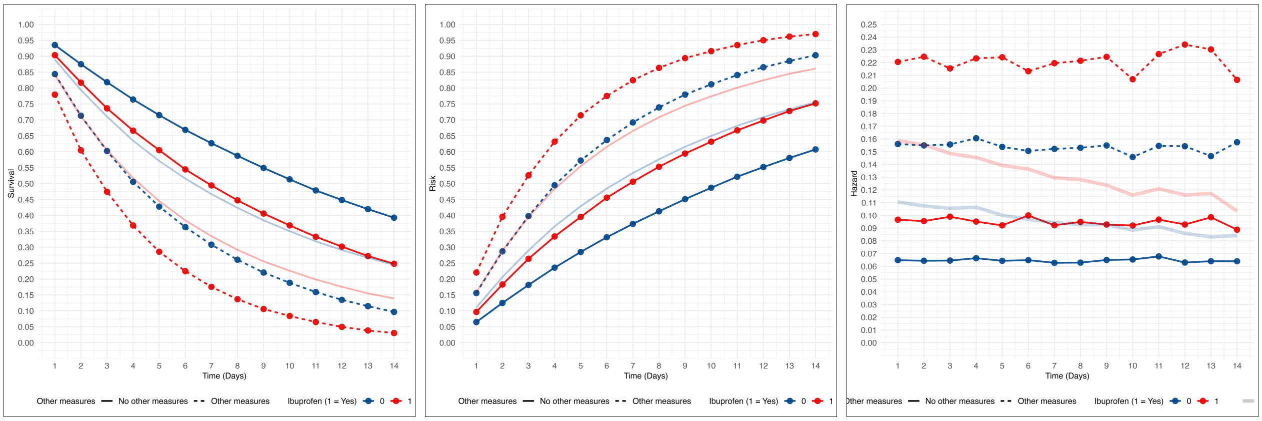
Closely related to the Kaplan-Meier estimator is the Nelson-Aalen estimator of the cumulative hazard function, $H(t)$. Defined in discrete time by the increments of the probability of event as the hazard by the ratio of events at time point to the risk set at the time point. Denoted as, $h(T_i) = \frac{d_t}{R(T_i)}$ at each time point T_i. The Nelson-Aalen estimator is then defined as the sum of these increments up to time t,,

$$\widehat{H}(t) = \sum_{T_i \leq t} \frac{d_t}{R\left(T_i\right)}. \tag{8}$$

2.2.1.2 Survival, Risk and Hazard Plots Regardless of whether a classical or counterfactual approach is used, survival, risk, and hazard are defined in the same way, see equations (7) and (8) for for survival and hazard where risk is simply one minus survival. Therefore, we can plot the results of the ibuprofen simulation study using this counting process and examine the conditional and marginal estimates for each metric, in this case with no censoring so all events in the first 14 days are observed.

The counterfactual survival times for each individual in the study are known to us, as this is a simulated study with no censoring. Therefore, we can eliminate the bias caused by other measures affecting the use of ibuprofen by using both survival times per person. Creating this type of counterfactual dataset, where counter to the fact that only some people took ibuprofen, we have data as if everyone, or no one, took it, is a core goal of the counterfactual approach.

To further describe the three metrics of survival, risk and hazard we illustrate them in Figure 3 as a function of time which is especially common practice to do for resulting estimates of the survival and risk. Creating a comparison of the survival curves for treated and untreated individuals gives a clear visual overview of a treatment effect. We plot the survival, risk, and hazards per other measures and treatment group based on the number of participants at the start of the study, the number of events observed each day and the subsequent changes to the risk set over the first 14 days. By equations (7) and (8) we estimate and plot the groups by other measures and treatment, so per metric, we have four lines. Shaded in the background are the marginal estimates per treatment group.

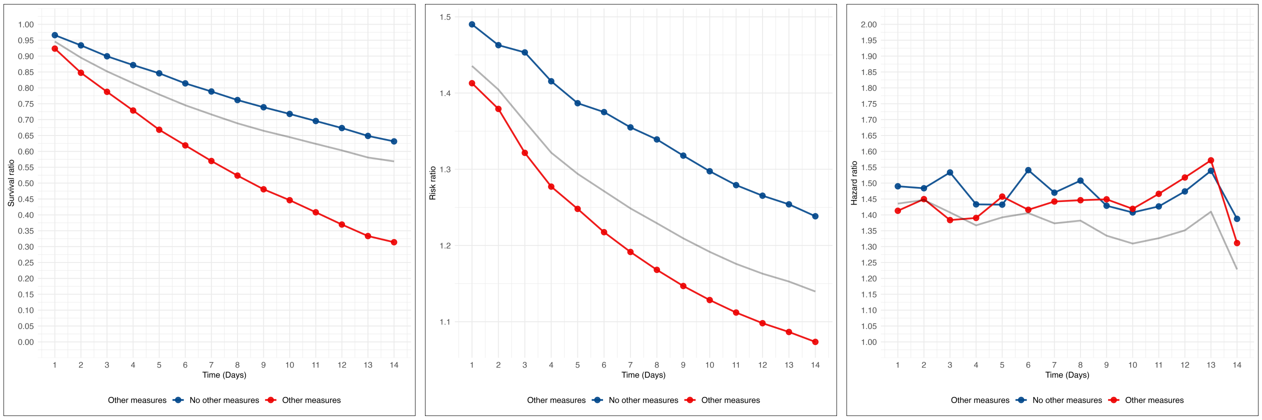


(a) Survival plot **(b)** Risk plot **(c)** Hazard plot

Figure 3: Survival, risk and hazards

The survival and risk in Figures 3 (a), (b) mirror each other since risk is simply the opposite of survival. The fastest decline in survival probability is, as expected, among individuals who take other measures and ibuprofen. The non-conditional estimates in the background fall right between the two conditional estimates. The hazard plot in Figure 3 (c) more clearly shows how the conditional estimates stay constant over time, while the marginal estimates per treatment vary.

2.2.1.3 Survival, Risk and Hazard Ratio Plots For the ratios, we plot the same groups, but now show the ratios between treated and untreated for other measures, along with the marginal estimates in the background in grey.



(a) Survival ratio plot **(b)** Risk ratio plot **(c)** Hazard ratio plot

Figure 4: Survival ratio, risk ratio and hazard ratio plots

Since the hazard ratio is dependent on the risk sets per day and not the start of study risk set it is much more unstable over time. Notably, for the survival and risk ratios we see in Figure 4 the marginal estimates fall between the conditional ratios but for the hazard ratio we see the marginal estimate go below the two conditional estimates and drop over time.

2.2.2 Classical and Counterfactual Perspectives

Comparable to the contingency table where the standard estimate was the odds ratio, in classical survival analysis the standard estimate is the ratio between hazards, the hazard ratio. A ratio under one indicates treated individuals who have not yet had an event have a lower hazard of the event happening at any time point. For the ibuprofen example that would mean treated individuals are reporting slower healing times. Ratio of one would mean that there is no difference and over one would mean that treated individuals are reporting faster healing times. The hazard ratio is like the odds ratio non-collapsible, and we can have a marginal hazard ratio as well as conditional hazard ratios but not collapse them into each other. This means that the classical framework aims to be accurate for a specific individual who goes to the pharmacy and asks on day 4 how likely it is that the swelling will heal over the coming night, given all other factors we know about that individual. The counterfactual framework prefers to aim at the first two questions as accurately as possible, given the population as a whole. As in for everyone who goes to the pharmacy the answers would be the same regardless of who they are but they would also favour stating the survival ratios and differences of the treatment regardless of the individuals framing of the question. For example someone asks how likely it is that the swelling goes down within two weeks if I take ibuprofen? The counterfactual framework could be the standardized analysis of the two week survivals and their differences in the population.

Both approaches still aim to draw a population wise causal effect of treatment conclusion.

2.2.3 Definition of a Causal Effect of Treatment

In the classical approach, the hazard ratio is the standard estimate of treatment effect; as such an $HR \neq 1$ indicates a treatment effect.

In the counterfactual approach, a causal estimate is based on a difference of potential outcomes under each treatment strategy denoted by Y^a. Counter to the factual event in observational studies that some individuals are not treated is the fictional event of all individuals in the population being treated. This is the counterfactual risk of event. For an individual i the causal effect of treatment A on outcome Y is defined by the two counterfactual outcomes of Y as, $Y_i^{a=1} \neq Y_i^{a=0}$.

In a study with n individuals $i = 1, 2, ..., n$ the causal effect of A on outcome Y is defined by the probabilities of the counterfactual outcomes. $\mathrm{P}\left[Y^{a=1} = 1\right] \neq \mathrm{P}\left[Y^{a=0} = 1\right]$. This is more commonly written as $\mathrm{E}\left[Y^{a=0}\right]$ for no one receiving treatment and $\mathrm{E}\left[Y^{a=1}\right]$ for everyone. A causal effect is then the difference between the two. If and only if $\mathrm{E}\left[Y^{a=0}\right] \neq \mathrm{E}\left[Y^{a=1}\right]$ is there a nonnull average causal effect on outcome Y. Average treatment effect is then defined as $\mathrm{E}\left[Y^{a=1}\right] - \mathrm{E}\left[Y^{a=0}\right]$.

For a causal survival effect by the counterfactual time T^a the estimate can be a comparison of the survival probabilities a function of time t as $\mathrm{P}\left[T^{a=1} > t\right] \neq \mathrm{P}\left[T^{a=0} > t\right]$ for $t = 1, 2, ... t_{\text{end}}$

Survival or risk differences based on treatment as a function of time t can provide an apparent causal effect over time. The use of ratios and hazard ratios is less common in the counterfactual framework than the classical one.

2.3 Treatment Assignment

For Swedish healthcare registers that rely solely on data without any contact with the participant, the observed treatment assignment A is determined by pharmacy prescription dispensation. The indicator for dispensation, and consequently the treatment group, is A (1: dispensed/treated, 0: undispensed/untreated). A swedish observational pharmacoepidemiology study's starting point is then defined as the date of dispensation (Humphreys et al., 2025). For simplicity, in this work we will assume a binary treatment assignment.

2.3.1 Probability of Treatment

For each study with just two treatment strategies, we can define an overall probability of treatment assignment A as a random Bernoulli variable with parameter $p = \mathrm{P}(A = 1)$. Additionally, treatment assignment can depend on covariates X, which influence whether individuals receive treatment or not.

We set the confounder U as a covariate that affects treatment assignment and, in doing so, introduces selection bias among participants in a study, since the treated group may not be representative of the overall population. Another random variable affecting treatment assignment is called an instrumental variable, I, which influences treatment assignment but does not causally affect the outcome, Y. Distance to the nearest pharmacy could serve as an example of an instrumental variable; people might be more likely to get treatment if they live close to a pharmacy. We illustrate the variables we have described in the following DAG.

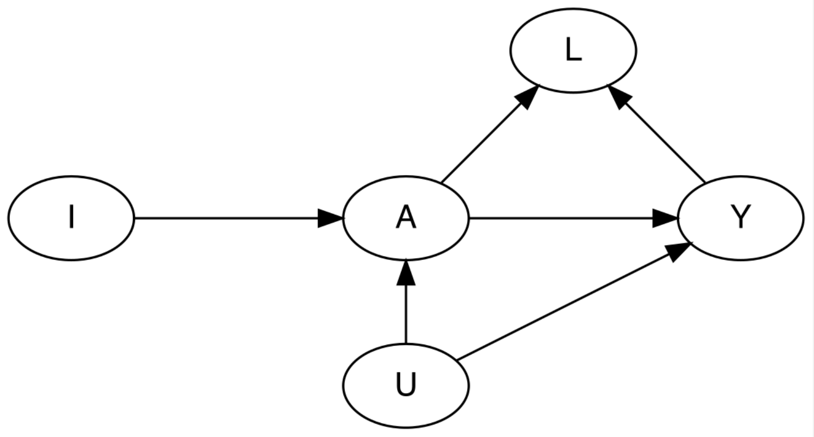


Figure 5: DAG of the described observational study

Arrows point toward A from I and U, and it is necessary to estimate the treatment assignment probabilities based on the covariates to control for confounding. Each individual's probability of receiving treatment is traditionally referred to as the individual's propensity score.

2.3.2 Propensity Scores

Propensity scores are used to assess bias based on treatment assignment and are used in both classical and counterfactual methods. The propensity scores are built on finding the conditional probability of recieving treatment given observed covariates. The scores then make it possible to adjust for confounding by using them as weights to equalize the treatment groups or to match together individuals based on their propensity scores.

This work will focus on the more common way of using them as weights, which provides the foundation for most counterfactual survival analysis work.

In 1983 Rubin and Rosenbaum published (ROSENBAUM & RUBIN, 1983) where they introduced and formalized the propensity score. The propensity score was defined as the conditional probability of assignment to a particular treatment given a vector of observed covariates. By this, given the propensity score, the treatment assignment and observed covariates are conditionally independent. They display this with the following, where x represent the covariates, a observed treatment and $e(x)$ the propensity score.

$$x \perp a \mid e(x)$$

The conditional independence was proven to hold regardless of the distribution of covariates x. If x is sufficient to adjust for confounding and selection bias, then $e(x)$ is sufficient too.

In more recent counterfactually aimed work, (Hernán & Robins, 2025), the propensity score is denoted by $\pi(x)$ and in terms of conditional probabilities as,

$$\pi(X) = \mathrm{P}[A = 1 \mid X]$$

and

$$x \perp A \mid \pi(x).$$

An ideal randomized trial would have $\pi(x) = 0.5$ for all participants and in observational studies the propensity score, $\pi(x)$ can balance the covariates between the treated and the untreated to make them conditionally independent.

2.3.3 Simulated Treatment Assignment

If treatment assignment depends on covariates it can be expressed by covariates and a baseline probability of treatment assignment via parametric logistic regression of $\pi(X) = \mathrm{P}(A = 1 \mid X)$. The log odds are modelled by α_j as coefficients of the logistic regression (Robertson et al., 2022)

$$\pi(X) = \operatorname{expit}\left(\alpha_0 + \sum_{j=1}^{p} \alpha_j X_j\right) \tag{9}$$

$$\alpha_0 = \operatorname{logit}(\mathrm{E}[A]) - \sum_{j=1}^{p} \alpha_j \mathrm{E}\left[X_j\right] \tag{10}$$

with the expit function defined as $\operatorname{expit}(x) = \frac{1}{1+\exp(-x)}$ and logit function as $\operatorname{logit}(p) = \log\left(\frac{p}{1-p}\right)$, $p \in (0, 1)$.

The marginal probability of treatment assignment is then $\mathrm{P}(A = 1) = \mathrm{E}\left[\pi(X)\right]$ and baseline probability of treatment assignment is α_0. If propensity scores are used to simulate treatment assignment, either baseline probability of treatment or the marginal probability of treatment can be used to define the per subjects treatment assignment in a study. Unless otherwise specified this work will specify the baseline probability of treatment assignment and the marginal will follow.

The relationship between, the baseline probability of treatment α_0 and the marginal $\mathrm{E}[A]$ is derived from the law of total expectation. The expected value of a random variable, A, can be expressed as the integral of its conditional expectation over the distribution of the conditioning variable, $\mathrm{E}[A] = \int \pi(x) d\mathrm{F}_X(x)$.

Notably, propensity scores only adjust for observed covariates so they do not help with finding if there is unknown confounding. Rubin recommends using past knowledge to compare if results are reasonable and secondly to run sensitivity analysis to check if unknown confounding could drastically change estimates (Agresti, 2013).

2.4 Classical Approach to Survival Time Analysis

Following the classical approach, we use individual healing times to determine the conditional adjusted hazard ratio of treatment given measured confounders and the conditional hazard ratio of the confounders.

For each measured confounding variable U, classically defined as a variable that meets the following three conditions.

(1) The confounder is associated with the treatment.

(2) The confounder is associated with the outcome conditional on the treatment.

(3) The confounder does not lie on a causal pathway between treatment and outcome. The classical approach estimates the hazard ratio of treatment conditionally on the confounding variables.

2.4.1 The Proportional Hazards Cox Model

The hazard ratios are estimated using the semiparametric regression model introduced in Section 1, known as the proportional hazards Cox model. The model assumes that the hazard rates remain proportional over time, making the hazard ratios stay constant over time.

The proportional Cox regression model uses survival event times and a set of covariates x_j to estimate the log hazard ratio β for each covariate, given the others as $\hat{\beta}_j$. It is indifferent about the baseline hazard function of the model, $h_0(t)$, meaning the intercept is left unspecified. $h(t)$ can vary over time but the hazard ratios are assumed to be independent of time.

The model is defined as,

$$h(t \mid x_j) = h_0(t) \exp\left(\beta^\top x_j\right). \tag{11}$$

Traditional residuals used for validation in regression models are based on the difference between observed and predicted values. The Cox model is semiparametric and does not assume a specific distribution for survival times, so there are no unique predicted values for each observed survival time. Residuals are still used in classical survival analysis to assess whether the proportional hazards assumption holds, particularly with Schoenfeld residuals and the Schoenfeld test.

2.4.1.1 Schoenfeld Residuals Schoenfeld residuals are defined as the difference between the observed covariate value x_i for individual i and the expected conditional mean of the covariate values at each ordered time point T_i, given their relative hazard from $\hat{\beta}$, and the risk set R with individuals $j \in R(T_i)$. In the simplest form, it is assumed there are zero tied event times and thus time points are unique per individual. This work will not delve into the details of handling tied event times for the Schoenfeld residuals; unique event times are assumed throughout. The Schoenfeld residuals r_i for each individual i are denoted as,

$$r_i = x_i - \bar{x}\left(\hat{\beta}, T_i\right),$$

where $\bar{x}$ is defined as,

$$\bar{x}\left(\hat{\beta}, T_i\right) = \frac{\sum_{j \in R(T_i)} x_j e^{x_j^\top \hat{\beta}}}{\sum_{j \in R(T_i)} e^{x_j^\top \hat{\beta}}}$$

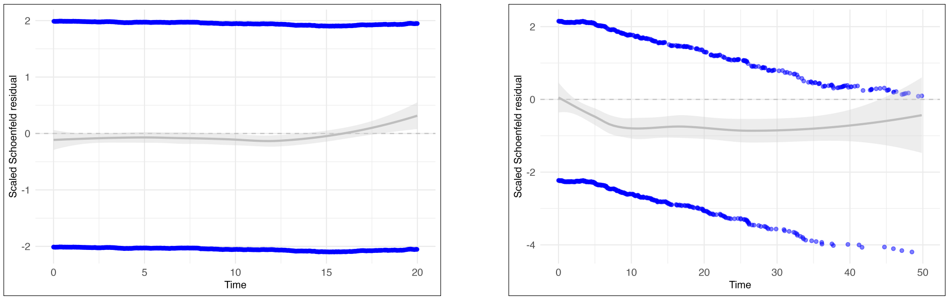
The above definition was first stated by David Schoenfeld (Schoenfeld, 1982) and subsequently extended to scaled residuals by Grambsch and Therneau (Patricia M. Grambsch, 1994). The scaled residuals, denoted by, $\hat{\mathbf{r}}_i^*$, were proven by Grambsch and Therneau to have greater diagnostic power than the unscaled ones and can be easily approximated by multiplying the estimator of the covariance matrix of estimated coefficients $\hat{\beta}$ by the number of uncensored events, m. The following is the approximation of the scaled form, which most traditional software packages use, and which will be used in this work.

$$\hat{\mathbf{r}}_i^* = \left[\widehat{\text{Var}}\left(\hat{\mathbf{r}}_i\right)\right]^{-1} = m\,\text{Var}(\hat{\beta})\hat{\mathbf{r}}_i.$$

Overview of the variance approximation can be found in, (David W. Hosmer, 2011, p. 200).

Plotting the residuals against time and visually examining the slope is a common method for assessing the proportional hazards assumption. The residuals' slope should be flat and centered around zero if the assumption is valid.

The following Figure in 6 are two examples of Schoenfeld residual plots, one with the proportional hazards assumption set to hold by simulating a constant hazard ratio and one where the assumption is violated by simulating a hazard ratio that changes over time. The Schoenfeld residuals of treatment for each of the 1000 simulated individuals are plotted against time as blue dots and a smoothed grey line is added to visualize the slope of the residuals.



(a) No slope and prop. hazards assumption holds

(b) Slope and broken prop. hazards assumption

Figure 6: Schoenfeld residuals

For Figure 6 (b), the dots are even around zero and stay in such a straight line that the specific dots become hard to see. For Figure 6 (a), the slope changes after time point 5, indicating that the proportional hazards assumption is violated.

2.4.2 Adjusted and Unadjusted Cox Model Hazard Ratios

If a confounder covariate is present in a study, the Cox model is adjusted by including the covariate, and in turn, becomes conditional on the covariate. In (Helen Bian, 2024) the authors present the following equation to summarize the two-part differences between an adjusted, (HR_{Adj}) and an unadjusted hazard ratio, (HR_{Unadj}). The first part of the equation states the differences between a conditional and a true marginal estimate, (HR_M) and the second is the confounding bias.

$$\log\left(HR_{Adj}\right) - \log\left(HR_{\text{Unadj}}\right) = \underbrace{\left[\log\left(HR_{\text{Adj}}\right) - \log\left(HR_M\right)\right]}_{\text{Non-collapsibility}} + \underbrace{\left[\log\left(HR_M\right) - \log\left(HR_{\text{Unadj}}\right)\right]}_{\text{Confounding}}$$

The non-collapsibility part of the equation is to capture the difference between an estimate given a covariate and the marginal estimate and the second part captures the confounding bias present in an unadjusted estimate, if a confounder is present. When estimating differences or bias of hazard ratios, it is like in this case common practice to work on the log scale.

2.4.3 Cumulative Hazard Function Estimation

Given an adjusted Cox model we can estimate the cumulative hazard function $\widehat{H}(t)$ given the estimated hazard ratios by weighing the risk set in the Nelson-Aalen (8) estimator by the estimated hazard ratios. This estimate is the Breslow estimator of the cumulative hazard function (Odd Aalen, 2008, p. 141), defined for covariate vector x_j as

$$\widehat{H}(t \mid x_j) = \sum_{T_i \leq t} \frac{d_{T_i}}{\sum_{j \leq R(T_i)} \exp\left(\widehat{\beta}^\top x_j\right)}. \tag{12}$$

2.4.4 Survival Function Estimation

Recall the identity $\widehat{S}(t|x_j) = \exp\left\{-\widehat{H}(t|x_j)\right\}$. Plugging in an estimator for the cumulative hazard function $\widehat{H}(t|x_j)$ yields the survival function over time based on the estimated adjusted Cox model. This is known as,

$$\mathrm{S}(t \mid x_j) = \exp\left[-\mathrm{H}_0(t)\exp\left(\widehat{\beta}^\top x_j\right)\right]$$

Based on this definition of the survival function we can generate event times for simulated studies.

