A connection between survival multistate models and causal inference for external treatment interruptions

Alexandra Erdmann[1] (iD)**, Anja Loos[2], and Jan Beyersmann[1]**

Abstract

Recently, treatment interruptions such as a clinical hold in randomized clinical trials have been investigated by using a multistate model approach. The phase III clinical trial START (Stimulating Targeted Antigenic Response To non-small-cell cancer) with primary endpoint overall survival was temporarily placed on hold for enrollment and treatment by the US Food and Drug Administration (FDA). Multistate models provide a flexible framework to account for treatment interruptions induced by a time-dependent external covariate. Extending previous work, we propose a censoring and a filtering approach both aimed at estimating the initial treatment effect on overall survival in the hypothetical situation of no clinical hold. A special focus is on creating a link to causal inference. We show that calculating the matrix of transition probabilities in the multistate model after application of censoring (or filtering) yields the desired causal interpretation. Assumptions in support of the identification of a causal effect by censoring (or filtering) are discussed. Thus, we provide the basis to apply causal censoring (or filtering) in more general settings such as the COVID-19 pandemic. A simulation study demonstrates that both causal censoring and filtering perform favorably compared to a naïve method ignoring the external impact.

1 Introduction

A substantial amount of treatment interruptions during a randomized clinical trial, not expected at the planning stage, might cause that the planned statistical analysis does not address the original study objective anymore. The phase III clinical trial START (Stimulating Targeted Antigenic Response To non-small-cell cancer) had to deal with a large number of treatment interruptions as it continued after a clinical hold (CH) was lifted.[1] A CH order issued by the FDA (US Food and Drug Administration) to the sponsor of a clinical trial entails stop of enrollment and that patients may not receive the investigational drug.

The START study served as a motivating example for Nießl et al.[2] to evaluate the potential implications of the CH on the treatment effect on overall survival (OS) and to suggest analysis methods to account for treatment interruptions induced by the CH. They showed the multistate model framework as a suitable and flexible tool to investigate the impact of the CH on the treatment effect supporting discussions around appropriate analysis methods. To compensate for a potential negative impact of the CH on the treatment effect Nießl et al.[2] suggested a censoring approach which censors patients at the start of the CH. They showed that this approach provides reliable estimates of the treatment effect on OS preserving the initial objective of the trial for a causal interpretation: estimation of initial treatment effect in the absence of the CH. Their

[1]Institute of Statistics, University of Ulm, Ulm, Germany
[2]Global Biostatistics and Epidemiology, Merck Darmstadt, Darmstadt, Germany

Corresponding author:
Alexandra Erdmann, Institute of Statistics, University of Ulm, Helmholtzstrae 20, Ulm 89069, Germany.
Email: alexandra.erdmann@uni-ulm.de

argumentation is based on the fact that the CH order is an external event. It is important to note that censoring by CH is independent in a counting process sense. This is more subtle than the common random censoring model that assumes stochastically independent death and censoring times. Assuming a beneficial treatment effect, censored and uncensored patients cannot be assumed to have the same hazard of death because censored patients may have to suspend their treatment which is expected to be harmful for the time to death.

The considerations of Nießl et al.[2] serve as motivation to provide an in-depth discussion on the connection of censoring by treatment interruption and causal inference for assessing treatment effects under hypothetical interventions.

In this article, we suggest an enhanced CH-censoring approach, which censors only patients that actually had to interrupt their treatment due to the CH. These are the progression-free patients in the treatment group, because in the START trial, treatment is administered only before disease progression. However, this method implies that we cannot simply use a Cox model to determine the treatment effect.

Moreover, we transfer our considerations about causal censoring to the more general concept of filtering making use of information collected after the end of the CH. These two novel methods to account for the CH have the major benefit that more observed events are included in the analysis compared to the censoring approach of Nießl et al.[2] However, it is less intuitive whether these two new methods provide causal estimates, as the censoring (and filtering) now depends on individual covariate values. To gain a better understanding of the relationship between independent censoring in a counting process sense and censoring as a causal intervention, we examine, based on the example of the CH situation, the implications of the censoring and filtering approaches on the matrix of transition probabilities from both a causal perspective and a multistate model perspective. In doing so, we also address the connection of "censoring by treatment interruption due to CH" to the g-computation formula.[3–6] Moreover, we discuss the assumptions for "causal censoring," that is, for censoring that lead to an identifiable causal effect. Our goal with this paper is to provide a conceptual multistate framework that enables us to address treatment interruptions or discontinuations and to identify a causal treatment effect in general settings with an external time-dependent covariate inducing a time-varying treatment.

Another example for a possible application are clinical trials affected by the COVID-19 pandemic. The current COVID-19 pandemic and subsequent restrictions have various consequences on planned and ongoing clinical trials. Its direct and indirect effects might lead to intercurrent events and missing data potentially leading to biased study results. The multistate model approach could support discussions and decision making on how to cope with the COVID-19 pandemic from a statistical point of view.[7] van Geloven et al.[8] discuss censoring of treatment as a hypothetical strategy for answering the question of how likely an event would be if no one received treatment.

In contrast to Nießl et al.,[2] this article will focus on the estimation of probabilities using the Aalen-Johansen estimator rather than the estimation of hazard ratios to be more in the spirit of causality.[9]

The remainder of this article is structured as follows. Section 2 introduces the theoretical background including survival multistate models as well as some key aspects of causal inference. Section 3 recapitulates multistate modeling of a CH by the example of the START trial and suggests two novel methods to account for the impact of the CH. Section 4 set out the link to causal inference and the identification of causal treatment effects. In Section 5., simulation studies are performed comparing