2.4.5 Simulate Data for Cox Proportional Hazards Model

We use inverse transformation sampling to simulate survival times based on the estimated Cox model parameters, as outlined in (Ralf Bender, 2005). The method is also extended to simulate time-varying covariates (Austin, 2012). In this work, we will restrict ourselves to time-fixed covariates.

We set the hazard rate at time $t = 0$ to $h_0(t) = \lambda$, the hazard function is then $h(t \mid x) = \lambda \exp\left(\beta^\top x\right)$. We denote the uniform distribution on interval $[0, 1]$ by, $G = \text{Unif}[0, 1]$ to create a distinction from confounder U and use it to generate random numbers between 0 and 1. The event times T_i are simulated by, $T_i = H_0^{-1}\left[-\log(U)\exp\left(-\beta^\top x\right)\right]$ which simplifies to,

$$T_i = -\frac{\log(G)}{\lambda \exp\left(\beta^\top x\right)}. \tag{13}$$

From equation (13), we can simulate studies with known true hazard ratios for treatment and covariates and a guarantee that the proportional hazards assumption holds, given the model is correctly specified.

2.4.6 Cox Model Data Simulation of Ibuprofen Data

For the simulated ibuprofen dataset, we simulate two potential survival times for each individual based on equation (13) corresponding to being treated and untreated, we denote these by treatment $A = a$, $T_i^{a=1}$ and $T_i^{a=0}$, corresponding to simulating 100.000 survival times for $N = 50.000$ individuals. Treatment and covariate, other measures U, affect the hazard rate by the constant conditional hazard ratios of $\exp(\beta_1)$ and $\exp(\beta_2)$ respectively. The baseline hazard rate λ is set to $\lambda = 0.1$ and the parameters are set to $\beta_1 = \log(1.5)$ and $\beta_2 = \log(2.5)$. The two simulated survival times for each individual i are then,

$$T_i^a = -\frac{\log(G)}{\lambda \exp\left(\beta_1 a + \beta_2 u\right)}.$$

For each individual, we only observe one of the two survival times based on their observed treatment assignment A. In Section 2.3 we described how treatment assignment is determined by a baseline probability of treatment assignment α_0 and by any covariate j that has $\alpha_j \neq 0$. We set the baseline probability of treatment assignment to $\alpha_0 = \text{logit}(0.5)$ and the coefficient of other measures to $\alpha_1 = \log(3)$ for observed $U = u$. The treatment assignment is then simulated for each individual i as a bernoulli random variable with parameter $\pi(x)_i$ by equation (9) as,

$$\pi(X)_i = \text{expit}\left(\alpha_0 + \alpha_1 u_i\right).$$

2.4.7 Cox Model Results on Ibuprofen Simulated Observational Study

To interpret the results of the Cox model on the ibuprofen data we need to state our model assumptions. We assume that (a) no one was lost to follow-up, (b) that the requirement of proportional hazards is fulfilled given an adjusted model, the hazard ratio is therefore independent of time, and (c) that there is no interaction between treatment and other measures.

The resulting estimate is the adjusted hazard ratio of ibuprofen given taking other measures, $\exp\left(\hat{\beta}_1\right) = 1.45$ (95% CI 1.42–1.48) and the hazard ratio of taking other measures, $\exp\left(\hat{\beta}_2\right) = 2.35$ (95% CI 2.30–2.40). People who take ibuprofen report 45% higher constant instantaneous rate of recovery times than people who do not take ibuprofen, conditional on not using other measures to speed healing. For an individual who takes ibuprofen until day 4 of swelling and asks about the likelihood of healing overnight, we cannot specify the exact probability, but we can say that the instantaneous healing rate is 45% higher than if he were not taking ibuprofen. Additionally, if he also uses other measures, the hazard increases by 3.4075 compared to not taking ibuprofen and not using other measures.

Below in Figure 7, we plot the survival function over time per treatment group and other measures based on the adjusted Cox model. Since the risk is the opposite of the survival, we add the risk faintly in the background as well.

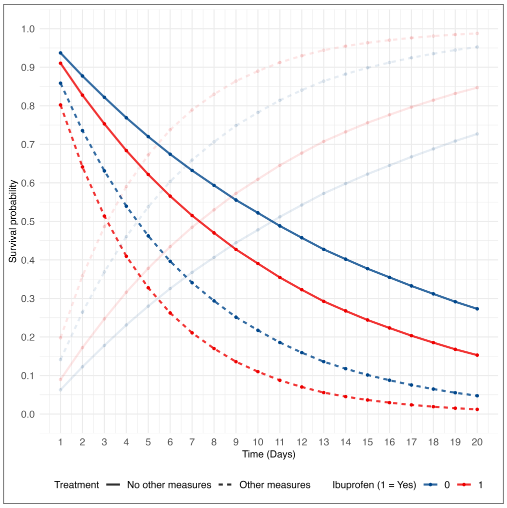


Figure 7: Survival probabilities, classical adjusted Cox. Risk faintly in background

Probability of surviving to day 10 given taking ibuprofen but not other measures is then $S(10|U = 0, A = 1) = P(T > 10) \approx 0.1$ and without both, $S(10|U = 0, A = 0) = P(T > 10) \approx 0.52$.

2.5 Counterfactual Approach

For an estimate similar to the single unconditional odds ratio in Section 2.1, we can use the individual survival times to estimate a single marginal hazard ratio HR_M and move into the causal framework. Like in the classical approach, we first need to establish the assumptions for our study and the general assumptions of the counterfactual approach.

2.5.1 Counterfactual Assumptions, or Identifiability Conditions

Validity of causal inference estimates is based on the following assumptions, also referred to as identifiability conditions (Hernán & Robins, 2025).

- Exchangeability. $Y^a \perp A \mid X$. The treated and untreated group are exchangeable. Adjusting for any confounders U leads to exchangeability and no unmeasured confounding is required. Exchangeability and no confounding are often used as interchangeable terms.

The causal definition of a confounder is based on changes to the classical definition.

(1) There exist variables X and U such that there is conditional exchangeability within their joint levels $Y^a \perp A \mid X, U$. X is then not on a causal pathway between A and Y, and $\mathrm{E}\left[Y^a \mid X = x, U = u\right]$ is identified by $\mathrm{E}[Y \mid X = x, U = u, A = a]$.

(2) U can be decomposed into two disjoint subsets U_1 and U_2 such that (i) U_1 and A are not associated within strata of X, and (ii) U_2 and Y are not associated within joint strata of A, X, and U_1. The variables in U_1 may be associated with the variables in U_2. U_1 can always be chosen to be the largest subset of U that is unassociated with treatment (Robins, 1997).

- Positivity. $\mathrm{P}[A = a \mid X = x] > 0$ for all values x with $\mathrm{P}[X = x] \neq 0$. No study participants can have a zero probability of receiving either treatment level, given their covariates. If a subgroup never receives treatment, no causal effect can be estimated in the subgroup.
- Consistency. $\mathrm{P}[Y^a = Y \mid A = a, X = x]$. Treatment is well defined. The potential outcome Y^a is the same as the observed outcome if the treatment received is the same as the treatment defined. Each participant has two potential outcomes. One becomes observed and the other counterfactual.

2.5.2 The Three Counterfactual g-methods

Causal inference has three primary methods to determine causal effects; IP-weighting of marginal structural models, the parametric g-formula for standardization, and g-estimation of structural nested models. The g simply stands for generalized. Their combined aim is to estimate the causal effect of treatment A on outcome Y for a population or a subset of it.

This work will focus on the most commonly used, first two approaches within the counterfactual framework, the MSM models and the parametric g-formula. The first part of this thesis will cover MSM models before moving on to the g-formula in the second part. For a brief description of g-estimation of structural nested models, see appendix.

2.5.3 Inverse Probability of Treatment Weights and Marginal Structural Models

To determine a marginal causal effect we use a regression model known as a marginal structural model, which is used for the marginal counterfactual potential outcome means of $\mathrm{E}\left[Y^a\right]$. Given a dichotomous treatment, the counterfactual potential outcomes are modeled by first reaching exchangeability by constructing a pseudo-population, where individuals are weighted by their modelled propensity scores or probability of receiving treatment, before estimating the outcome via an MSM model to control for confounding. The weights are referred to as inverse probability of treatment weights, IPTW (Williamson & Ravani, 2017).

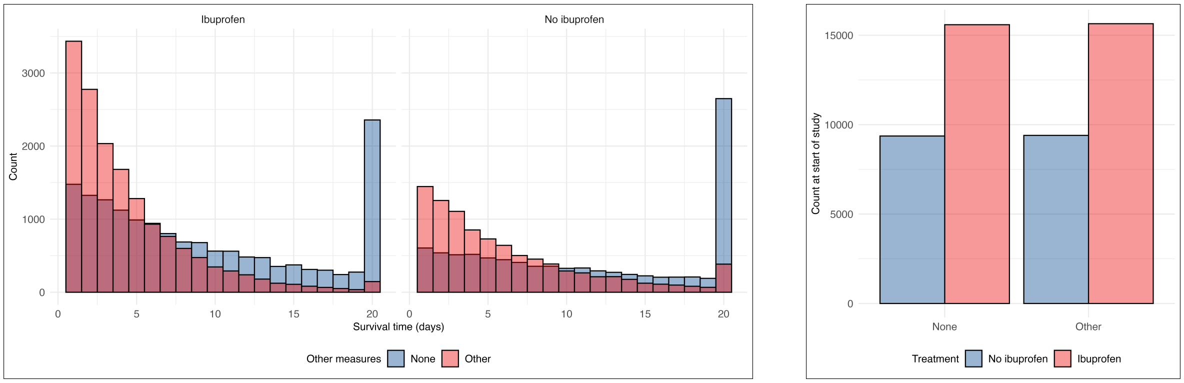
2.5.3.1 Inverse Probability of Treatment Weights The weights are defined as the inverse probability, or propensity scores, of receiving observed treatment $A = a$. To stabilize the weights their numerator is set to the marginal probability of treatment $\mathrm{P}[A = 1]$. The goal of the weights is to use them as the conditional probability of treatment given the confounding covariates $\mathrm{P}[A \mid U]$. The weights are sensitive to extreme values and outliers, which are commonly addressed by setting a threshold for the weights, such as truncating them at the 99th and 1st percentile (Williamson & Ravani, 2017). In functional form, the stabilized weights are defined as,

$$\mathrm{IPTW} = \frac{f(A = 1)}{f(A = a \mid U)}. \tag{14}$$

The conditional probability density function of treatment given confounders $f(A \mid U)$ is estimated by fitting a logistic regression model of treatment assignment. The predicted probabilities from the treatment model are then used to calculate the weights for each individual in the study.

Individuals who received treatment, unlikely given their confounders are assigned a higher weight and vice versa. This creates a pseudo-population where treatment is independent of confounders, making treated and untreated individuals exchangeable, (Cole & Hernán, 2008).

To illustrate the effect of weighting data by IP-weights we replot the survival time distributed by treatment group and other measures in figure 2 but now each individual survival time is weighed by their IP-weight.



(a) Histograms of healing time per group

(b) Total counts per other measures

Figure 8: IP-weighted healing times by ibuprofen and taking other measures

Figure 8 (a) now shows a more even distribution of survival times across the treatment groups compared to Figure 2 (a). IP-weights have balanced the confounding effect of other measures. Since our data is simulated, we know the true marginal and conditional probabilities of treatment. Therefore, for this example, we are using the true weights instead of the estimated weights from a treatment model.

2.5.3.2 Marginal Structural Models Counterfactual estimates have the objective of $\mathrm{E}\left[Y^{a=1}\right]-\mathrm{E}\left[Y^{a=0}\right]$. MSMs model this by

$$E[Y^a]=\beta_0+\beta_{1a}.$$

This refers to the baseline probability of outcome $P[Y^a=1|a=0]=\beta_0$ with a linear addition of the marginal causal effect of treatment in the population as β_1 (Robins et al., 2000).

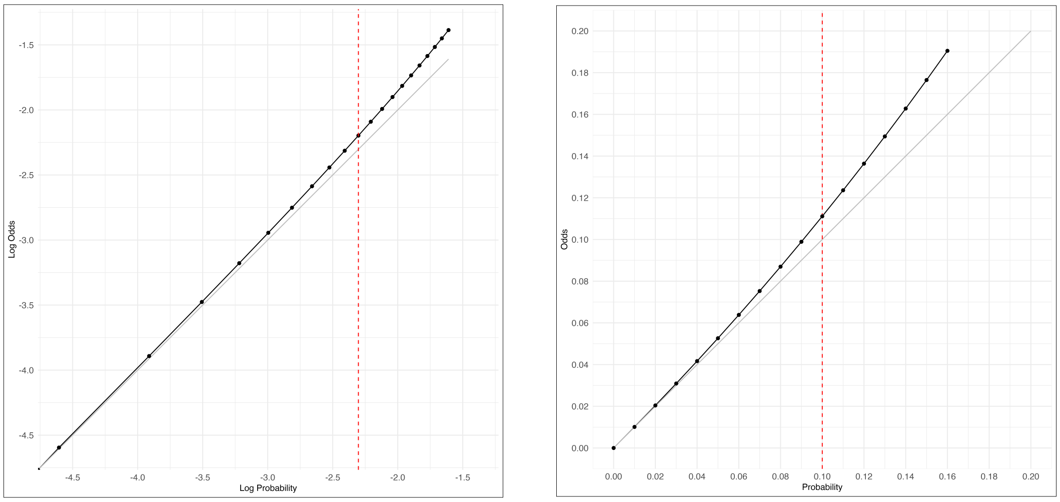
A strength of MSMs is that given correctly specified IP-weights, they have the property of null preservation, if the true causal effect in the population is null, the outcome model can be misspecified and still return a null causal effect (Hernán & Robins, 2025). This is not the case for Cox models or the parametric g-formula where misspecification of the outcome model can lead to biased estimates even if the true causal effect is null.

2.5.3.3 Counterfactual Hazard Ratio by a Logistic MSM Now if the goal is to estimate the causal hazard ratio we set the MSM as a logistic regression model for $\mathrm{E}\left[Y_t^{a=1}\right]/\mathrm{E}\left[Y_t^{a=0}\right]$,

$$\log\frac{P\left[Y_{t+1}=1\mid Y_t=0,A\right]}{P\left[Y_{t+1}=0\mid Y_t=0,A\right]}=\operatorname{logit}P\left[Y_{t+1}=1\mid Y_t=0,A\right]=\beta_0+\beta_1a+\beta_2t. \tag{15}$$

The causal hazard ratio of treatment is then $\exp(\beta_1)$, the baseline hazard rate is $\exp(\beta_0)$ and the hazard ratio for each unit increase in time is $\exp(\beta_2)$. The inclusion of time is based on the presumption that hazard ratios are time varying. Modeling the hazard ratios by a logistic regression is based on the assumption that the odds are approximately equal to the probabilities and thus the hazard given rare events. The time unit t, f.x. hours, weeks, days, is required to be short enough for $P\left[Y_{t+1}=1\mid Y_t=0,A\right]<0.1$ to hold, ensuring the odds–hazard approximation (Hernán & Robins, 2025, p. 227).

To visualize the approximation required we plot the relationship of odds $=\frac{p}{1-p}$ between probabilities, odds, and their respective logs. To do this we simulate a sequence of probabilities from 0 to 0.2 with 0.01 increments. For each probability we find the corresponding odds and plot the results, since the logistic model works on the log scale we also display the logs. The cutoff for the approximation is added as a vertical red line at $p=0.1$.



(a) Log of prob. and odds **(b)** Odds as a function of probability

Figure 9: Probability approximation by odds

The closer to 0.1 from zero the probability gets in figure 9, the further the odds and probability diverge.

This estimation of the hazard ratio in causal inference is not typically used or advised in classical survival analysis. Notes on this are found in, (David W. Hosmer, 2011, p. 116).

Looking back at figure 3, we see that the hazard rates are too high for our unit time interval of days to meet the approximation requirement. The next logical step is to shorten the time interval to hours instead of days and recalculate the survival times accordingly. A complication is that when that is done, we will quickly run into computational limitations since the dataset size grows drastically. To explain this, we need to define the type of data structure used for survival analysis within the counterfactual framework.

2.5.3.4 Choice of Data Structure for Counterfactual Survival Analysis Classically, study data is one row per subject given one observation. When we created the counting process, we went from one row per subject to one row per subject at each time point they are in the study. We will refer to this as person-time. For example, if the time unit of study is days and a participant is censored at day 10, the subject will have 10 rows in the dataset. Expansion from this format is then each row representing a unique time point for each subject and treatment strategy, both before and after they experience an event. If study time is 100 days and treatment is dichotomous, all subjects will have 200 rows in the dataset. This data structure will be referred to as person-time expanded. A similar but more simple data structure is study-time expanded. Here, each row represents a unique time point for each treatment strategy, regardless of the number of subjects in the study. A study over 100 days with dichotomous treatment will have 200 rows in total.

Predictions can then be made on the expanded datasets to compare predictions as the counterfactual event of everyone treated versus everyone untreated.

Considering our example of an ibuprofen study with 50,000 participants observed over 20 days, in the person-time format, each individual can have up to 20 rows, allowing us to potentially expand to 1,000,000 rows. This is still manageable, but switching the time unit to hours instead of days would produce a dataset with up to 24,000,000 rows. Such a large dataset can quickly become computationally difficult due to the study size and total observation period. Therefore, it is important to minimize data expansion while ensuring the rare event assumption for the odds-hazard approximation is met.

In the ibuprofen example, switching to 12-hour intervals instead of daily ones is sufficient to satisfy the rare event assumption for the odds-hazard approximation. Replotting the hazard rates per treatment group and other measures with the new time unit shows that the marginal hazard rates remain consistently below 0.1.

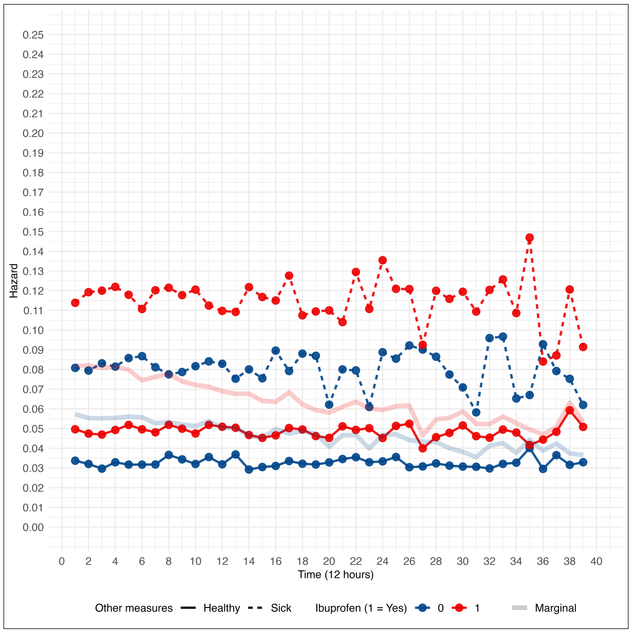


Figure 10: Hazard rate per 12 hour intervals

According to Figure 10, the approximation stays valid as long as the marginal estimate is the target.

2.5.3.5 IP-weighted Pooled Logistic Regression Models as Marginal Structural Models

Given the new data structure of person time the logistic MSM is fit as a pooled logistic regression model. Pooled refers to the fact that the data is in the person-time format for each subjects time in study, across multiple time intervals and fit as a single logistic regression model. This introduces correlated clusters for each individual. The within subject correlation makes standard error estimates inaccurate, to account for this cluster-robust standard errors are used when estimating the model parameters with each individual as a cluster. The resulting robust standard errors are guaranteed to provide a asymptotically 95% confidence intervals, covering the true β at least

95% of the time in large samples, (Hernan et al., 2000). A rigorous definition of the cluster-robust standard errors can be found in (Li & Redden, 2015).

For the ibuprofen study we fit a IP-weighted pooled logistic regression model to estimate the discrete-time $OR \approx$ hazard ratio in each 12 hour interval as,

$$\operatorname{logit} P\left[Y_{t+1} = 1 \mid Y_t = 0, A, t\right] = \beta_0 + \beta_1 a + \beta_2 t + \beta_3 t a. \tag{16}$$

The resulting HR estimates are, $\exp(\beta_1) = 1.496$ (95% CI 1.447–1.548) and the average change in hazard for each increase in time is, $\exp(\beta_2) = 0.989$ (95% CI 0.987–0.990). The baseline hazard at time t_0 is then, $\exp(\beta_0) = 0.061$ (95% CI 0.059–0.063).

The nonparametric Kaplan-Meier estimator of survival over time has a parametric counterpart for the IP-weighted pooled logistic regression model. Given study-time expanded data we can predict the hazard of event in each time interval t for everyone treated and everyone untreated. Since survival at time t is the product of surviving each prior time interval, we can estimate the survival function over time as the product of one minus the predicted hazard in each unit time interval up to time t, (Hernán & Robins, 2025, p. 224). In the Kaplan-Meier notation this is defined as,

$$\widehat{S}(t) = \prod_{m \leq t} \left\{1 - \hat{h}(m)\right\}. \tag{17}$$

We can also express the survival function in terms of probabilities,

$$\mathrm{P}\left[Y_t = 0\right] = \prod_{m \leq t} \mathrm{P}\left[Y_m = 0 \mid Y_{m-1} = 0\right]. \tag{18}$$

Marginal survival probabilities over time can thus be estimated from the IP-weighted pooled logistic regression model by predicting the hazard at each time interval under everyone treated and everyone untreated and calculating the survival function by equation (17) or (18).

2.5.3.6 Effect Modification and Conditional Marginal Structural Models MSMs include parameters of treatment A, any variables that interact with treatment, and, based on the research question at hand, variables that only affect outcome to improve model prediction. Variables that interact with treatment are called effect modifiers, as they modify the effect of the treatment. We will refer to effect modifiers with V. The choice of V does not worsen the accuracy of the marginal estimates, but it does make them conditional on V, as the numerator of the IP-weights is conditioned on V. The numerator gets set to $\mathrm{P}[A = 1 \mid V]$, so the more variables included, the closer the estimate gets to a classical conditional estimate. The MSM, including effect modification by V is then denoted as, $\mathrm{E}\left[Y^a \mid V\right] = \beta_0 + \beta_1 a + \beta_2 a V + \beta_3 V$.

It is not possible to demonstrate effect modification using a DAG, as DAGs display associations, not interactions. This is a fallback of the DAGs representation.

2.5.3.7 Faux Marginal Structural Models Marginal structural models that condition on all covariates X to adjust for confounding are referred to as Faux Marginal Structural Models. As stated, examining the effect modification by covariates can be achieved by conditioning the numerator on V, $f[A \mid V]$. Stabilized IP-weights are defined as, $IPTW(V) = \frac{f[A=1|V]}{f[A|X]}$. It is clear that if $V = X$ then $IPTW(V) = 1$. Thus, IP-weights equal 1 when the numerator and denominator are equal, which is the case when V contains all covariates X. This is the case for Faux Marginal Structural Models.

This helps us understand the difference between the classical approach and the counterfactual approach. The closer we get to the conditional nature of the classical approach, the less difference there is between the two approaches. The counterfactual approach to find IP-weights for adjusting confounding in a Cox model approaches the classical Cox model as the number of covariates conditioned on by the choice of $V \in X$ for the numerator increases.

It is common practice in causal survival analysis to condition the MSM on baseline covariates to improve model prediction, this could be gender or nationality for example. Since this makes the estimate conditional on these covariates and thus closer to a classical estimate we do not explore this further in this work.

2.5.4 IP-weighted Cox Model as an MSM

IP-weights can be used with the proportional hazards Cox model to adjust for confounding by IP-weighting the risk sets, instead of including the confounder in the outcome model.

Marginal structural models, model the marginal counterfactual potential means, which, if weighted, the Cox model does as well. Therefore, an IP-weighted Cox model is a marginal structural model for survival data, (Hernan et al., 2000). The marginal IP-weighted Cox model can be represented as, $h(t) = h_0(t) \exp(\beta_1 a)$.

The causal interpretation is then that $\exp(\beta_1)$ is the ratio of the hazard rate at time t if everyone in the population were treated to the hazard rate at time t if no one in the population were treated. As with the classical Cox model, the hazard ratio is constant over time.

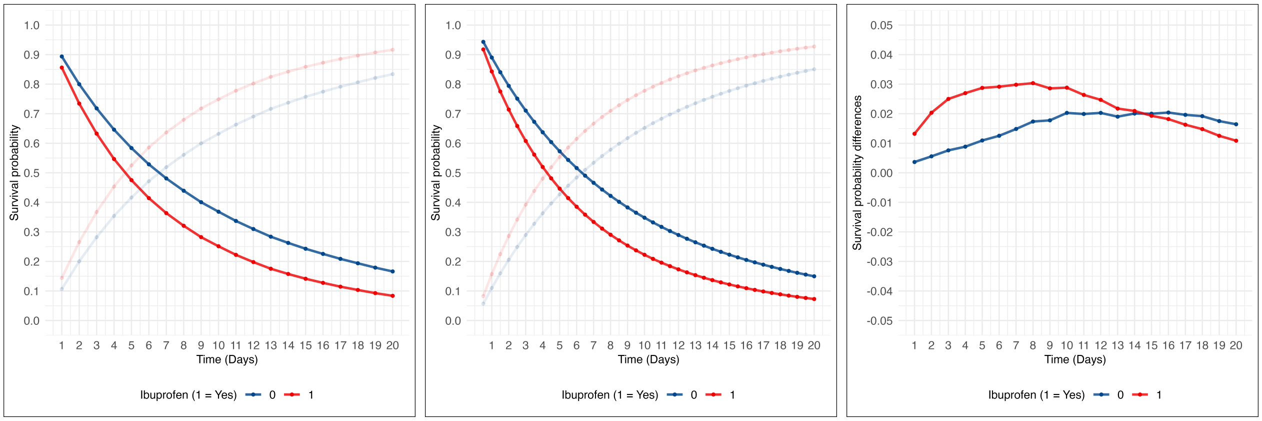
Similar to the pooled logistic model, standard errors for Cox models are not reliable with IP-weights. Robust-cluster (sandwich) standard errors are used to account for the per person variability introduced by weighting the data by IP-weights. The definition and formula for the robust standard errors are not straightforward and in general left out of counterfactual literature. The importance of using them is stressed in (Hernan et al., 2000) and a detailed mathematical background can be found in (Wang et al., 2023).

We refit the Cox model, now unconditional on the IP-weighted pseudo-population, and find that taking ibuprofen increases by, on average, $= 1.38$ (95% CI 1.36–1.41) the instantaneous rate of recovery, regardless of other measures.

2.5.5 Marginal Causal Survival and Risk Probabilities

The IP-weighted Cox model can be used to plot the marginal survival and risk probabilities over time as we did with the adjusted Cox model.

We plot the survival and risk from the IP-weighted logistic regression model and compare them to the IP-weighted Cox model survival estimates by illustrating their differences over time.



(a) IPTW-Cox MSM

(b) IPTW-GLM MSM

(c) Survival differences between the first two

Figure 11: Survival probabilities

Figures 11 tell us that given both models, the probability of surviving to day 10 is $S(10) = P(T > 10) \approx 0.25$ with ibuprofen and $S(10) = P(T > 10) \approx 0.35$ without ibuprofen. The estimate is slightly lower for the IP-weighted MSM pooled logistic regression model, but the difference is minimal.

The comparison of the two models shows that they provide very similar estimates of survival probabilities over time.

2.6 Proportional Hazards Assumptions

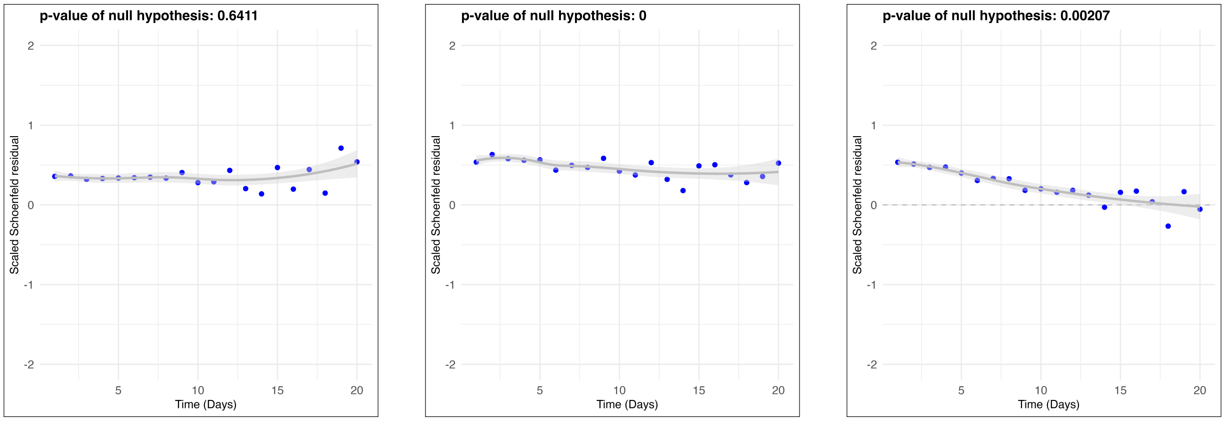
Returning to the assumptions we set for the study, we confirmed, in Section 2.4.5, that the proportional hazards assumption was fulfilled, given the conditional model. What about the marginal model? We jumped to the causal framework and set up the marginal Cox model without testing if the assumption of proportional hazards was fulfilled there as well. In fact, that is rarely the case; like the odds ratio, the hazard ratio is non-collapsible and thus if the conditional hazard ratio is constant over time, the marginal hazard ratio is not guaranteed to be constant over time as well. The classical test for checking if the proportional hazards assumption holds is based on the scaled Schoenfeld residuals; if they are statistically significantly correlated with time, the assumption is violated. That is, in fact, the case for the marginal estimate.

2.6.1 Schoenfeld Residual Test for Proportional Hazards

The proportional hazards assumption is tested by a hypothesis test called the Schoenfeld test, which evaluates the null hypothesis of zero slope for each covariate for significance. A rigorous mathematical background for the Schoenfeld test of a non-zero slope can be found in (Patricia M. Grambsch, 1994).

The Schoenfeld test assesses the null hypothesis of proportional hazards. It is based on checking whether there is a zero-slope in the Schoenfeld residuals over time using a chi-squared test of null hypothesis. If a significant slope is found at the 5% level, it indicates that the covariate's effect on the hazard rate varies over time, violating the proportional hazards assumption. The test is performed both globally for the entire model and for each covariate.

We apply the Schoenfeld test to the three discussed Cox models: adjusted, unadjusted, and IP-weighted marginal Cox models. Below are the plots of the Schoenfeld residuals over time for each model and their respective p-values.



(a) Adjusted Cox model **(b)** Unadjusted Cox model **(c)** IP-weighted marginal Cox model

Figure 12: Test of proportional hazard assumptions for three different Cox models on same data

Based on the Schoenfeld residual test in Figure 12, the only Cox model that meets the proportional hazards assumption is the adjusted Cox model. Both marginal models violate the assumption.

2.6.2 Proportional Hazards Assumption under the Causal Framework

From a classical perspective, the marginal hazard ratio is now inaccurate because the assumption is violated. However, from a causal perspective, the assumption holds too rarely in practice to be useful or meaningful. The Schoenfeld test is stated to lack enough power to detect violations, and for both small and large datasets, it can reject the null hypothesis too easily (Mats J. Stensrud, 2020). From a causal perspective the assumption is that the hazard ratio varies over time and should be modelled accordingly. Authors of (Mats J. Stensrud, 2020) follow up with (Stensrud & Hernán, 2025) where they argue that methods not requiring the proportional hazards assumption should be used instead. One of the authors of both papers, Mats J. Stensrud, also wrote (Dumas

& Stensrud, 2025), pointing out how the use of hazard ratios for causal inference can generally be misleading.

In our ibuprofen example, the estimated constant marginal hazard ratio is closer to the counterfactual aim of the unconditional ibuprofen treatment effect in the entire population than the conditional estimate. However, it can be regarded as an average of a time-varying hazard ratio, based on the causal assumption of time-varying hazard ratios.

2.6.3 Hazards, Hazard Ratios and Risk Set Over Time

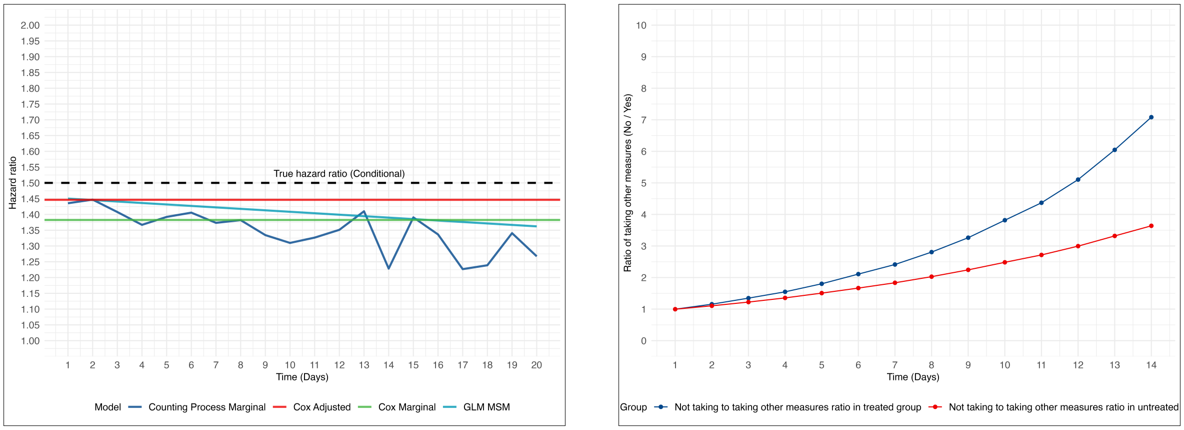
For survival and risk probabilities at specific time points t_k, we assess the probability of an event happening within the time interval from the start of study to time t_k.

The hazard rate and hazard ratio at timepoint t_k are on the other hand conditional on surviving up to the timepoint.

This bias, which occurs when conditioning on survival, is known in the counterfactual framework as survivor bias, healthy worker bias, or, in the context of the Cox model, the built-in selection bias inherent to the Cox model.

In the ibuprofen example, to adjust for other factors influencing ibuprofen intake, we both weighted individuals by their probability of taking the drug and conditioned on other measures, both variables based on all individuals at the start of the study. The hazard ratio is derived from hazards, or the rate at which events occur per day, which is conditional on the set of people still at risk on that specific day, known as the risk set. Essentially, the risk set on day 1 can look quite different from the risk set on day 20.

To visualize this, we plot in Figure 13 the marginal hazard ratio and risk set for each day. This is possible because we have all the individuals observed and counterfactual healing times. By design, we have adjusted for other factors that influence the use of ibuprofen and can determine the marginal hazard ratio over time by examining, each day the number of individuals in the risk set and the number of events that occur. To compare the true counting process hazard ratio, with the classical or counterfactual framework estimates, we plot the estimates as additional lines in the true hazard ratio graph.



(a) Ibuprofen effect over time

(b) Changes in risk set over time

Figure 13: Changes and constants over the first 14 days

The marginal estimate in Figure 13 decreases over time as the proportion of people taking only ibuprofen increases. The adjusted estimate remains the most technically accurate reflection of the true constant hazard ratio at the start of the study. Since the hazard ratio varies significantly over time as the risk sets shrink, repeated studies are needed to plot a stable line. In the appendix, we have added a Monte Carlo simulation of the marginal hazard ratio over time, based on 100 simulated studies of the ibuprofen study.

2.6.4 HR Over Time Given Various Treatment and Confounding Effects

The ibuprofen study is simulated with a true conditional hazard ratio, dependant on, other measures as, $\exp(\beta_1) = 1.5$ and the hazard ratio of the, other measures covariate is $\exp(\beta_2) = 2.5$. Half of the individuals are set to take other measures and the baseline probability of taking ibuprofen is set to $\alpha_0 = 0.5$. The odds of taking ibuprofen, given taking other measures, are set to 3 to introduce confounding.

This simulated study is straightforward and chosen to illustrate how the risk set changes over time. We examined how the combined effects of non-collapsibility and built-in selection bias influence the marginal hazard ratio estimate over time. Creating a study where we know there is a strong positive effect of treatment and a even stronger positive effect of a confounding covariate results in a risk set that changes significantly over time.

What about other scenarios?

To visualize the hazard ratio over time for different combinations of treatment effects and confounders, we create Figure 14 where we expand the previous ibuprofen example to analyze the first 20 days after a swelling event affecting $N = 50,000$ people. For each individual, we eliminate confounding using the counterfactual approach, which involves hypothetically assuming that, contrary to reality, everyone was observed both with and without ibuprofen. We simulate ibuprofen as having a positive effect with $HR = 1.5$ (faster healing), no effect $HR = 1$, or a negative effect on healing time $HR = 0.5$, combined with the binary other measures under the same conditions, positive as \$HR = 2.5, no effect as $HR = 1$, or a negative effect as $HR = 0.626$.

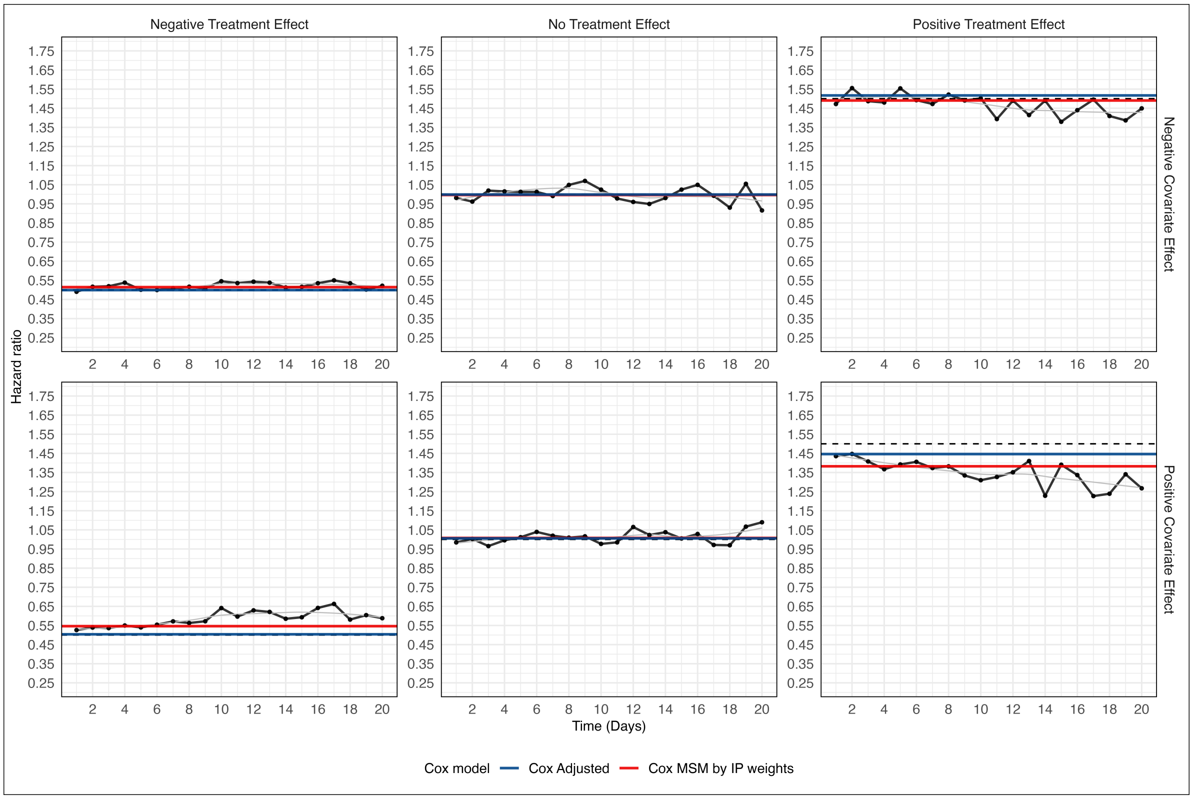
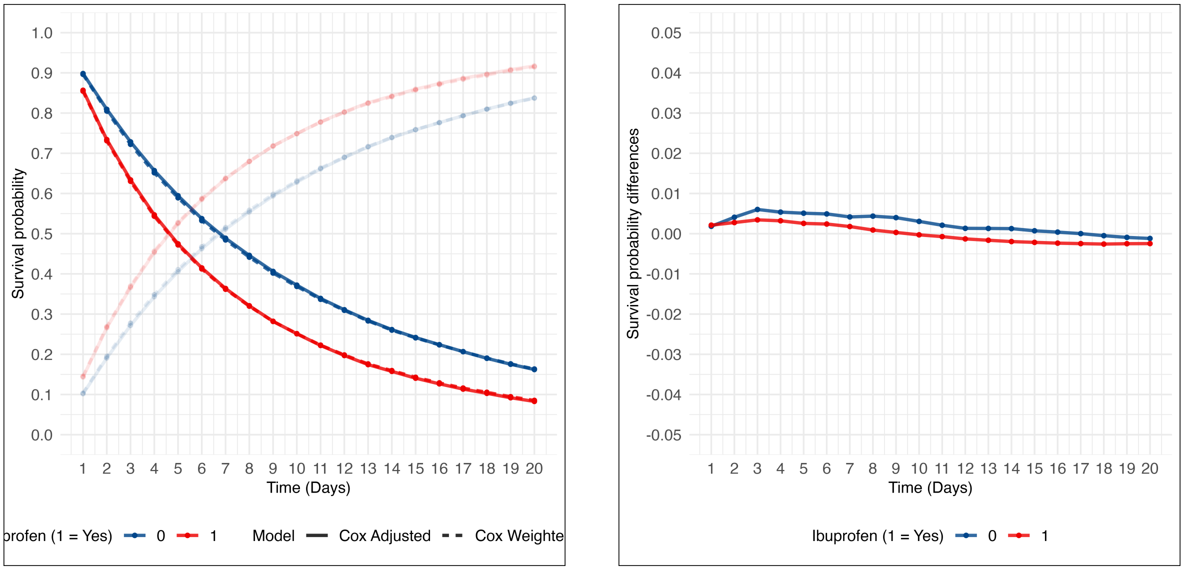


Figure 14: Hazard ratio over time given pos., neg. or no effect of treatment or other measures.

In Figure 14 we see that for the other combinations of treatment and confounding effects, the marginal hazard ratio over time behaves more stably or even stays approximately constant over time. The two-part problem of non-collapsibility and built-in selection bias of the Cox model is most pronounced when both treatment and confounder have a strong positive effect on the hazard rate. A Monte Carlo simulation of all six combinations of treatment and confounder effects is presented in the appendix.

2.7 Marginal Survival Probabilities from an Adjusted Cox Model

Survival probabilities are not non-collapsible; therefore, to compare classical and counterfactual estimates of survival probabilities, the classically or counterfactually estimated marginal survival curves offer a good comparison of the two approaches. Especially since the counterfactual framework stresses the use of marginal estimates and survival probabilities over adjusted hazard ratios. The conditional adjusted Cox model can be collapsed to marginal survival probabilities since the Cox model estimates survival for each individual. The marginal survival curves of the ibuprofen study described in Section 2.4.6 are plotted below in Figure 15 for the adjusted Cox model and the IP-weighted Cox model to visualize the outcomes and their comparisons.



(a) Survival probabilities

(b) Survival differences

Figure 15: Survival probabilities over time, IP-weighted Cox model and adjusted Cox model, perfect overlap

The differences visualized by Figure 15 show that the two survival curves overlap almost perfectly.

The collapse is possible because Cox model estimated survival probabilities can be standardized by predicting the survival probabilities for each individual in the study under everyone treated and everyone untreated and finding the average survival probabilities over time. This method is called standardization and is denoted in Section 2.7.1.

Using the adjusted Cox model to predict survival probabilities for each individual under both treatment strategies is computationally intensive and proved to be infeasible for the $n = 50.000$ individuals in our ibuprofen study. To be able to find the marginal survival probabilities we sampled 5.000 individuals from the study and used them to predict and standardize the survival probabilities over time.

2.7.1 Standardization of a Conditional Estimate.

Standardizing a conditional probability to find a marginal estimate is the foundation of the counterfactual method called the parametric g-formula which we mentioned as one of the g-methods in Section 2.5.2. As shown in Figure 15, standardization of the adjusted Cox model can closely approximate the marginal survival probabilities estimated by the IP-weighted Cox model MSM.

The two methods of making a marginal model or standardizing a conditional model estimate are, in fact, stated as theoretically equivalent in (Hernán & Robins, 2025, p. 173), given the models are correctly specified and censoring is correctly accounted for. In Sections 2.9 and 2.10 we will further define the parametric g-formula, the stated equivalence of methods and the common deviance from it in observational studies, where some model misspecification is assumed unavoidable.

The ibuprofen study example has no censoring, and we have focused on comparisons between the classical and counterfactual frameworks with correctly specified models before, in the next Section 2.8 moving on to more complex scenarios with misspecification and censoring.

This allows us to introduce the parametric g-formula under conditions where we can explore how and if the two counterfactual methods differ.

2.8 Censoring and Competing Events

As mentioned in Section 1, survival time analysis must take into account censoring and competing events that block the observation of the events of interest. So far, we have defined and examined survival times within classical and counterfactual frameworks without accounting for these factors. Censoring and competing events are crucial in survival analysis. A participant in a study might drop out, making it impossible to determine if they will eventually experience the event of interest. For example, individuals may be censored due to the study ending, loss to follow-up, or death. When death occurs, it is a competing event that prevents the event of interest from happening; in other cases, the event may still occur, but the researcher cannot observe it, meaning they are censored.

A perfect study scenario, like the one we examined with the simulated ibuprofen example, involves no individuals lost to follow-up or censored. However, this is rarely the case in real life. Simulating survival times for a study, considering possible competing events, administrative censoring, and loss to follow-up, produces a stochastic process of random variables indexed by time t. The earliest of these random variables represents the observed event times, denoted as T_i for individual i. For simplicity, in this work, we treat competing events as censoring events. We refer to censoring with the random variable $C\{0, 1\}$, if $C = 1$, the survival time is a censored time. In fact, Y and C are now complementary events; if one occurs, the other cannot.

Although the survival time is censored, the censoring time still offers essential information about survival up to that point.

2.8.1 Censoring Mechanisms

To draw causal conclusions from censored survival data, it is important to consider the reasons and circumstances under which participants were censored. In addition to administrative censoring at the end of the study or competing events, there are three main censoring mechanisms:

1. Administrative censoring.
    - Censoring by the end of the study, a single date is set for all participants, marking the end of the study. The time spent in the study can then vary between participants.
    - By the end of the study, participants may have different start and end dates, but each individual has a maximum time limit.
2. Death from age and other competing events.

- The event of death from natural causes. For simulations, a maximum age must be set for all participants. When subjects reach this age, they are considered to have died from natural causes.

3. Censoring mechanisms.

   - Censoring completely at random.

   Censoring is independent of all other variables. This is often called loss to follow-up and can happen for unknown reasons. When censoring is fully random, standard survival analysis methods can be used without adjustments because censoring does not cause bias.

   - Censoring at random.

   Covariate X can influence censoring, which is random depending on measured covariates. For example, death due to age is a censoring event that clearly depends on age. When censoring is at random, it skews the risk set by removing individuals with specific covariate values more often than others. This creates selection bias in methods that rely on the risk set at each time point, such as the Kaplan-Meier estimator, the Cox model, and MSM models.

   - Censoring not at random.

   Outcome influences censoring; how effectively treatment impacts someone can determine whether they continue treatment. This is also called informative censoring and is the hardest type of censoring to recognize and correct for. This discussion will not go into how to adjust for non-random censoring; however, it is important to be aware of its presence and the limitations it causes for both approaches.

2.8.2 Left Truncation

Left truncation, also known as delayed entry, occurs when individuals commence participation in a study at various times. In a straightforward survival analysis example, it is often assumed that all participants begin at $t = 0$ and that the hazard rate is calculated by considering the lesser of their individual right-censoring or event times, represented as the ratio of the number of events to those at risk at each specific time point. Under conditions of left truncation, individuals may continue to enter the study and begin contributing to the hazard rate at different times. If all individuals share a final date of study then left truncatation causes some people to be administratively censored after a shorter time than other and thus less likely to have an event in the time frame of the study.

The models or counting process approach we implement in this thesis are equally valid for studies with left truncation, right-censoring or both and the adjustements made to account for both censoring and delayed entry are built on the same foundation for all methods or approaches. For the remainder of the thesis we will emphasize the bias caused by righ-censoring times but we note that left truncation would not require a different set up (Odd Aalen, 2008. p. 6).

2.8.3 Simulating Censoring Times

Let T^c be an exponentially distributed random variable representing censoring time. Hazard rate of censoring is denoted as $h_c(t) = \lambda_c$. Possible censoring mechanisms are categorized into three versions: censoring completely at random (MCAR), censoring at random (MAR), or censoring not at random (MNAR). The abbreviations refer to missing completely at random, missing at random, and missing not at random, respectively. We therefore need to create a censoring simulation that accounts for these mechanisms so that we can explore the different results from a classical or counterfactual survival analysis under these scenarios.

Missing completely at random (MCAR) is independent of covariates and therefore does not bias estimates. The baseline censoring of survival times T^c is denoted, $T^c \sim \text{Exponential}(\lambda_c)$, $\text{E}[T^c] = \frac{1}{\lambda_c}$. The parameter λ_c is chosen to be a function of the desired censoring level and the minimum of the baseline hazards in the treated and untreated groups, $\lambda_c = \frac{\min(\lambda_1,\lambda_0)(1-p_c)}{p_c}$ where p_c is the desired censoring level.

Missing at random (MAR) can be created by an accelerated failure time (AFT) model. It is used to simulate covariate conditional survival times. The AFT model is defined for a given censoring time T^c and a vector of covariates in $X = (X_1, \dots, X_k)$ with assigned scale parameters $w = (w_1, ..., w_k)^\top$, on the log-scale. If a covariate has multiple levels then each level is assigned a different w_j. For each X_j, there then is a constant $w(x_j) = w_j$. Unless otherwise specified, all $w_j = 0$.

The AFT model assumes that covariates have a multiplicative effect on the survival time and an additive effect on log survival time (Majeed, 2020) and (Odd Aalen, 2008, p. 443).

$$T^c_{i,MAR} = \exp(-\log(\lambda_c) + w^\top X) = \lambda_c^{-1} \exp\left(w^\top X\right)$$

By using an AFT model, we have guaranteed that by setting $w \neq 0$ for some $x \in X$, the censoring is at random but not completely.

Now each individual can have a censoring time T^c based on the chosen censoring mechanisms. If we, for now, ignore survival times, and think of a study where we would only observe censoring times we can still see that these simulated censoring times would not be observed for all individuals. That is because it is not only the random censoring times that can occur. The random censoring times can themselves be censored by other censoring times f.x. if a simulated censoring time is after 20 days in a 10 day study, the censoring time is administratively censored by the end of study. This can also come up due to death from age since we have to choose a maximum age of individuals to prevent immortality in our study. To define the final censoring time of individual i we refer to it as the observed censoring time. The observed censoring time is now defined as a random variable defined as the minimum of the random variables defined by the possible censoring times of an individual in the study. Given a censoring mechanism that is completely at random we denote this as,

$$T_i^c = \min(T_{i,MCAR}^c, T_{i,age=100}^c, T_{i,administrative}^c). \tag{19}$$

The final observed survival time of individual i is then the minimum of the censoring time T_i^c and uncensored survival time T_i^a. If the observed survival time is $T_i = T_i^c$, then $Y_i = 0$ and $C_i = 1$.

2.8.4 Adjustments for Censoring Bias with the Proportional Hazards Cox Model

If censoring is at random in a study, the Cox model is adjusted by including the covariate that influences censoring in the model, just as if it were a confounder. If the covariate only affects censoring and not the outcome or treatment assignment adjustment becomes optional since the censoring does not induce bias on the treatment effect estimation. Censoring completely at random does not need any adjustment since the risk sets are not biased by selection bias from censoring. If censoring is not at random, more advanced methods are needed to adjust for the bias it introduces and that will be outside the scope of this work for all methods.

2.8.5 MSMs with Inverse Probability of Censoring Weights (IPCW)

Just as we modeled the propensity score of treatment to generate an inverse probability of treatment weight, we can similarly model the probability of censoring, given treatment and covariates X, to create the inverse probability of censoring weights, (IPCW).

The joint modeling of the inverse probability of treatment weights and the inverse probability of censoring weights can be used to meet the three assumptions of the causal inference framework: exchangeability $Y^{a,c=0} \perp (A, C) \mid X$, positivity, and consistency. The joint modeling of treatment and censoring probabilities creates a pseudo-population where censoring occurs completely at random (Hernán & Robins, 2025). Constant stabilized censoring weights are defined as,

$$IPCW = \frac{\mathrm{P}[C = 0 \mid A]}{\mathrm{P}[C = 0 \mid A, X]}. \tag{20}$$

For treatment weights we estimate the probability of treatment given confounders as a constant given a non-time varying dichotomous treatment. For the censoring weights we estimate the probability of staying uncensored given covariates in each time interval k. The censoring weights are therefore more complex to estimate than the constant treatment weights. Methodologically the probability of being uncensored is estimated in the same way as estimating the hazard over time by a logistic regression model, instead of finding the probability of event occurance we find the probability of remaining uncensored in each time interval for each individual in person-time format. Just as the survival probability in interval k is defined as the conditional probability of no event in intervals $[1, k-1]$ up to interval k, the censoring probabilities are defined as the conditional probability of not being censored in intervals $[1, k-1]$ up to interval k. The censoring weights, $IPCW_i(t)$, are the inverse of the estimated probability of remaining uncensored up to time t, stabilized by the marginal probability of not being censored given observed treatment a_i. Defined as follows for each individual i and intervals $1 \leq k \leq t < T_i$ where $C(k)$ is the censoring random variable in interval

k and $c_i(k) = 0$ denotes that individual i is uncensored in interval k, T_i and t are the observed survival time.

$$IPCW_i(t) = \prod_{k=1}^{\text{int}(t)} \frac{P\left(C(k) = c_i(k) \mid C(k-1) = c_i(k-1), A = a_i\right)}{P\left(C(k) = c_i(k) \mid C(k-1) = c_i(k-1), A = a_i, X(k-1) = x_i(k-1)\right)}. \tag{21}$$

Since each individual has a final observed survival time of T_i the censoring weights are at most defined at $t = T_i$ which is then the censoring weight that can be used for a time-to-event analysis by a Cox model as an IP-weighted Cox MSM for a marginal estimate. The individual model weight is then defined as the constant, $w_i = IPTW_i \cdot IPCW_i(T_i)$ and for each interval in person time the weights are $w_i(t) = IPTW_i \cdot IPCW_i(t)$. The definition of the longitudinal censoring weights, $IPCW_i(t)$ is given in (Odd Aalen, 2008, p. 381) as sw_i for the time-varying treatment weights case and used here for the censoring weights case described in (Cole & Hernán, 2008).

2.9 Standardization or the Parametric g-formula.

Now we are ready to go over the parametric g-formula which is the model that fully utilizes a comparison of two potential outcomes rather than adjusted hazard ratios or IP-weighted hazard ratios. Standardization or the parametric g-formula for a mean potential outcome of event for everyone, uncensored, treated, and untreated, $\mathrm{E}[Y^{a,c}]$ is denoted as,

$$\mathrm{E}\left[Y^{a,c=0}\right] = \sum_x \mathrm{E}[Y \mid A = a, C = c, X = x] \times \mathrm{P}[X = x] \tag{22}$$

for a time-fixed treatment A. If the counterfactual assumptions of exchangeability, positivity, and consistency, defined in 2.5.1, hold, the standardized mean outcome in the uncensored and untreated group equals the mean potential outcome had everyone been uncensored and untreated. We refer back to Section 2.5.3.5 on MSM models for the overview of pooled logistic regression models that can be used to estimate $\mathrm{E}[Y \mid A = a, C = 0, X = x]$. The difference is that equation (15) is now, like the Cox model, adjusted for confounding and censoring, instead of weighted by IPTW and IPCW. The standardization is then done by predicting for each individual in the study under each treatment strategy over all time intervals k and averaging the predicted outcomes over all individuals using the law of total expectation. This can be denoted as $\frac{1}{N}\sum_{i=1}^{N} \widehat{\mathrm{E}}\left[Y \mid A = a, C = 0, X = x\right]$ for all N individuals and chosen time intervals of the study since equation (22) can also be written as the double expectation $\mathrm{E}[\mathrm{E}[Y \mid A = a, C = 0, X = x]]$. Equation (15) can be expressed in terms of the parametric g-formula for each individual as,

$$\text{logit } \mathrm{P}\left[Y_{k+1,i} = 1 \mid Y_{k,i} = 0, A = a_i, X = x_i\right] = \beta_{0,k} + \beta_1 a_i + \beta_2 a_i \times k + \beta_3 a_i \times k^2 + (\beta_4, \ldots, \beta_j)^\top \mathbf{X}_{ij} \tag{23}$$

We include $\beta_{0,k}$ and set the linear and quadratic time-by-treatment terms to capture changes in the baseline hazard over time and to allow the treatment effect to vary over time.

The pooled logistic regression model in equation (23) is commonly referred to as the Q-model or the outcome model. To distinguish it as a middle step to the parametric g-formula, we refer to it as the Q-model.

In practice, covariates are often continuous, and in that case the standardized mean is written as an integral with the joint cumulative distribution function of the random variables in X instead of a sum, $\int \mathrm{E}[Y \mid A=a, C=0, X=x] dF_X(x)$.

The discrete-time survival is as in equation (17) the product of one minus the, in this case, conditional hazards up to time t_{end},

$$\prod_{m=1}^{t_{\text{end}}} \mathrm{P}\left[Y_m = 0 \mid Y_{m-1}=0, C_m=0, A=a, X=x\right]. \tag{24}$$

We can standardize the discrete-time survival over the covariates in X in the same way as in equation (22).

By first constructing a conditional model, as in Cox regression, and then applying standardization, the parametric g-formula produces a middle stage counterfactual model, the Q-model that yields conditional effect estimates comparable to those obtained from the Cox model. This allows us to compare a conditional estimate obtained by both the classical and counterfactual approaches.

2.10 IP-weighted MSM and the Parametric g-formula Equivalence

The IP-weighted mean and the standardized mean of Y^a are theoretically the same. This means the parametric g-formula and IP-weights are two methods within the causal framework that, in theory, should give identical results if there is no model misspecification, no unmeasured/mismeasured confounding or positivity violations. The proof of this (Hernán & Robins, 2025, p. 25), is as follows.

Assume treatment A defined as a finite discrete random variable A and that positivity, consistency and exchangeability given covariates $X \in \left\{x_1, ..., x_j\right\}$ is assumed to hold under the density function of A given X, $f(a \mid x) > 0$ for all x such that $\mathrm{P}[X=x] > 0$. The standardized mean of outcome Y as the weighted average of the conditional means of Y for observed treatment level a is defined as,

$$\mathrm{E}\left[Y\right] = \sum_x \mathrm{E}[Y \mid A=a, X=x]\,\mathrm{P}[X=x]. \tag{25}$$

The IP-weighted mean of Y for treatment level a is defined as,

$$\mathrm{E}\left[Y\right] = \mathrm{E}\left[\frac{\mathbf{1}_A(a) Y}{f(A \mid X)}\right] \tag{26}$$

With the indicator function $\mathbf{1}_A(a) := \begin{cases} 1 & \text{if } a = A, \\ 0 & \text{if } a \neq A. \end{cases}$

In what follows, we will display the proof from (Hernán & Robins, 2025, p. 25) of the equality of the IP-weighted and standardized mean counterfactual outcome of Y^a.

$$\mathrm{E}\left[Y^a\right] = \sum_x \mathrm{E}[Y^a \mid A = a, X = x] \, \mathrm{P}[X = x] = \mathrm{E}\left[\frac{\mathbf{1}_A(a) Y^a}{f(A \mid X)}\right].$$

By the law of total of expectation we derive the standardized mean from the IP-weighted mean,

$$\begin{aligned} \mathrm{E}\left[\frac{\mathbf{1}_A(a) Y}{f(A \mid X)}\right] &= \sum_x \frac{1}{f(a \mid x)} \mathrm{E}[Y \mid A = a, X = x] f(a \mid x) \, \mathrm{P}[X = x] && (27) \\ &= \sum_x \mathrm{E}[Y \mid A = a, X = x] \, \mathrm{P}[X = x]. && (28) \\ & && (29) \end{aligned}$$

X is not constrained as discrete, if it is continuous, then the sum over X is replaced with an integral.

Conditional exchangeability in addition to the previous conditional positivity ensures that both the IP-weighted and the standardized means are, by consistency, equal to the counterfactual mean $\mathrm{E}\left[Y^a\right]$.

$$\begin{aligned} \mathrm{E}\left[Y^a\right] &= \sum_x \mathrm{E}\left[Y^a \mid X = x\right] \mathrm{P}(X = x) && (30) \\ &= \sum_x \mathrm{E}\left[Y^a \mid A = a, X = x\right] \mathrm{P}(X = x) && (31) \\ &= \sum_x \mathrm{E}\left[Y \mid A = a, X = x\right] \mathrm{P}(X = x). && (32) \end{aligned}$$

IP-weighted mean $\mathrm{E}\left[\frac{\mathbf{1}_A(a) Y}{f(A|X)}\right]$ is then also equal to the counterfactual mean of $\mathrm{E}\left[Y^a\right]$. To show that by conditional exchangeability this statement holds we derive the following,

$$\begin{aligned} \mathrm{E}\left[\frac{\mathbf{1}_A(a) Y}{f(A \mid X)}\right] &= \mathrm{E}\left\{\mathrm{E}\left[\left.\frac{\mathbf{1}_A(a) Y}{f(a \mid X)}\right| X\right]\right\} && (33) \\ &= \mathrm{E}\left\{\mathrm{E}\left[\left.\frac{\mathbf{1}_A(a)}{f(a \mid X)}\right| X\right] \mathrm{E}\left[Y^a \mid X\right]\right\} && (34) \\ &= \mathrm{E}\left\{\mathrm{E}\left[Y^a \mid X\right]\right\} && (35) \\ &= \mathrm{E}\left[Y^a\right]. && (36) \end{aligned}$$

Since this is a theoretical proof of the equivalence between IP-weighting and standardization, it does not necessarily reflect practical application in epidemiology. In reality, the equivalence does not translate to practice when models and datasets are imperfect. Bias may not develop similarly in effect estimates from both methods since they are using different models but close results can suggest that the models are well specified given data, but that is not a guarantee while large differences in results do indicate misspecification in the treatment or outcome models.

Impact from the in practice, common deviations from the positivity assumption varies between IP-weighting and standardized estimates that depend on parametric models. Standardization, can handle the lack of positivity by parametric extrapolation but the IP-weighted mean loses its causal interpretation. If we fit a model for $\mathrm{E}[Y \mid A, X]$ that will smooth over X with structural zeroes the smoothing will introduce bias into the estimation, and therefore the 95% confidence intervals around the estimates will cover the true effect less than 95% of the time. This does not have to be an issue of lack of data since data can have structural non positivity and thus positivity might never be reached, no matter the amount of data collected (Chatton et al., 2020). IP-weights cannot be estimated in the presence of non-positivity and are more susceptible to violations or near-violations of positivity in comparison to standardization which can though give small variance estimates with large bias.

To conclude, these are two different ways to reach the same exact goal of estimating the counterfactual mean outcome Y^a, each way has their own complications and biases that can arise. If we assume that in practice things do generally not go perfectly we also have to assume that whichever way we go there will be some bumps along the way that affect the final estimates.

2.11 Flow Diagram and Summary of Methods

To demonstrate the differences between the classical and counterfactual approach, the standard pathways in each are compared in the following flow diagram.

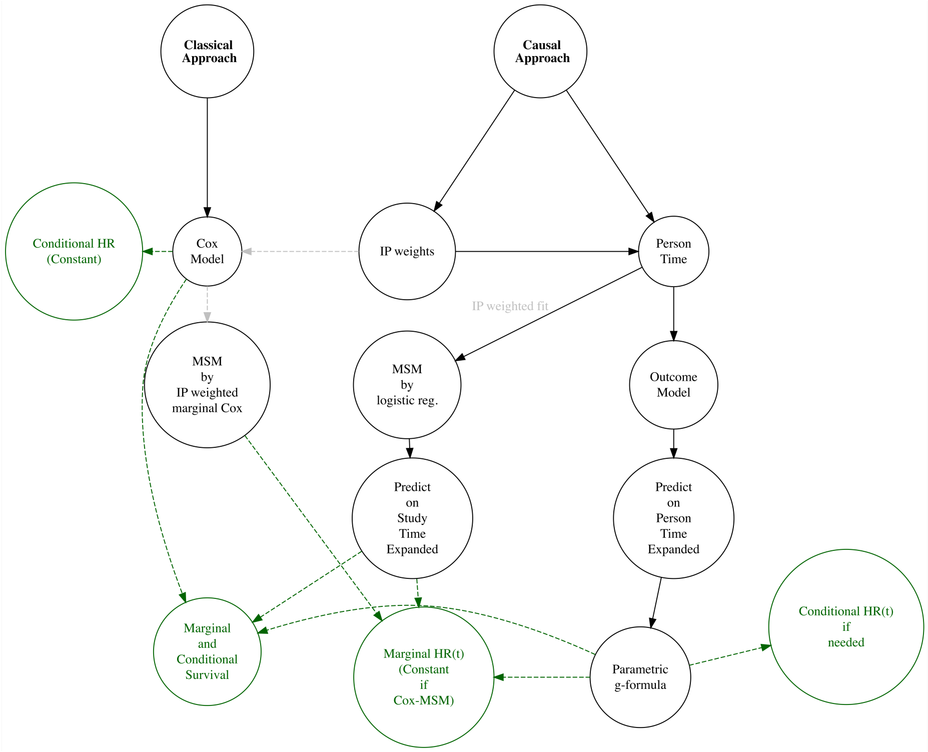

Classical Approach
Causal Approach
Conditional HR (Constant)
Cox Model
IP weights
Person Time
IP weighted fit
MSM by IP weighted marginal Cox
MSM by logistic reg.
Outcome Model
Predict on Study Time Expanded
Predict on Person Time Expanded
Conditional HR(t) if needed
Marginal and Conditional Survival
Marginal HR(t) (Constant if Cox-MSM)
Parametric g-formula

Figure 16: Flow diagram of methods

Figure 16 summarizes the two approaches and the different methods used to adjust for bias from confounding and censoring that we have discussed in this work. The assumed starting point is an observation dataset with time-to-event data, one row per individual and a dichotomous treatment variable.

Based on the diagram the choice of approach should depend on the research question at hand, the data available and the choice of aiming for a conditional estimate or a marginal one. Careful consideration of the assumptions, possible biases and limitations of each approach should also be a factor in the decision making. For example if there is doubt that the proportional hazards assumption holds for the Cox model then the counterfactual approach should be preferred since it does not rely on a constant hazard ratio over time.

2.12 Model Validation

For this work, where we have simulated data with known true parameters, we choose to validate the classical and counterfactual models by a Monte Carlo simulation. The Monte Carlo confidence intervals are found by continuously generating datasets to estimate the hazard ratios and survival curves for each. For each dataset we estimate the parameters of interest and compare them to the true parameters used to generate the data.

We visualize the distribution of estimates over time using the 95% quantiles of estimates and their corresponding 95% Monte Carlo confidence intervals. The confidence intervals are calculated by assuming that the estimates converge to a normal distribution as the number of simulations increases, in accordance with the central limit theorem (Batyrbekova et al., 2024).

The mean and 95% Monte Carlo confidence intervals for $\hat{\beta}$ are then calculated for m simulations as $\hat{\beta} \pm 1.96 \cdot \frac{s}{\sqrt{m}}$, where s is the standard deviation of the m estimates.

Additionally, we assess the bias of the Monte Carlo mean estimates from the true values and the coverage of each study´s 95% confidence intervals. This is done by calculating the proportion of the m studies where the true parameter value lies within the per model estimated 95% confidence interval. Since for the marginal hazard ratio over time we estimate multiple time points and do not have a single true value to compare to, we find the bias-eliminated coverage, instead of finding the proportion of simulations with the true value within the confidence interval we find the proportion of simulations where the Monte Carlo mean estimate lies within the per model estimated 95% confidence interval. To estimate the power of the models we also calculate the proportion of simulations where the null hypothesis is rejected at a significance level of $\alpha = 0.05$.

The performnance measures are summarized in Table 2 and can be found in more detail in (Morris et al., 2019).

Table 2: Monte Carlo performance measures.

Performance	Definition	Estimate	Monte Carlo SE of estimate
Bias	$\mathrm{E}[\hat{\beta}] - \beta$	$\frac{1}{m}\sum_{i=1}^{m} \hat{\beta}_i - \beta$	$\sqrt{\frac{1}{m(m-1)}\sum_{i=1}^{m}\left(\hat{\beta}_i - \bar{\beta}\right)^2}$
Coverage	$\Pr\left(\hat{\beta}_{\text{low}} \ \leq \beta \leq \hat{\beta}_{\text{upp}}\ \right)$	$\frac{1}{m}\sum_{i=1}^{m} 1\left(\hat{\beta}_{\text{low}\ ,i} \leq \beta \leq \hat{\beta}_{\text{upp}\ ,i}\right)$	$\sqrt{\frac{\text{Cover }.\widehat{\times(1-}\ \widehat{\text{Cover}}.\ )}{m}}$
EmpSE	$\sqrt{\mathrm{Var}(\hat{\theta})}$	$\sqrt{\frac{1}{n_{\text{sim}}-1}\sum_{i=1}^{n_{\text{sim}}}\left(\hat{\theta}_i - \bar{\theta}\right)^2}$	$\frac{\widehat{\text{EmpSE}}}{\sqrt{2(n_{\text{sim}}\ -1)}}$
Bias-eliminated coverage	$\Pr\left(\hat{\beta}_{\text{low}} \ \leq \bar{\beta} \leq \hat{\beta}_{\text{upp}}\ \right)$	$\frac{1}{m}\sum_{i=1}^{m} 1\left(\hat{\beta}_{\text{low}\ ,i} \leq \bar{\beta} \leq \hat{\beta}_{\text{upp}\ ,i}\right)$	$\sqrt{\frac{\text{Cover }.\widehat{\times(1-}\ \widehat{\text{Cover}}.\ )}{m}}$
Rejection % (power or type I error)	$\Pr\left(p_i \leq \alpha\right)$	$\frac{1}{n_{\text{sim}}}\sum_{i=1}^{n_{\text{sim}}}\ 1\left(p_i \leq \alpha\right)$	$\sqrt{\frac{\widehat{\text{Power}}\ \times(1-\ \widehat{\text{Power}}\ )}{n_{\text{sim}}}}$

3 Monte Carlo Simulations of Classical and Counterfactual HR Estimates

Based on the mehods described in the previous Sections we set up a Monte Carlo simulation study to compare the classical Cox model to the counterfactual methods of IP-weighted Cox MSM and the parametric g-formula. We evaluate how well the models perform in estimating the hazard ratio of treatment over time and as a constant estimate. We compare the estimates from the MSM models to the estimates from the parametric g-formula to evaluate how close they are to each other and report all estimates with their Monte Carlo performance measures.

As an independent continuation of the first simulation, we set up a second simulation study in Section 3.2 to evaluate how different model parameter choices affect bias in the estimates from a classical Cox model and the IPTW Cox marginal structural model. The second simulation is evaluated visually by plotting the Monte Carlo mean estimates with their 95% Monte Carlo confidence intervals over time for different combinations of covariates included in the treatment model and outcome model.

To create an overview of the notation and variables used in both simulations we denote them in Table 3.

Table 3: Notation and variables used in the simulation.

Symbol	Name	Description	Purpose
Treatment and outcomes			
A	Treatment	(0 = Untreated, 1 = Treated)	Did people receive the treatment?
Y_k	Outcome	(0 = No event, 1 = Event)	Indicator of event occurrence in interval k
C_k	Censoring	(0 = Not censored, 1 = Censored)	Indicator of censoring occurrence in interval k
Baseline covariates			
X_1	Gender	Binary (0 = female, 1 = male)	Possible confounder or effect modifier
X_2	Age	Age in years (continuous)	Possible confounder
X_3	Instrument	Binary (0 = No, 1 = Yes)	Instrumental variable
X_4	Neighborhood	Categorical (factor)	Possible confounder
X_5	Health before	Binary (0 = Healthy, 1 = Sick)	Possible confounder
X_6	Income	Monthly income (continuous)	Collider (by study design)
Time-to-event quantities			
$T^{a=0}$	Survival time (Untreated)	Survival time under $A = 0$	Counterfactual survival time
$T^{a=1}$	Survival time (Treated)	Survival time under $A = 1$	Counterfactual survival time
T^c	Censoring time	Censoring time	Counterfactual censoring time

3.1 Treatment Effect Estimation

We generate M replicates of studies, each with N individuals. For each individual i we generate a vector of covariates X, a survival time given treatment T^a and censoring time T^c. To follow a Cox model we need to define the survival time T^a as a function of the covariates X and the treatment A. The survival time is defined as the minimum of the event time and the censoring time, i.e. $T_i^a = \min(T^a, T^c)$. The event time is defined as a function of the covariates and treatment, i.e. $T^a = f(X, A)$. To create data for classical survival analysis and causal survival analysis we will use the same data generation process but we simulate both $T^{a=1}$ and $T^{a=0}$ as a function of the covariates X and treatment A for each individual, the ratio between the two is the true hazard ratio.

The following DAG in Figure 17 illustrates the chosen causal structure of the simulated observational study.

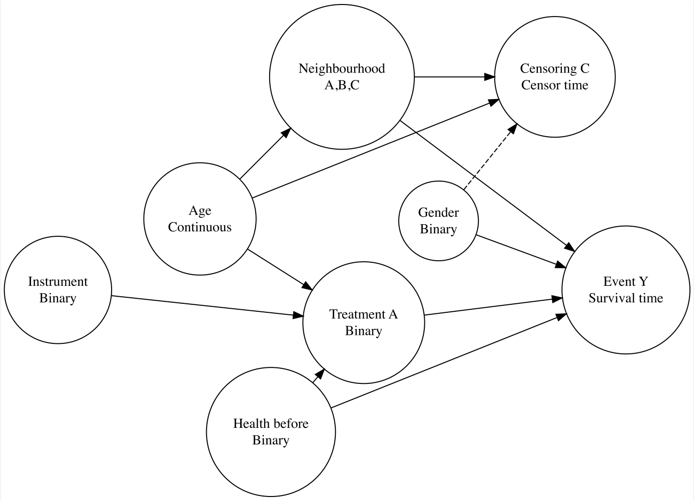


Figure 17: DAG of causal relationships

Figure 17 shows how we have built the data generation process to reflect a realistic observational study for an undisclosed drug treatment A. To indicate a hypothetical unknown relationship we use a dashed line from gender to censoring C.

To draw a fair comparison between methods we assume that the same covariates are confounders to be adjusted for, regardless of approach. In Section 1.2 we discussed how a box around a variable on a DAG indicates that it is adjusted for to eliminate bias and obtain exchangeability between treatment groups. A DAG that reflects the chosen covariates is shown in Figure 18.

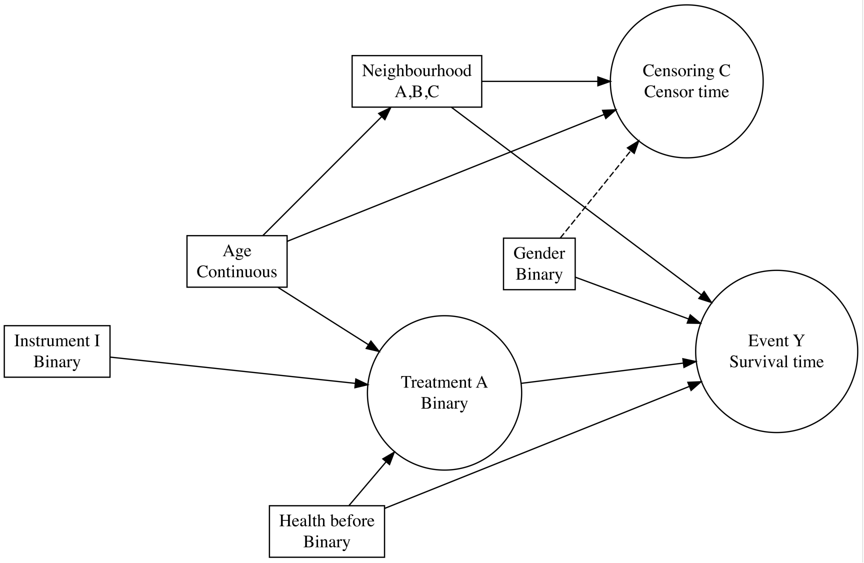


Figure 18: DAG of adjusted causal relationships

The choice of covariates in Figure 18 is based on the clear confounders of X_5 as the health before treatment and the joint confounding from X_2 and X_4 as age and neighborhood. Only adjusting for X_4 since it acts as a confounder by mediating the effect of age on treatment could be enough to adjust for confounding but to be either safe or overadjust we include both. We choose to include the instrument X_3 in all models to to allow for the possible bias that the inclusion of an instrument can introduce and see how the methods compare under this scenario. The choice of gender X_1 as a covariate that afects outcome directly and "unknownly" affects censoring is to create a realistic scenario where researcher will not be able to know or measure all relationships in a study.

3.1.1 Set Parameters of Monte Carlo Simulation

We generate the study described in 3.1 by simulating for each random variable based on the defined mechanisms for censoring times in Section 2.8.3, survival times in Section 2.4.5 and treatment assignment in Section 2.3.3 based on variables denoted in Table 3. The parameter values are summarized in Table 4.

Table 4: Parameters of the Monte Carlo simulation.

Notation	Value	Description
Study design		
N	5 000	Number of individuals
M	500	Number of Monte Carlo iterations
T_{study}	20	Time of study (years)
Treatment assignment		
α_0	$p = (0.5)$	Baseline probability of treatment, Bernoulli
α_1	3	Log odds ratio for health before variable
α_2	4	Log odds ratio for the instrumental variable
α_3	0.98	Log odds ratio for unit increase in age
Covariate distributions		
X_{age}	$\mathcal{TN}(50, 30, 20, 90)^1$	Age (years), truncated normal
$p_{\text{health before}}$	$p = (0.5)$	Health before, Bernoulli
p_{gender}	$p = (0.5)$	Gender, Bernoulli
$p_{\text{neighborhood}}$	$p = (0.4,\ 0.4,\ 0.2)$	$\Pr\big(X_{\text{neigh}} = 1, 2, 3 \mid \text{age} < 70\big)$
$p_{\text{neigh}\mid\text{senior}}$	$p = (0.4,\ 0.3,\ 0.3)$	$\Pr\big(X_{\text{neigh}} = 1, 2, 3 \mid \text{age} \geq 70\big)$
Survival time simulation		
λ_0	1/15	Baseline hazard in untreated T^0
λ_1	1/10	Baseline hazard in treated T^1
$\exp\,(\beta_{\text{health}})$	2.5	HR for sick vs. healthy
$\exp(\beta_{\text{gender}})$	0.909	HR for men vs. women
$\exp(\beta_{\text{age}})$	1	HR per unit increase in age
$\exp(\beta_{\text{neighborhood}})$	0.833	HR for neighborhood C vs. others
Censoring model		
T_c^{age}	100	Administrative censoring age (years)
T_c^{admin}	T_{study}	Administrative censoring time (years)
$w_{\text{neighborhood}}$	0.6	Neighborhood C shift on censor hazard
w_{gender}	0.7	Gender men shift on censor hazard
λ_c	$0.4\lambda_1$	Baseline censoring hazard

3.1.2 Simulated Observational Study Overview

We take one random study out of the 500 to visualize the simulated data and display in Figure 19. The plot is colored by treatment group. Displayed survival times are simulated survival times based on treatment assignment before cut off by censoring. The survival times are therefore the true survial times of T^a and T^c and the event time is the observed times. We can see how the covariates are associated with treatment, outcome and censoring. As we defined in Table 4 health before and the instrumental variable are the strongest predictors of treatment. Age affects both treatment and censoring times vaguely but since age is truncated at 90 years old at the start of study and the study is over 20 years few individuals reach this age. Age also has an effect on

[1] Truncated normal distribution; see [@TruncatedNormal].

survival times indirectly through neighborhood since older individuals are defined as more likely to live in one neighborhood. Health before is a clear confounder as it affects both treatment and survival times.

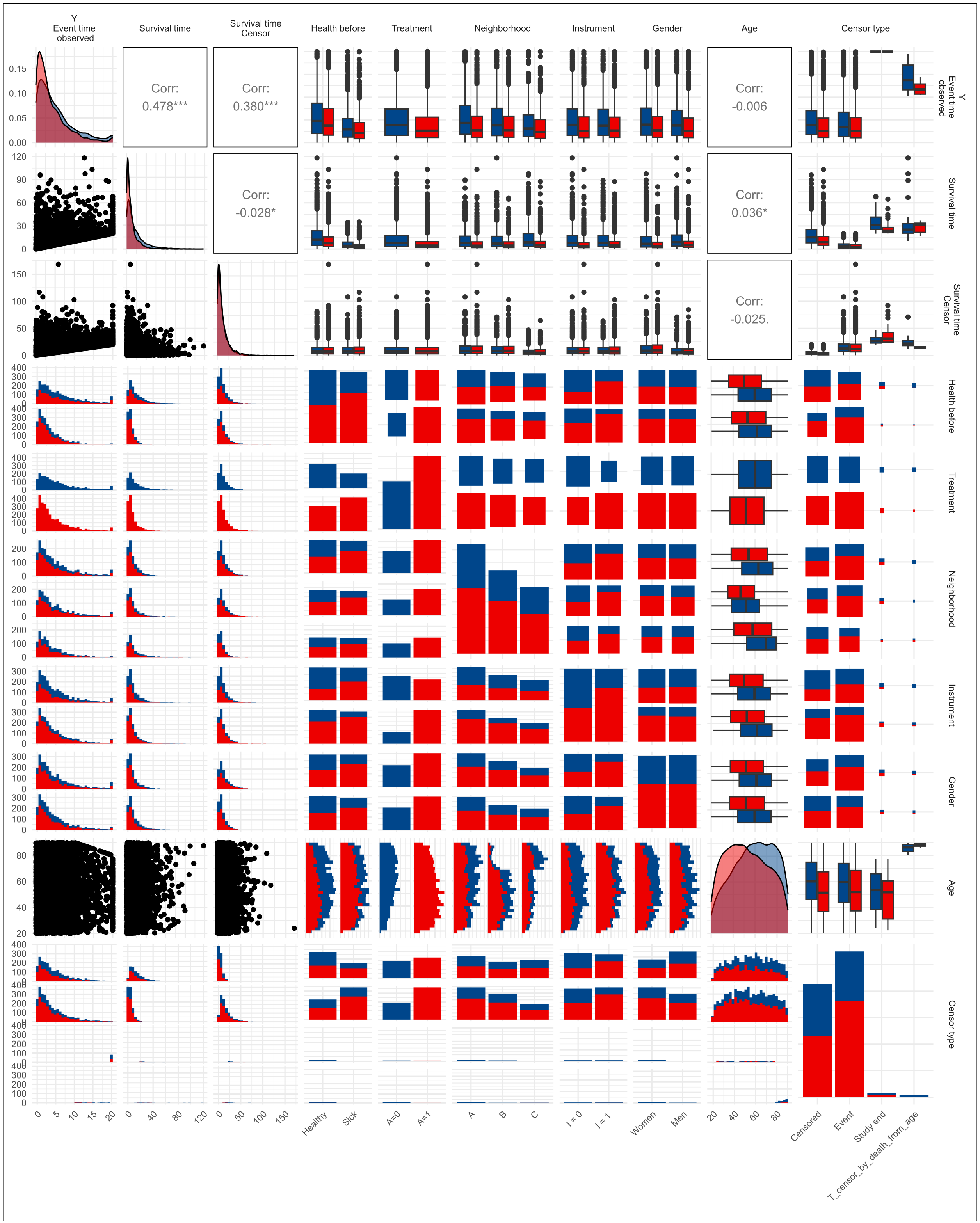


Figure 19: All covariates and relationships

3.1.3 Models Used for Estimation

Based on Figure 18 we choose parameters for all models across methods. We allow for some misspecification by including the instrument parameter in all models and assume the effect of gender to censoring to be unknown.

3.1.3.1 Cox Model For the Cox model, based on Section 2.4.1 we define the hazard function as,

$$h(t \mid A, x_1, \dots, x_5) = h_0(t) \exp\left(\beta_1 A + \beta_2 x_1 + \beta_3 x_2 + \beta_4 x_3 + \beta_5 x_4 + \beta_6 x_5\right). \tag{37}$$

All covariates that affect the outcome and treatment or censoring are included in the Cox model to adjust for confounding. The proportional hazards assumption is assumed to hold in this model and we will not inlude any time-varying effects or delve into testing the assumption further for the conditional model since that would be trivial given the data simulation process.

The goal is to estimate the accuracy of the Cox models estimation of the conditional treatment effect as β_1.

3.1.3.2 Q-model and the Parameteric g-formula The outcome model or the counterfactual Q-model is estimated based on Section 2.9 as,

$$\begin{aligned}\text{logit } \mathrm{P}\left[Y_{k+1,i} = 1 \mid Y_{k,i} = 0,\, A = a_i,\, X = x_i\right] = \beta_{0,k} + \beta_1 a_i + \beta_2 a_i k + \beta_3 a_i k^2 \\ + \beta_4 x_{i1} + \beta_5 x_{i2} + \beta_6 x_{i3} + \beta_7 x_{i4} + \beta_8 x_{i5}.\end{aligned} \tag{38}$$

We include all covarites that affect the outcome and treatment or censoring in the Q-model to adjust for confounding. To fully explore the time-varying marginal treatment effect we include time intervals k and squared time intervals k^2 with an interaction with treatment A to allow for a flexible time-varying treatment effect.

To standardize the estimation, we predict over person-time expanded format for all individuals their counterfactual potential outcome at each time interval k under both treatment and no treatment. The marginal discrete-time hazards, hazard ratio and survival curves are then estimated over each interval over the study period.

3.1.3.3 Weighted MSM The weights are defined first in Section 3.1.3.3.1 based on the definitions of inverse probability weights in Section 2.5.3.1 for the treatment weights and Section 2.8.5 for the censoring weights. The pooled logistic regression marginal structural model is defined based on the modeled weights as described in Section 2.5.3.5.

3.1.3.3.1 IPCW and IPTW The time-varying inverse probability of censoring weights (IPCW) are estimated as,

$$\text{IPCW}_i(t) = \prod_{k=1}^{T_i} \frac{P(C_i(k) = 0 \mid C_i(k-1) = 0,\ A_i = a_i)}{P(C_i(k) = 0 \mid C_i(k-1) = 0,\ A_i = a_i,\ x_{i2}, x_{i4})}. \tag{39}$$

We include age X_2 and neighborhood X_4 in the censoring model as well as treatment A and time intervals k and squared time intervals k^2 to allow for time-varying censoring.

The censoring weights had a few extreme values that skewed the estimates. In 2.5.3.1 we discuss the common practice to cut weights off at the 1st and the 99th percentile to address this issue, we used this approach to stabilize the weights.

The constant inverse probability of treatment weights (IPTW) are estimated as,

$$\text{IPTW}_i = \frac{f(A=1)}{f(A = a_i \mid x_2, x_3, x_5)}. \tag{40}$$

We include age X_2, the instrumental variable X_3 and health before X_5 in the treatment model. The inverse probability of treatment weights had a stable mean around 1 and showed no sign of needing to be stabilized.

3.1.3.3.2 Pooled Logistic Regression MSM We estimate the pooled logistic regression MSM by,

$$\text{logit } \text{P}\left[Y_{k+1,i} = 1 \mid Y_{k,i} = 0, A = a_i, X = x_i\right] = \beta_{0,k} + \beta_1 a_i + \beta_2 a_i k + \beta_3 a_i k^2 \tag{41}$$

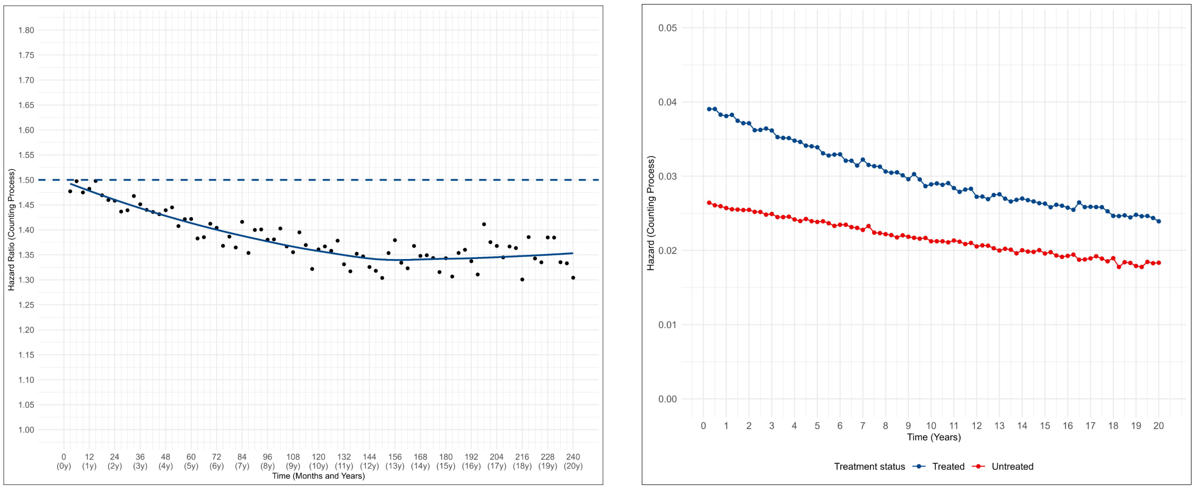
With the same time intervals k inclusion as in the Q-model. The product of the IP-weights is applied to the model and each individual is modeled in person-time format with one row per interval k until event or censoring. Each individual is modeled as a cluster to obtain robust standard errors.

The hazards and hazard ratios are then estimated by predicting the marginal counterfactual potential outcome under each treatment level for each time interval k in study-time expanded format.

3.1.3.3.3 Cox Regression MSM We also estimate the marginal treatment effect by an IP-weighted Cox MSM, using the same weights as for the pooled logistic regression MSM and the same cluster approach to get the robust standard errors. The use of the time-varying weights for the constant Cox model is described in 2.8.5. Since the Cox model does not estimate the time-varying effect only a constant marginal hazard ratio is estimated for $\beta_1 A$.

3.1.4 True Marginal Hazards and Hazard Ratios

We use the counting process approach defined in 2.2.1.1 to estimate the nonparametric discrete-time marginal hazards and the the hazard ratio over time regardless of method. To do this we expand the simulated dataset into person-time format based on the two potential survival time of treatment $A = a$ for each individual so we have one row per time interval under both treatment and no treatment. We find the hazard in each k interval as the probability of event by the ratio of the number of events d_k and the number of individuals at risk $R(k)$. The subsequent ratio between the two estimated hazards is then an estimation of the true simulated marginal hazard ratio where there is no censoring. We chose an interval length of three months for this estimation and ran one study for $N = 1.000.000$. Resulting discrete-time marginal hazard ratio in each interval k and the marginal discrete-time hazards are visualized in Figure 20.



(a) Marginal hazard ratio

(b) Marginal Hazards

Figure 20: Marginal hazards and hazard ratio for a large N

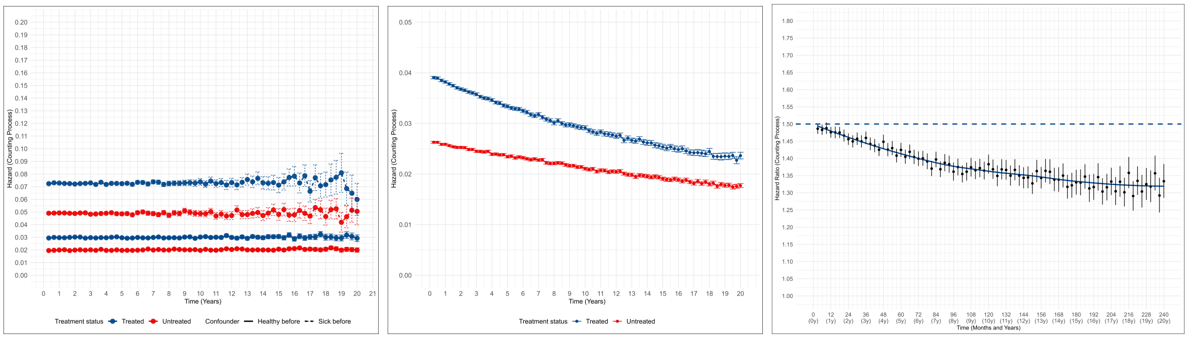
The hazard ratio in Figure 20 (a) is more unstable so we draw a smoothed line to visualize the trend over time. We are simulating data based on a true conditional hazard ratio so the marginal hazard ratio is in this case not constant over time but decreases as seen in Figure 20. The true conditional value of the simulation is set to $\exp(\beta_1) = 1.5$ and visualized by a constant dashed line but the marginal hazard ratio is not a constant or stable enough over time to reasonably estimate it and use as a comparison value for the Monte Carlo results. Therefore, we compare the Monte Carlo estimations of a conditional hazard ratio to the true hazard ratio and explore visually and by model evaluation how well the marginal hazard ratio is estimated by each method. Figure 20 surves as a reference for the marginal hazard ratio estimation.

3.1.4.1 Choice of Time Intervals k The choice of using a three month interval length for the truth estimation is done to improve the accuracy of the counting process estimation but we acknowledge that this choice is somewhat unhelpful since interval length should ideally be the same across all estimations. Rather than changing all estimation to three month intervals which would be too computationally heavy for the Monte Carlo estimation, or to change the truth estimation to four month intervals and loose accuracy and rerunning time consuming simulations we choose to keep the different interval lengths and acknowledge this as a limitation where now interval k does not align perfectly across simulations. This is a notational issue not a conceptual one since the total time is the same, we just have more intervals for the truth estimation. Unless specifically stated all other estimations use four month intervals for k.

3.1.5 Monte Carlo Simulation Results

We simulate the $M = 500$ studies with $N = 5.000$ individuals each and estimate the treatment effect, for the iterations of estimated values we choose a time length of four months as the intervals k and for the iterations of the marginal true values of the hazard ratios we use the same interval length as in Section 3.1.4 of 3 months for improved accuracy and consistency.

First we visualize in Figure 21 (c) the same metric of a true marginal hazard ratio as in Figure 20 but now as the Monte Carlo mean of the estimate with the 95% Monte Carlo confidence intervals. Using the same counting process mechanism of events divided by the number at risk we plot in Figure 21 (a) the discrete-time conditional hazards based on the observed survival times, this time including censoring and only one row per individual for the observed treatment level and in Figure 21 (b) we plot the true marginal hazards. Figure 21 (a), can be used to validate the chosen time length of four months for intervals k that allows us to stay under the required probability of event lower than 0.1 approximation of the counterfafctual approach as discussed in Section 2.5.3.3.



(a) Hazards in 4 month interv. **(b)** True marginal hazards in 3 month interv. **(c)** True discrete-time marginal HR

Figure 21: Monte Carlo mean and CI of hazards per time interval to validate using logistic regression as a hazard approximation. Figure (c) is the estimated true marginal hazard ratio.

3.1.5.1 Classical Cox Model and the Counterfactual Parametric g-formula Q-model

The estimated adjusted hazard ratios, from the classical Cox Model and the estimation of the approximated adjusted hazard ratio from the pooled logistic regression by the counterfactual Q-model are compared to each other and the true value of the conditional hazard ratio.

We report the Monte Carlo mean and the performance metrics denoted in Section 2.12 by a boxplot in Figure 22 and in Tables 5, 6.

The estimated conditional hazard ratios by the classical Cox model and the counterfactual Q-model are displayed by a boxplot in Figure 22. The dashed line indicates the true hazard ratio of treatment set in the simulation as HR = 1.5 and the boxplots show the 75th and 25th percentiles of the Monte Carlo estimates.

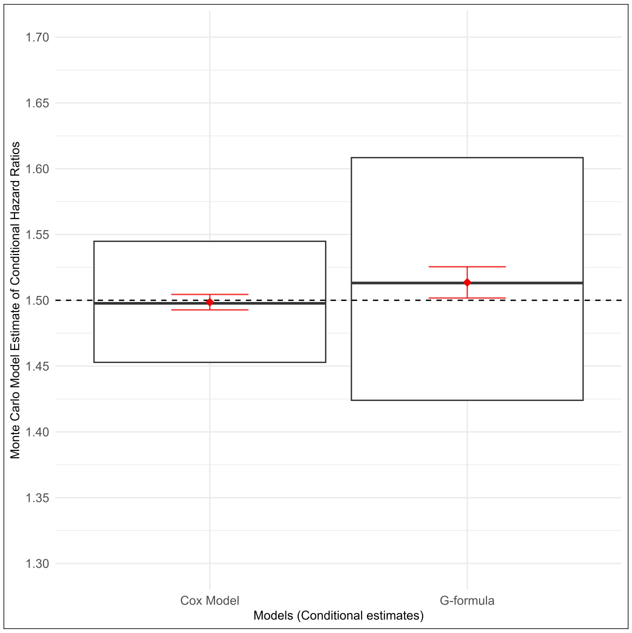


Figure 22: Adjusted hazard ratio - Model estimates

In Figure 22 we can see that both models are close to the true value of the conditional hazard ratio. The Cox model appears to be more stable since the Monte Carlo CI and the Monte Carlo percentiles are more narrow around the true value. The Cox Model appears to have close to no bias at all, while the narrow Monte Carlo CI of the Q-model (referred to as the G-formula in the figure) seem to land just above the true value. To further compare the estimates we report all main numerical results as exponentiated in Table 5 and further analyze the performance in Table 6.

Table 5: Monte Carlo model estimation of conditional treatment effect

Model	Parameter	MC E(HR)	Bias HR	MC CI 95%	MC CI width
Cox Model	age	1.000	1.000	(1.000, 1.000)	0.000
Cox Model	drug1	1.499	0.999	(1.493, 1.504)	0.012
Cox Model	genderMen	0.908	0.998	(0.905, 0.911)	0.006
Cox Model	health_beforeSick	2.508	1.003	(2.499, 2.517)	0.018
Cox Model	instrument	1.001	1.001	(0.997, 1.004)	0.007
Cox Model	neighborhoodB	0.999	0.999	(0.995, 1.002)	0.008
Cox Model	neighborhoodC	0.835	1.002	(0.831, 0.839)	0.008
Q-model	(Intercept)	0.022	-	(0.022, 0.022)	0.000
Q-model	age	1.000	1.000	(1.000, 1.000)	0.000
Q-model	drug1	1.514	1.009	(1.502, 1.525)	0.024
Q-model	drug1:time	1.000	-	(0.999, 1.001)	0.002
Q-model	drug1:timesq	1.000	-	(1.000, 1.000)	0.000
Q-model	genderMen	0.900	0.990	(0.897, 0.903)	0.006
Q-model	health_beforeSick	2.560	1.024	(2.551, 2.570)	0.019
Q-model	instrument	1.000	1.000	(0.997, 1.004)	0.007
Q-model	neighborhoodB	0.999	0.999	(0.995, 1.003)	0.008
Q-model	neighborhoodC	0.824	0.989	(0.820, 0.828)	0.008
Q-model	time	1.000	-	(0.999, 1.001)	0.001
Q-model	timesq	1.000	-	(1.000, 1.000)	0.000

The parameter drug1 in Table 5 refers to the treatment effect. Both models are very close to the true effect, small confidence intervals and little bias. The Cox model is nearly perfect with estimated bias only around 0.1% while the Q-model has a slightly larger bias of around 1%. The Q-model confidence intervals falls just above the true estimate so the bias is statistically significant but very small still.

To further compare the models we report more performance metrics in Table 6.

Table 6: Monte Carlo model estimation of conditional treatment effect - Performance and log scale

Model	Parameter	MC E(log(HR))	Bias log(HR)	EmpSE log(HR)	Reject. %	MC perc 95%	Coverage
Cox Model	age	0.000	0.000	0.001	0.054	(0.998, 1.002)	0.946
Cox Model	drug1	0.405	-0.001	0.045	1.000	(1.376, 1.637)	0.946
Cox Model	genderMen	-0.097	-0.002	0.037	0.726	(0.845, 0.983)	0.948
Cox Model	health_beforeSick	0.920	0.003	0.041	1.000	(2.316, 2.715)	0.956
Cox Model	instrument	0.001	0.001	0.040	0.044	(0.927, 1.077)	0.956
Cox Model	neighborhoodB	-0.001	-0.001	0.044	0.062	(0.923, 1.094)	0.938
Cox Model	neighborhoodC	-0.180	0.002	0.052	0.944	(0.754, 0.919)	0.952
Q-model	(Intercept)	-3.805	-	0.102	1.000	(0.018, 0.027)	-
Q-model	age	0.000	0.000	0.001	0.052	(0.998, 1.002)	0.948
Q-model	drug1	0.414	0.009	0.090	0.992	(1.268, 1.784)	0.944
Q-model	drug1:time	0.000	-	0.010	0.044	(0.981, 1.020)	-
Q-model	drug1:timesq	0.000	-	0.000	0.048	(1.000, 1.000)	-
Q-model	genderMen	-0.105	-0.010	0.038	0.780	(0.838, 0.976)	0.938
Q-model	health_beforeSick	0.940	0.024	0.042	1.000	(2.364, 2.771)	0.898
Q-model	instrument	0.000	0.000	0.041	0.044	(0.925, 1.079)	0.956
Q-model	neighborhoodB	-0.001	-0.001	0.045	0.062	(0.918, 1.097)	0.938
Q-model	neighborhoodC	-0.194	-0.011	0.053	0.956	(0.743, 0.909)	0.934
Q-model	time	0.000	-	0.008	0.042	(0.984, 1.014)	-
Q-model	timesq	0.000	-	0.000	0.042	(1.000, 1.000)	-

In Table 6 we see that the rejection % and coverage for each model is also showing very accurate and similar results. the coverage and power (rejection %) are nearly identical between the two models and the empirical standard deviation (EmpSE) is also similar and low, but even lower for the Cox model. Both models estimate the log treatment effect very well but the Cox model performs slightly better.

Since the models give such identical results creating a new simulation under more difficult to interpret conditions, like much higher censoring and reestimating if they again give close results would be of interest.

3.1.5.2 Comparison of Marginal Treatment Effect Estimation Now we compare the estimates from the Parametric g-formula (3.1.3.2), the IP weighted pooled logistic regression marginal structural model (3.1.3.3) (IP-MSM) and the IP weighted Cox marginal structural model (3.1.3.3.1) (IP-Cox-MSM)to each other and the simulated true marginal hazard ratio defined in Figure 20.

The study time period of 20 years is divided into four month intervals indexed by k and the discrete-time marginal hazard ratio is estimated in each interval by the parametric g-formula and the IP weighted pooled logistic regression MSM model (IP-MSM) as described in 3.1.3. We report the Monte Carlo mean and confidence intervals denoted in 2.12 by a faceted plot in Figure 23 and for the in IP-MSM and IP-Cox-MSM marginal parameter estimates we report them in Tables 7, 8.

Resulting discrete-time marginal hazard ratios are visualized for each four month interval k in Figure 23. The purple dots indicate the Monte Carlo mean estimate of the true, uncensored marginal hazard ratio in each three month interval and the colored lines and connecting dots are the Monte Carlo mean of the model estimates. The errorbars indicate the 95% Monte Carlo confidence intervals. The dashed line indicates the true conditional hazard ratio.

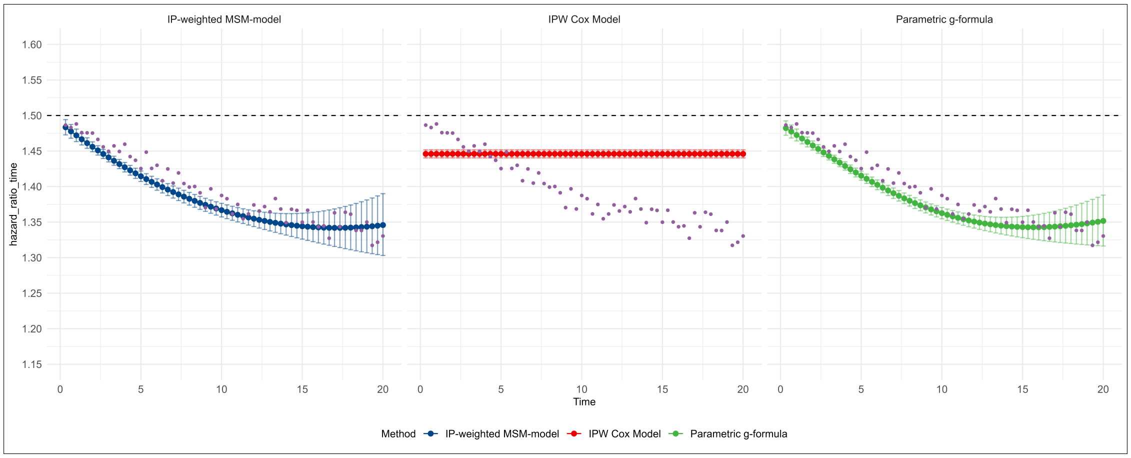


Figure 23: Monte Carlo mean of estimated marginal hazard ratios over time intervals, x axis over years

In Figure 23 the IP-MSM model and the parametric g-formula appear close to identical. The confidence intervals are slightly wider around the 20 year mark for the IP-MSM model and the first interval under all models does not have the true conditional hazard ratio in the confidence intervals. The IPW Cox MSM displays a constant marginal hazard ratio around HR ≈ 1.45. The parametric g-formula appears to have a curved up again at the end which indicates a stronger k^2 effect, we saw a possible hint of that in the large N simulation in Figure 20 (a) as well but in Figure 21 there is no indication of that and this could just be random variation.

To directly compare the parametric g-formula and MSM estimates better we plot their subsequent survival curves by (7) and visualize in Figure 24. We plot the Monte Carlo mean of the estimated marginal survival curves under treatment and no treatment with the 95% Monte Carlo percentile intervals.

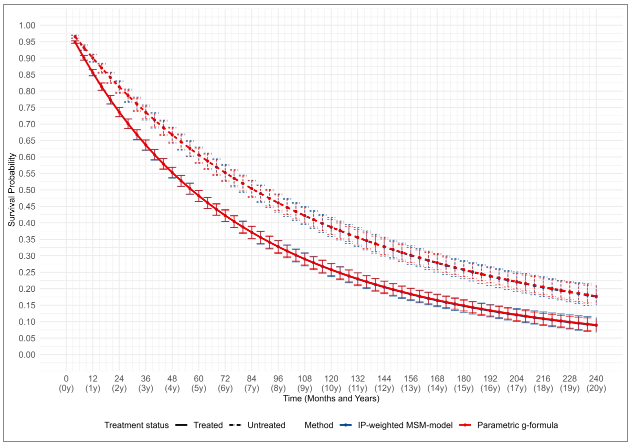


Figure 24: Marginal survival probabilities

The survival curves in Figure 24 are nearly identical for the IP-MSM and the parametric g-formula. When plotting the Monte Carlo confidence intervals they overlapped completely so we chose to display the percentiles instead for better visualization. Creating the survival curves from the IP-Cox-MSM proved too computationally heavy so we do not include them here.

Table 7: Monte Carlo mean of model estimated discrete-time marginal treatment effect

Model	Parameter	MC E(HR)	MC CI width	MC CI 95%
IP-Cox-MSM	drug1	1.446	0.011	(1.440, 1.452)
IP-MSM	(Intercept)	0.036	0.000	(0.036, 0.036)
IP-MSM	drug1	1.516	0.025	(1.503, 1.529)
IP-MSM	drug1:time	0.996	0.002	(0.995, 0.997)
IP-MSM	drug1:timesq	1.000	0.000	(1.000, 1.000)
IP-MSM	time	0.992	0.002	(0.991, 0.993)
IP-MSM	timesq	1.000	0.000	(1.000, 1.000)

The results in Table 7 confirm the visual results in Figure 23 that the IP-MSM has wider confidence intervals then the IP-Cox-MSM and and varies significantly by time interval k. The IP-Cox-MSM gives a marginal hazard ratio estimate that is lower than the other two methods are in the first few years but has narrow confidence intervals and appears to be stable and a reasonable estimate of an average of the marginal hazard ratio over time.

Table 8: Monte Carlo mean of model estimated discrete-time marginal treatment effect - Performance

Model	Parameter	MC E(log(HR))	EmpSE log(HR)	Rejection %	Cover. (bias elim.)	MC perc 95%
IP-Cox-MSM	drug1	0.369	0.045	1.000	0.958	(1.332, 1.576)
IP-MSM	(Intercept)	-3.324	0.079	1.000	0.962	(0.031, 0.042)
IP-MSM	drug1	0.416	0.094	0.990	0.956	(1.248, 1.815)
IP-MSM	drug1:time	-0.004	0.011	0.068	0.952	(0.974, 1.019)
IP-MSM	drug1:timesq	0.000	0.000	0.056	0.952	(1.000, 1.001)
IP-MSM	time	-0.008	0.009	0.124	0.966	(0.975, 1.011)
IP-MSM	timesq	0.000	0.000	0.036	0.968	(1.000, 1.000)

In Table 8 we see that the IP-MSM and IP-Cox-MSM model have similarly high rejection % of the null hypothesis of no effect and a bias eliminated coverage of approximately 95% for the treatment effect parameter drug1. The IP-MSM model has much wider Monte Carlo percentile intervals and a high empirical standard deviation (EmpSE) of the log hazard ratio estimate. The IP-Cox-MSM model has a low empirical standard deviation and a narrow percentile intervals but since it estimates a constant marginal hazard ratio over time it is not directly comparable. The variability of the IP-MSM model over time is captured aswell by the wide percentile intervals for time parameter k and k^2.

To further explore the differences between a classical Cox model and a counterfactual IP-Cox-MSM we explore a similar observational study in Section 3.2. We fit the models under various combinations of model covariates and thus a range of misspecified models to see how well the models perform under poorly chosen model parameters.

3.2 Covariate Combinations for Models

In Section 3.1.1 we chose which covariates to model on based on the causal assumption we stated by Figure 18. The choice of parameters controls how good the subsequent estimate of the hazard ratio can be. We allowed for some possible model misspecification and rationalized that for a fair comparison we choose one set up and use for all models. The comparison is then between the methods but for that specific covariate combination.

To further analyze the Cox proportional hazards model under the classical framework and the IP weighted Cox MSM of the counterfactual framework we estimate the treatment effect for each model under all possible combinations of covariates. This allows us to see how well the models perform under misspecification and how forgiving they are to the choice of covariates included in the model.

3.2.1 Binomial Theorem

The binomial theorem is used to find the number of possible combinations of covariates in the models. The binomial coefficient is denoted as,

$$\binom{n}{x} = \frac{n!}{x!(n-x)!}$$

.

And the binomial theorem is,

$$(x+y)^r = \sum_k \binom{r}{k} x^k y^{r-k}, \qquad \text{integer } r \geqslant 0. \quad \text{ or } |x/y| < 1. \tag{42}$$

When $x = y = 1$ we have 2^r as the number of possibilities. We use this to compute the total number of combinations for each set up (Bhavanari Satyanarayana, 2019).

Based on Table 3, the number of possible combinations of covariates in our study is $x_1, \dots, x_6$, $2^{|X|} = 64$.

3.2.2 Data and Modeling Scenario Combinatorix

We briefly summarize below how quickly the computations can grow if the goal is to model all covariate combinations, all simulation combinations and censoring mechanisms. The total number of scenarios is the product of the number of data simulations, models and censoring mechanisms where we assume one out of four mechanisms are possible.

Data simulation	Models	Censoring	Total scenarios
$2^6 = 64$	$2^6 = 64$	$4 \times 4 = 16$	$64 \times 64 \times 16 = 65\,536$

Exploring all combinations across simulations and models would become difficult to manage or interpret so we choose to focus on one simulation set up, and censoring mechanism and explore all covariate combinations for that set up only. Further research could explore more simulations and censoring mechanisms.

3.2.3 Simulation Set Up

We use the same simulation set up as in Section 3.1.1 but we simplify the censoring mechanism to be completely at random. We make gender a stronger predictor of outcome to increase the possibility of collider bias when conditioning on income, defined as a collider in Figure 25. We choose to keep the true conditional hazard ratio at 1.5 for comparable results to Section 3.1.1 but we lower the hazards by setting the baseline hazards to $\lambda_0 = 1/45$ and $\lambda_1 = 1/30$ for untreated and treated respectively. Since we keep the time of study at 20 years this will increase administrative censoring significantly from the previous simulation.

The causal relationships are visualized in the DAG in Figure 25.

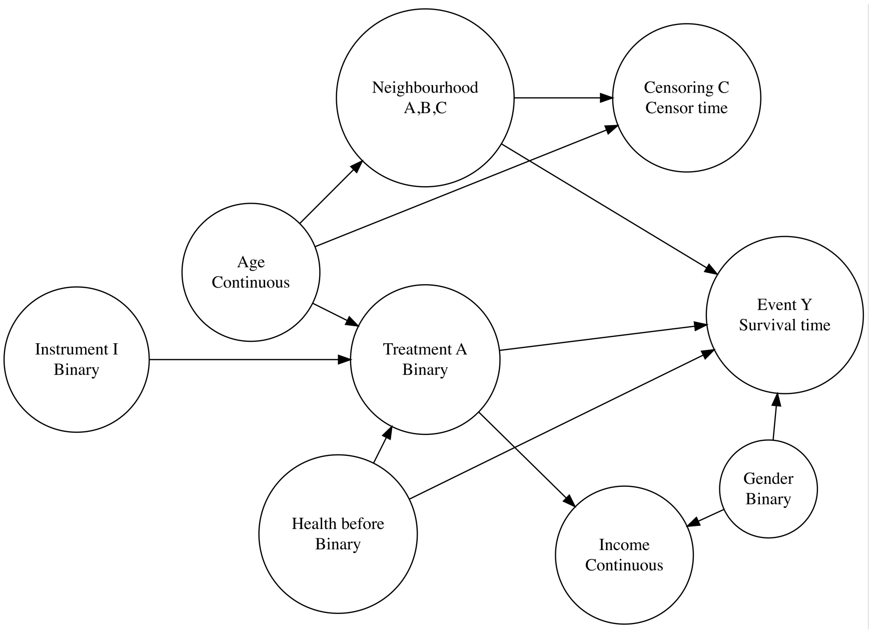


Figure 25: DAG of causal relationships for combinations simulation

Income becomes a collider in Figure 25 since it is affected by both treatment and gender which also affects the outcome. Notably, income is commonly used as a confounder in observationsal studies but in this study over 20 years we assume that treatment could have an effect on an individual's income rather than the other way around.

The Monte Carlo simulation study parameters are summarized in Table 10.

Table 10: Parameters of the Monte Carlo simulation study of covariate combinations

Notation	Value	Description
Study design		
N	1 000	Number of individuals
M	400	Number of Monte Carlo iterations
Treatment assignment		
α_0	$p = (0.5)$	Baseline probability of treatment, Bernoulli
α_1	3	Log odds ratio for health before variable
α_2	4	Log odds ratio for the instrumental variable
α_3	0.98	Log odds ratio for unit increase in age
Covariate distributions		
X_{age}	$\mathcal{TN}(55, 35, 20, 90)$	Age (years), truncated normal
X_{income}	$\mathcal{N}(35 + 10 * A + 5 * \text{gender}, 5) \times \mathcal{U}(0.95, 1.05)$	Income, normal
$p_{\text{health before}}$	$p = (0.5)$	Health before, Bernoulli
p_{gender}	$p = (0.5)$	Gender, Bernoulli
$p_{\text{neighborhood}}$	$p = (0.4,\ 0.4,\ 0.2)$	$\Pr\big(X_{\text{neigh}} = 1, 2, 3 \mid \text{age} < 70\big)$
$p_{\text{neigh\|senior}}$	$p = (0.4,\ 0.3,\ 0.3)$	$\Pr\big(X_{\text{neigh}} = 1, 2, 3 \mid \text{age} \geq 70\big)$
Survival time simulation		
λ_0	1/45	Baseline hazard in untreated T^0 (Years)
λ_1	1/30	Baseline hazard in treated T^1 (Years)
$\exp\big(\beta_{\text{gender}}\big)$	0.7692	HR for men vs. women
$\exp\big(\beta_{\text{neighborhood}}\big)$	0.833	HR for neighborhood C vs. others
$\exp\ \big(\beta_{\text{health}}\big)$	2.5	HR for sick vs. healthy
$\exp\big(\beta_{\text{age}}\big)$	1	HR per unit increase in age
Censoring model		
T_c^{age}	100	Administrative censoring age (years)
T_c^{admin}	T_{study}	Administrative censoring time (years)
λ_c	$0.4\lambda_1$	Baseline censoring hazard

In Figure 26 we visualize the simulated observational study.

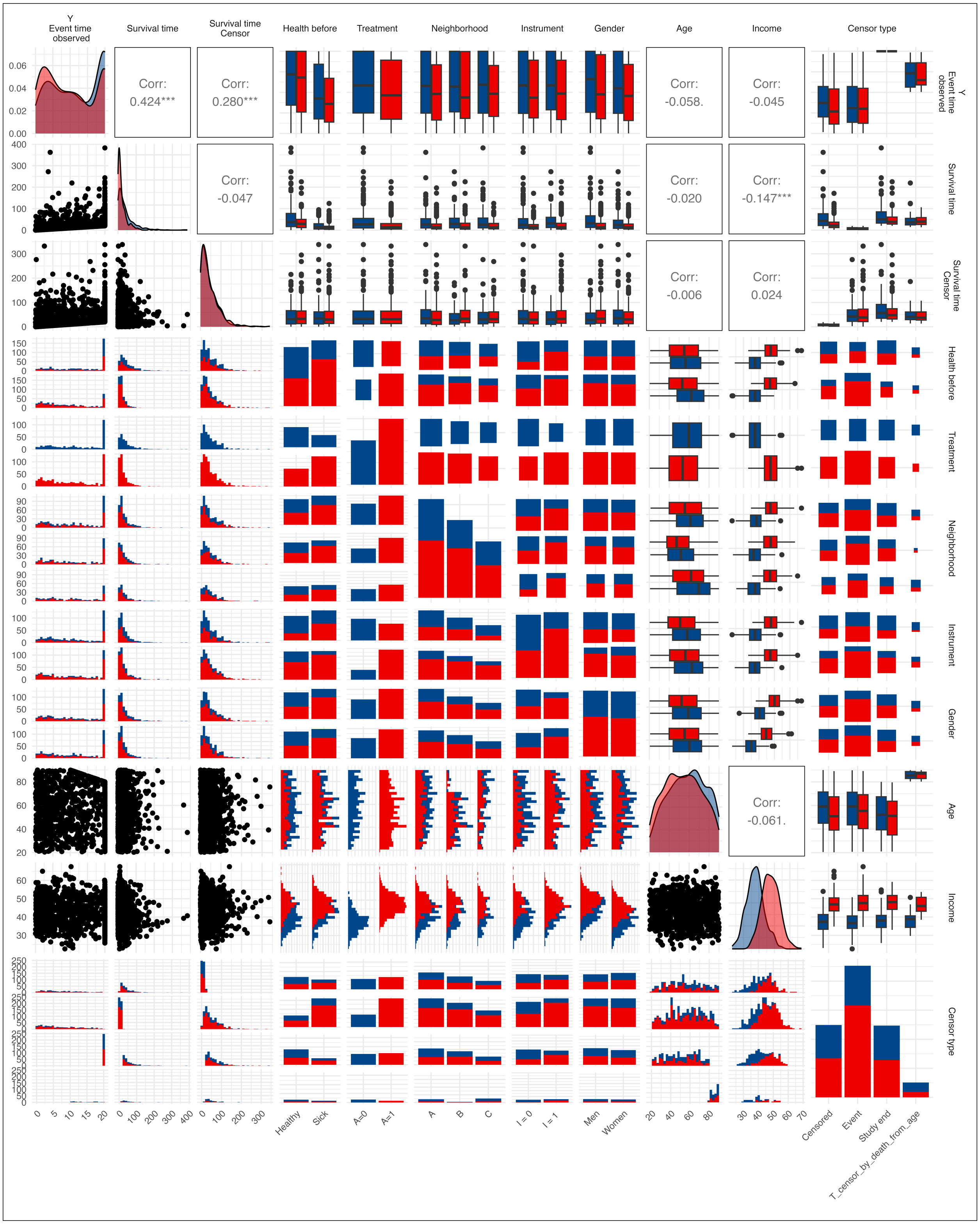


Figure 26: All covariates and relationships

3.2.4 Model Combinations and Results

We use the Cox model defined in Section 3.1.3.1 and the IP-weighted Cox MSM model defined in Section 3.1.3.3 but now fit across all possible combinations of covariates X denoted in Table 3. Since the censoring is now completely at random we do not need to model the censoring mechanism for the IP weights and can rely on a logistic regression model for treatment probability.

The results are explored visually in Figures 27 and 28 for the Cox model and the IP weighted Cox MSM respectively. The combinations numbers correspond to the combinations stated in Tables 11 and 12 below. The boxplots show the 75th and 25th percentiles of the Monte Carlo mean estimates and the 95% Monte Carlo confidence intervals as error bars with the mean dot in the middle. The true conditional value of the simulation is set to $\exp(\beta_1) = 1.5$ and displayed by a dashed red line. The combination numbers on the x-axis appear a little further to the left of the boxplot they correspond to, if this causes confusion please refer to the tables 11, 12 for clarity.

The estimates in Figure 27 can serve as a reference point for the marginal estimates in Figure 28 since the combination numbers correspond to the same covariate sets in both figures.

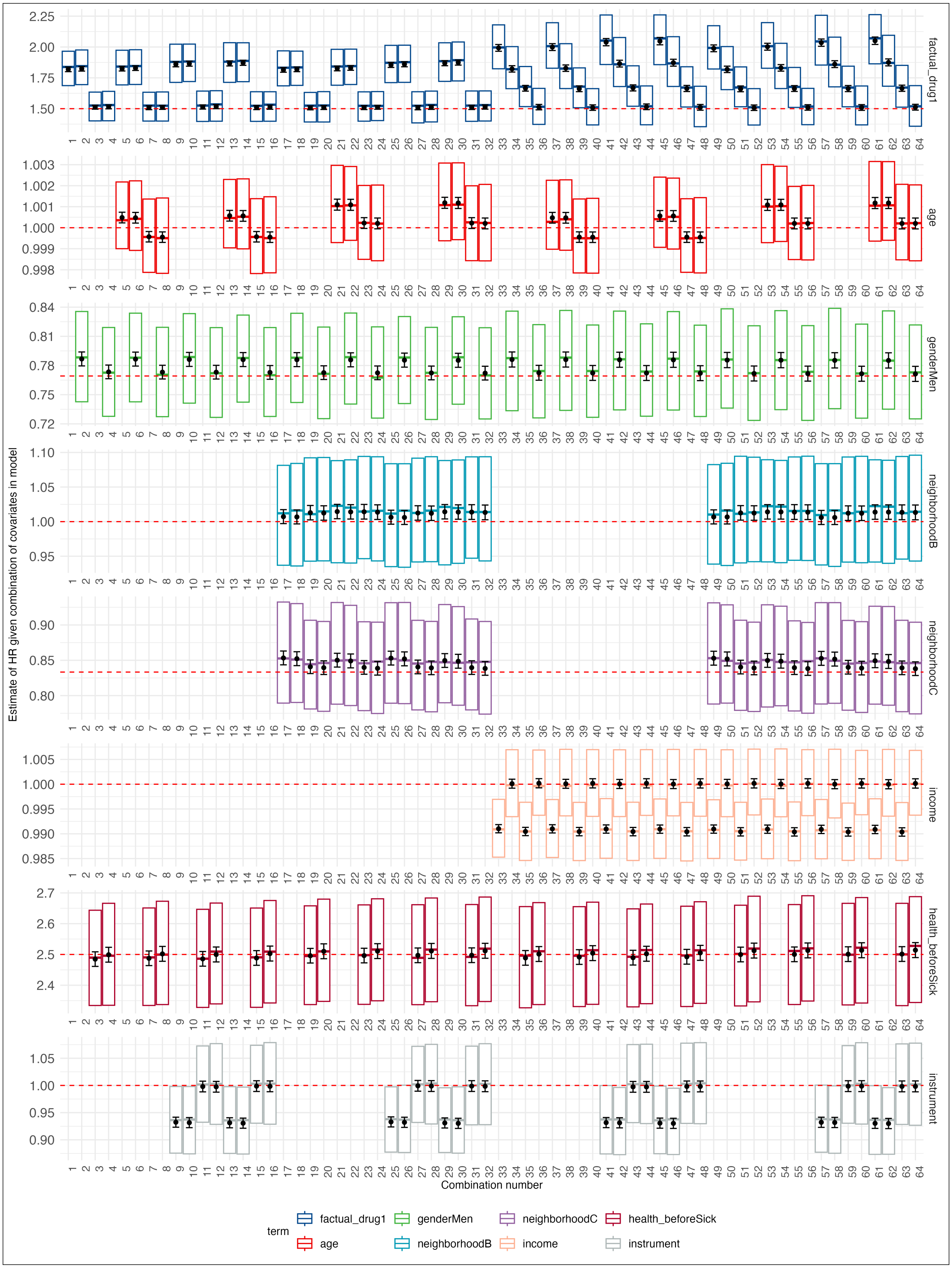


Figure 27: Cox regression hazard ratio estimates

The inclusion of neighborhood in Figure 27 appears to have little effect on the treatment effect estimate and the effect of neighborhood itself is consistently well estimated across combinations. Including income skewes the treatment effect estimate but that can be adjusted for by including gender aswell. the health before variable is necessary to include for a valid estimate of the treatment effect as expected for a strong confounder. No clear pattern appears for the inclusion of age or more interestingly for the instrumental variable. Since Figure 27 proved hard to read down to all 64 combinations we picked two combinations to highlight the effect of including or excluding certain covariates and plot the Monte Carlo 25th and 75th percentiles in Figure 30, this plot is included in the appendix.

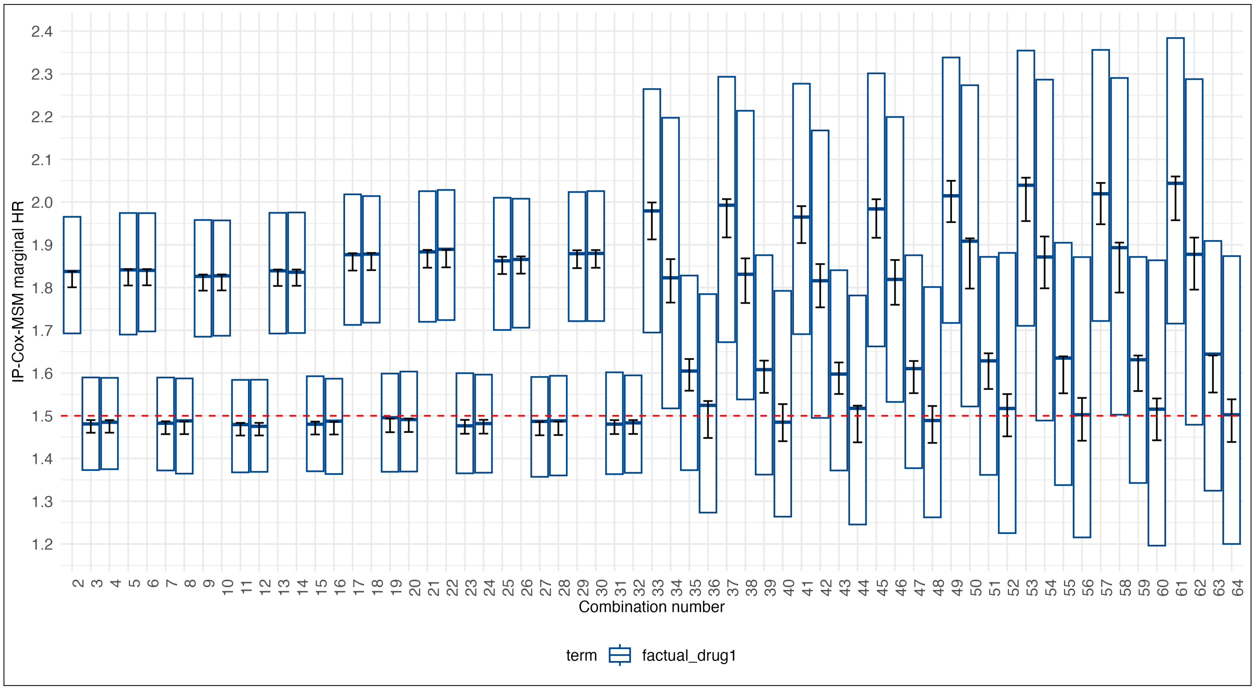


Figure 28: IP weighted Cox MSM hazard ratio estimates

Figure 28 shows that the IP weighted Cox MSM estimates vary a lot based on what combination of covariates is used. As for the Cox model in Figure 27 including health before is necessary for a valid estimate of the treatment effect. Including income without gender, gives very high estimates from combination 30 to 64 but including gender fixes that issue a bit. Even though we have added the dashed line of the true conditional hazard ratio as a reference, the MSM model is estimating a marginal hazard ratio. Based on the consistency of estimates close to $HR \approx 1..475$ we can assume that the marginal hazard ratio is close to that value for this simulation set up. Notably the inclusion of the instrument appears to have little if any, biasing effect on the estimates.

Overview of the fitted models can be found in the appendix in tables 11, 12 down to the combination number per model that links back to the plots of results.

4 Conclusion

Goal of the simulations was to compare the two theoretically equivalent methods of IP weighting and standardization by the parametric g-formula and to also compare them to the classical Cox model. The results show that the three methods produce very similar results for the specified models as long as the desired estimand is clearly defined as conditional, marginal or marginal and time-varying.

The IP-MSM model seems to be the least stable and further work on this subject could look into how the methods compare under close to broken positivity assumptions or under scenarios with extreme weights where more work would be needed to stabilize the weights. The ability of the model to model the time-varying hazards is a strength but if the model is too unstable to give precise estimates that strength is more so a weakness and a stable Cox model or the IP-Cox-MSM could be preferred if an interpretation of an average marginal hazard ratio is enough given the research question.

The parametric g-formula appears to be a strong estimator of both conditional and marginal hazard ratios under our chosen set ups. The limitation it has by being very computationally demanding does on the other hand require it to perform at least as well as the IP-MSM model since it requires more work to implement.

We explored and discussed in Section 2.6.4 how for a study with a strong positive effect of treatment, we expect shorter survival times for treated individuals and the same for a strong positive effect from a confounder variable, this combination creates a clear time-varying marginal hazard ratio as the risk set changes over time. Figure 29 gives an insight into how the discrete-time marginal hazard ratio behaves over time given other combinations of confounder and treatment effects and we can see that the combination we have explored is the most extreme case of this effect. Based on these results, it could be a better estimate of the marginal treatment effect over time to use the stable Cox model or the IP-weighted Cox model rather than to assume and model a time-varying treatment effect where there is none.

Further work could aswell delve into the power and accuracy of the Schoenfeld residuals test which we have mostly left out of this thesis.

4.1 Limitations

The expanded dataset framework that the counterfactual models rely on, especially the parametric g-formula are computationally extremely heavy. Running simulations or using real world data requires computational power and code built for speed. We had to limit this work to what could be accomplished within reason and for more extensive research on the parametric g-formula heavier computing power would be needed. The same was the case for the combination of covariates set up, running on many covariates becomes infeasible quickly.

Due to the vast material we had to limit ourselves to not dive into effect modifiers in the simulations which should have proven interesting but is left for further work. The contrasts of a marginal or a conditional estimate can become interestingly complicated with effect modifiers that can make a marginal treatment effect appear as zero but with strong conditional treatment effects.

The build up of simulated covariates happened over a period of time which caused some loss of bigger picture elements, using gender as a binary covariate with such a strong effect on income is more outdated than the simulation study should have been set as.

5 Appendix

5.1 G-estimation by Structural Nested Models

The third method within the causal framework is G-estimation by structural nested mean models, SNM for short.

Instead of estimating the marginal mean, SNM models estimate the conditional mean within levels of A and selected covariates. SNMs are agnostic about the intercept β_0 and the main effect of the mixed effect $V \in X$, making them semi-parametric. This is equivalent to removing β_0, β_3 from a mixed effect MSM model resulting in, $\mathrm{E}\left[Y^a - Y^{a=0} \mid A = a, X\right] = \beta_1 a + \beta_2 aX$.

The SNM model is therefore more similar to the traditional Cox model than to other causal models. It is conditional, non-collapsible, and not recommended when dealing with time-varying treatments (Hernán & Robins, 2025). Since they are rarely the focus in causal inference research, they are not discussed further in this work.

5.2 Treatment and Confounding Effects on a Marginal HR

In Section 2.6.4 we displayed Figure 14 for a single simulated study and here we extend the same study to 100 iterations and estimate the Monte Carlo mean estimates and confidence intervals. The confidence intervals are denoted in Section 2.12.

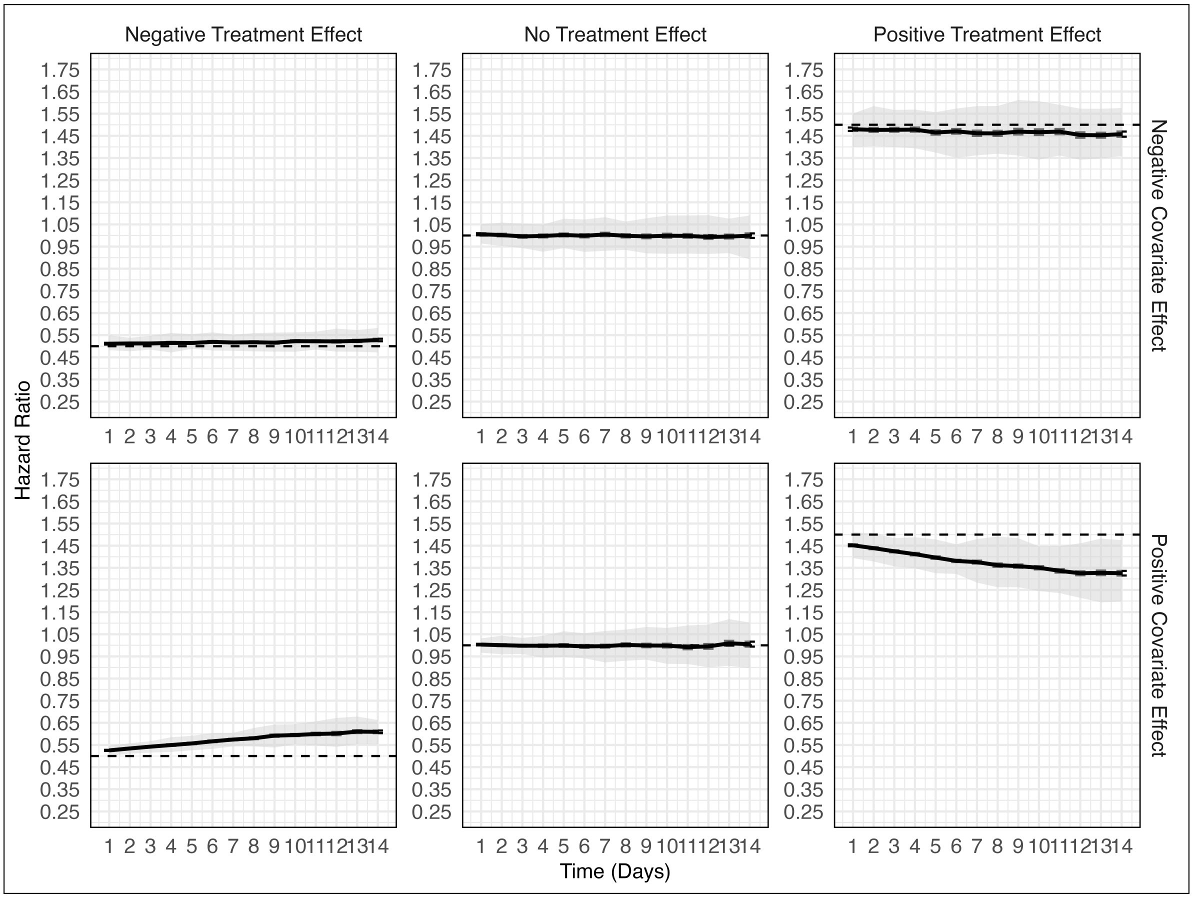


Figure 29: Hazard ratio over time given pos, neg or no effect of treatment or other measures.

The Figure 29 shows a Monte Carlo simulation of the marginalm discrete-time hazard ratio for various combinations of treatment and confounding effects, as described in Section 2.6.4. The 95% quantile intervals are shown as shaded areas around the estimated Monte Carlo mean hazard ratios. These are based on 100 simulated studies for each scenario and the Monte Carlo confidence intervals are calculated over the 100 studies, the intervals became very tight so they almost dissapear.

5.3 Good and Bad Covariate Combinations

Combinations 25 and 59 from Section 3.2.

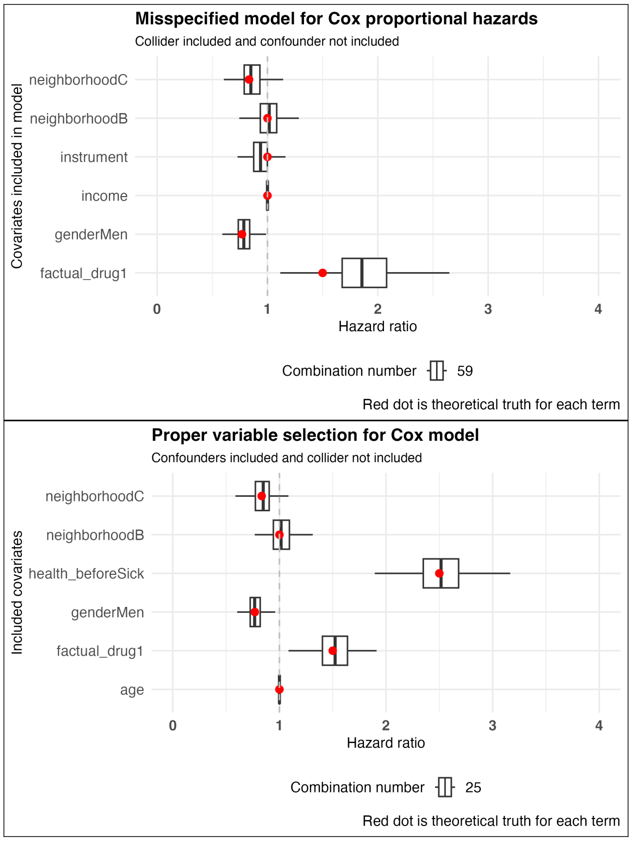


Figure 30: Cox regression hazard ratio estimates examples

Bibliography

Agresti, A. (2013). *Categorical data analysis*. Wiley. https://books.google.se/books?id=6PHHE1Cr44AC

Austin, P. C. (2012). Generating survival times to simulate cox proportional hazards models with time-varying covariates. *Statistics in Medicine*, *31*(29), 3946–3958. https://doi.org/10.1002/sim.5452

Austin, P. C., & Giardiello, D. (2025). The impact of violation of the proportional hazards assumption on the calibration of the cox proportional hazards model. *Statistics in Medicine*, *44*(13–14), e70161. https://doi.org/10.1002/sim.70161

Batyrbekova, N., Bower, H., Dickman, P. W., Szulkin, R., Lambert, P. C., & Andersson, T. M.-L. (2024). Potential bias introduced by not including multiple time-scales in survival analysis: A simulation study. *Communications in Statistics – Simulation and Computation*, *53*(2), 993–1006. https://doi.org/10.1080/03610918.2022.2038626

Bhavanari Satyanarayana, S. M. S., T. V. Pradeep Kumar. (2019). *Mathematical foundations of computer science*. CRC Press.

Burkardt, J. (2023). *The truncated normal distribution*. Department of Scientific Computing, Florida State University. https://people.sc.fsu.edu/~jburkardt/presentations/truncated_normal.pdf

Chatton, A., Le Borgne, F., Leyrat, C., Gillaizeau, F., Rousseau, C., Barbin, L., Laplaud, D., Léger, M., Giraudeau, B., & Foucher, Y. (2020). G-computation, propensity-score-based methods, and targeted maximum likelihood estimator for causal inference with different covariates sets: A comparative simulation study. *Scientific Reports*, *10*, 9219. https://doi.org/10.1038/s41598-020-65917-x

Cole, S. R., & Hernán, M. A. (2004). Adjusted survival curves with inverse probability weights. *Computer Methods and Programs in Biomedicine*, *75*(1), 45–49. https://doi.org/10.1016/j.cmpb.2003.10.004

Cole, S. R., & Hernán, M. A. (2008). Constructing inverse probability weights for marginal structural models. *American Journal of Epidemiology*, *168*(6), 656–664. https://doi.org/10.1093/aje/kwn164

Daniel, R., Zhang, J., & Farewell, D. (2021). Making apples from oranges: Comparing noncollapsible effect estimators and their standard errors after adjustment for different covariate sets. *Biometrical Journal*, *63*(3), 528–557. https://doi.org/10.1002/bimj.201900297

David W. Hosmer, S. L., Jr. (2011). *Applied survival analysis*. John Wiley & Sons.

Digitale, J. C., Martin, J. N., & Glymour, M. M. (2022). Tutorial on directed acyclic graphs. *Journal of Clinical Epidemiology*, *142*, 264–267. https://doi.org/https://doi.org/10.1016/j.jclinepi.2021.08.001

Dumas, E., & Stensrud, M. J. (2025). How hazard ratios can mislead and why it matters in practice. *European Journal of Epidemiology*, *40*(6), 603–609. https://doi.org/10.1007/s10654-025-01250-9

Gerds, T. A., Kattan, M. W., Schumacher, M., & Yu, C. (2013). Estimating a time-dependent concordance index for survival prediction models with covariate dependent censoring. *Statistics* *in Medicine*, *32*(13), 2173–2184. https://doi.org/10.1002/sim.5681

Greenland, S. (1996). Absence of confounding does not correspond to collapsibility of the rate ratio or rate difference. *Epidemiology*, *7*(5), 498–501. https://pubmed.ncbi.nlm.nih.gov/8862980/

Greenland, S., Robins, J. M., & Pearl, J. (1999). Confounding and collapsibility in causal inference. *Statistical Science*, *14*(1), 29–46. https://doi.org/10.1214/ss/1009211805

Helen Bian, G. W., Menglan Pang. (2024). Non-collapsibility and built-in selection bias of period-specific and conventional hazard ratio in randomized controlled trials. *BMC Medical Research Methodology*, *24*(1). https://doi.org/10.1186/s12874-024-02402-3

Hernan, M. A., Brumback, B., & Robins, J. M. (2000). Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. *Epidemiology*, *11*(5), 561–570. https://doi.org/10.1097/00001648-200009000-00012

Hernán, M. A. (2010). The hazards of hazard ratios. *Epidemiology*, *21*(1), 13–15. https://doi.org/10.1097/ede.0b013e3181c1ea43

Hernán, M. A. (2017). Invited commentary: Selection bias without colliders. *American Journal of Epidemiology*, *185*(11), 1048–1050. https://doi.org/10.1093/aje/kwx077

Hernán, M. A. (2022). Causal analyses of existing databases: No power calculations required. *Journal of Clinical Epidemiology*, *144*, 203–205. https://doi.org/https://doi.org/10.1016/j.jclinepi.2021.08.028

Hernán, M. A., & Robins, J. M. (2006). Estimating causal effects from epidemiological data. *Journal of Epidemiology & Community Health*, *60*(7), 578–586. https://pmc.ncbi.nlm.nih.gov/articles/PMC2652882/

Hernán, M. A., & Robins, J. M. (2025). *Causal inference: What if (the book) by miguel hernán and jamie robins*. https://miguelhernan.org/whatifbook.

Humphreys, A. B. C., Matthews, A. A., Young, J. C., Berglund, A., Lindahl, B., Wettermark, B., Dahabreh, I. J., Kahan, T., & Hernán, M. A. (2025). The definition of treatment assignment in observational emulations of target trials – an empirical examination in the swedish primary care cardiovascular database. *Annals of Epidemiology*, *108*, 56–62. https://doi.org/https://doi.org/10.1016/j.annepidem.2025.06.003

Igelström, E., Craig, P., Lewsey, J., Lynch, J., Pearce, A., & Katikireddi, S. V. (2022). Causal inference and effect estimation using observational data. *Journal of Epidemiology and Community Health*, *76*(11), 960–966. https://doi.org/10.1136/jech-2022-219267

Iketle Aretha Maharela, D.-G. C., Lizelle Fletcher. (2024). Modified cox models: A simulation study on different survival distributions, censoring rates, and sample sizes. *Mathematics*, *12*(18), 2903. https://doi.org/10.3390/math12182903

Kennedy, T. (2016). Basics of direct monte carlo. In *Monte carlo methods — a special topics course* (pp. 11–20). https://math.arizona.edu/~tgk/mc/book_chap2.pdf

Laura Pazzagli, M. Z., Marie Linder. (2018). Methods for time-varying exposure related problems in pharmacoepidemiology: An overview. *Pharmacoepidemiology and Drug Safety*, *27*(2), 148–160. https://doi.org/10.1002/pds.4372

Li, P., & Redden, D. T. (2015). Small sample performance of bias-corrected sandwich estimators for cluster-randomized trials with binary outcomes. *Statistics in Medicine*, *34*(2), 281–296.

https://doi.org/10.1002/sim.6344

Lousdal, M. L. (2018). An introduction to instrumental variable assumptions, validation and estimation. *Emerging Themes in Epidemiology*, *15*(1). https://doi.org/10.1186/s12982-018-0069-7

Majeed, A. F. (2020). *Accelerated failure time models: An application in insurance attrition.* HAL Open Science post-print. https://univ-pau.hal.science/hal-02953269v1

Mats J. Stensrud, M. A. H. (2020). Why test for proportional hazards? *JAMA*, *323*(14), 1401. https://doi.org/10.1001/jama.2020.1267

Michele Jonsson Funk, C. W., Daniel Westreich. (2011). Doubly robust estimation of causal effects. *American Journal of Epidemiology*, *173*(7), 761–767. https://doi.org/10.1093/aje/kwq439

Mitchell H. Katz, W. W. H. (1993). Proportional hazards (cox) regression. *Journal of General Internal Medicine*, *8*(12), 702–711. https://doi.org/10.1007/bf02598295

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

Odd Aalen, H. G., Ornulf Borgan. (2008). *Survival and event history analysis.* Springer Science & Business Media.

Patricia M. Grambsch, T. M. T. (1994). Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika*, *81*(3), 515–526. https://doi.org/10.1093/biomet/81.3.515

Patrick J. Heagerty, Y. Z. (2005). Survival model predictive accuracy and ROC curves. *Biometrics*, *61*(1), 92–105. https://doi.org/10.1111/j.0006-341x.2005.030814.x

Post, R. A. J., Heuvel, E. R. van den, & Putter, H. (2024). The built-in selection bias of hazard ratios formalized using structural causal models. *Lifetime Data Analysis*, *30*, 404–438. https://doi.org/10.1007/s10985-024-09617-y

Ralf Bender, M. B., Thomas Augustin. (2005). Generating survival times to simulate cox proportional hazards models. *Statistics in Medicine*, *24*(11), 1713–1723. https://doi.org/10.1002/sim.2059

Robertson, S. E., Steingrimsson, J. A., & Dahabreh, I. J. (2022). Using numerical methods to design simulations: Revisiting the balancing intercept. *American Journal of Epidemiology*, *191*(7), 1283–1289. https://doi.org/10.1093/aje/kwab264

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), 1393–1512. https://doi.org/https://doi.org/10.1016/0270-0255(86)90088-6

Robins, J. M. (1997). Causal inference from complex longitudinal data. In M. Berkane (Ed.), *Latent variable modeling and applications to causality* (Vol. 120, pp. 69–117). Springer. https://doi.org/10.1007/978-1-4612-1842-5_4

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

ROSENBAUM, P. R., & RUBIN, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41–55. https://doi.org/10.1093/biomet/7

0.1.41

Royston, P. (2015). Tools for checking calibration of a cox model in external validation: Prediction of population-averaged survival curves based on risk groups. *The Stata Journal: Promoting Communications on Statistics and Stata*, *15*(1), 275–291. https://doi.org/10.1177/1536867x1501500116

Schoenfeld, D. (1980). Chi-squared goodness-of-fit tests for the proportional hazards regression model. *Biometrika*, *67*(1), 145–153. https://doi.org/10.1093/biomet/67.1.145

Schoenfeld, D. (1982). Partial residuals for the proportional hazards regression model. *Biometrika*, *69*(1), 239–241. https://doi.org/10.1093/biomet/69.1.239

Stensrud, M. J., & Hernán, M. A. (2025). Why use methods that require proportional hazards? *American Journal of Epidemiology*, *194*(6), 1504–1506. https://doi.org/10.1093/aje/kwae361

Taubman, S. L., Robins, J. M., Mittleman, M. A., & Hernán, M. A. (2009). Intervening on risk factors for coronary heart disease: An application of the parametric g-formula. *International Journal of Epidemiology*, *38*(6), 1599–1611. https://doi.org/10.1093/ije/dyp192

Terry M. Therneau, P. M. G. (2000). *Modeling survival data: Extending the cox model*. Springer Science & Business Media.

Turchin, A., Petito, L. C., Hegermiller, E., Carnahan, R., DeVries, A., Goel, S., Lansang, M. C., McDonnell, M. E., Nair, V., Priest, E., Willey, V. J., Kaul, A. F., & Hernán, M. A. (2025). Cardiovascular events in individuals treated with sulfonylureas or dipeptidyl peptidase 4 inhibitors. *JAMA Network Open*, *8*(7), e2523067. https://doi.org/10.1001/jamanetworkopen.2025.23067

van Dijk, P. C., Jager, K. J., Zwinderman, A. H., Zoccali, C., & Dekker, F. W. (2008). The analysis of survival data in nephrology: Basic concepts and methods of cox regression. *Kidney International*, *74*(6), 705–709. https://doi.org/https://doi.org/10.1038/ki.2008.294

Wang, X., Turner, E. L., & Li, F. (2023). Improving sandwich variance estimation for marginal cox analysis of cluster randomized trials. *Biometrical Journal*, *65*(3), e2200113. https://doi.org/10.1002/bimj.202200113

Westreich, D., Cole, S. R., Young, J. G., Palella, F., Tien, P. C., Kingsley, L., Gange, S. J., & Hernán, M. A. (2012). The parametric g-formula to estimate the effect of highly active antiretroviral therapy on incident AIDS or death. *Statistics in Medicine*, *31*(18), 2000–2009. https://doi.org/10.1002/sim.5316

Williamson, T., & Ravani, P. (2017). Marginal structural models in clinical research: When and how to use them? *Nephrology Dialysis Transplantation*, *32*(suppl_2), ii84–ii90. https://doi.org/10.1093/ndt/gfw341

Young, J. G., Cain, L. E., Robins, J. M., O'Reilly, E. J., & Hernán, M. A. (2011). Comparative effectiveness of dynamic treatment regimes: An application of the parametric g-formula. *Statistics in Biosciences*, *3*(1), 119–143. https://doi.org/10.1007/s12561-011-9040-7

Table 11: Used models for IP-weights and Cox-MSM models

IPW formula	Combination number
factual_drug ~ 1	1
factual_drug ~ gender	2
factual_drug ~ health_before	3
factual_drug ~ gender + health_before	4
factual_drug ~ age	5
factual_drug ~ gender + age	6
factual_drug ~ health_before + age	7
factual_drug ~ gender + health_before + age	8
factual_drug ~ neighborhood	9
factual_drug ~ gender + neighborhood	10
factual_drug ~ health_before + neighborhood	11
factual_drug ~ gender + health_before + neighborhood	12
factual_drug ~ age + neighborhood	13
factual_drug ~ gender + age + neighborhood	14
factual_drug ~ health_before + age + neighborhood	15
factual_drug ~ gender + health_before + age + neighborhood	16
factual_drug ~ instrument	17
factual_drug ~ gender + instrument	18
factual_drug ~ health_before + instrument	19
factual_drug ~ gender + health_before + instrument	20
factual_drug ~ age + instrument	21
factual_drug ~ gender + age + instrument	22
factual_drug ~ health_before + age + instrument	23
factual_drug ~ gender + health_before + age + instrument	24
factual_drug ~ neighborhood + instrument	25
factual_drug ~ gender + neighborhood + instrument	26
factual_drug ~ health_before + neighborhood + instrument	27
factual_drug ~ gender + health_before + neighborhood + instrument	28
factual_drug ~ age + neighborhood + instrument	29
factual_drug ~ gender + age + neighborhood + instrument	30
factual_drug ~ health_before + age + neighborhood + instrument	31
factual_drug ~ gender + health_before + age + neighborhood + instrument	32
factual_drug ~ income	33
factual_drug ~ gender + income	34
factual_drug ~ health_before + income	35
factual_drug ~ gender + health_before + income	36
factual_drug ~ age + income	37
factual_drug ~ gender + age + income	38
factual_drug ~ health_before + age + income	39
factual_drug ~ gender + health_before + age + income	40
factual_drug ~ neighborhood + income	41
factual_drug ~ gender + neighborhood + income	42
factual_drug ~ health_before + neighborhood + income	43
factual_drug ~ gender + health_before + neighborhood + income	44
factual_drug ~ age + neighborhood + income	45
factual_drug ~ gender + age + neighborhood + income	46
factual_drug ~ health_before + age + neighborhood + income	47
factual_drug ~ gender + health_before + age + neighborhood + income	48
factual_drug ~ instrument + income	49
factual_drug ~ gender + instrument + income	50
factual_drug ~ health_before + instrument + income	51
factual_drug ~ gender + health_before + instrument + income	52
factual_drug ~ age + instrument + income	53
factual_drug ~ gender + age + instrument + income	54
factual_drug ~ health_before + age + instrument + income	55
factual_drug ~ gender + health_before + age + instrument + income	56
factual_drug ~ neighborhood + instrument + income	57
factual_drug ~ gender + neighborhood + instrument + income	58
factual_drug ~ health_before + neighborhood + instrument + income	59
factual_drug ~ gender + health_before + neighborhood + instrument + income	60
factual_drug ~ age + neighborhood + instrument + income	61
factual_drug ~ gender + age + neighborhood + instrument + income	62
factual_drug ~ health_before + age + neighborhood + instrument + income	63
factual_drug ~ gender + health_before + age + neighborhood + instrument + income	64

Table 12: Used models for Cox models

Cox model	Combination number
~ factual_drug	1
~ factual_drug + gender	2
~ factual_drug + health_before	3
~ factual_drug + gender + health_before	4
~ factual_drug + age	5
~ factual_drug + gender + age	6
~ factual_drug + health_before + age	7
~ factual_drug + gender + health_before + age	8
~ factual_drug + instrument	9
~ factual_drug + gender + instrument	10
~ factual_drug + health_before + instrument	11
~ factual_drug + gender + health_before + instrument	12
~ factual_drug + age + instrument	13
~ factual_drug + gender + age + instrument	14
~ factual_drug + health_before + age + instrument	15
~ factual_drug + gender + health_before + age + instrument	16
~ factual_drug + neighborhood	17
~ factual_drug + gender + neighborhood	18
~ factual_drug + health_before + neighborhood	19
~ factual_drug + gender + health_before + neighborhood	20
~ factual_drug + age + neighborhood	21
~ factual_drug + gender + age + neighborhood	22
~ factual_drug + health_before + age + neighborhood	23
~ factual_drug + gender + health_before + age + neighborhood	24
~ factual_drug + instrument + neighborhood	25
~ factual_drug + gender + instrument + neighborhood	26
~ factual_drug + health_before + instrument + neighborhood	27
~ factual_drug + gender + health_before + instrument + neighborhood	28
~ factual_drug + age + instrument + neighborhood	29
~ factual_drug + gender + age + instrument + neighborhood	30
~ factual_drug + health_before + age + instrument + neighborhood	31
~ factual_drug + gender + health_before + age + instrument + neighborhood	32
~ factual_drug + income	33
~ factual_drug + gender + income	34
~ factual_drug + health_before + income	35
~ factual_drug + gender + health_before + income	36
~ factual_drug + age + income	37
~ factual_drug + gender + age + income	38
~ factual_drug + health_before + age + income	39
~ factual_drug + gender + health_before + age + income	40
~ factual_drug + instrument + income	41
~ factual_drug + gender + instrument + income	42
~ factual_drug + health_before + instrument + income	43
~ factual_drug + gender + health_before + instrument + income	44
~ factual_drug + age + instrument + income	45
~ factual_drug + gender + age + instrument + income	46
~ factual_drug + health_before + age + instrument + income	47
~ factual_drug + gender + health_before + age + instrument + income	48
~ factual_drug + neighborhood + income	49
~ factual_drug + gender + neighborhood + income	50
~ factual_drug + health_before + neighborhood + income	51
~ factual_drug + gender + health_before + neighborhood + income	52
~ factual_drug + age + neighborhood + income	53
~ factual_drug + gender + age + neighborhood + income	54
~ factual_drug + health_before + age + neighborhood + income	55
~ factual_drug + gender + health_before + age + neighborhood + income	56
~ factual_drug + instrument + neighborhood + income	57
~ factual_drug + gender + instrument + neighborhood + income	58
~ factual_drug + health_before + instrument + neighborhood + income	59
~ factual_drug + gender + health_before + instrument + neighborhood + income	60
~ factual_drug + age + instrument + neighborhood + income	61
~ factual_drug + gender + age + instrument + neighborhood + income	62
~ factual_drug + health_before + age + instrument + neighborhood + income	63
~ factual_drug + gender + health_before + age + instrument + neighborhood + income	64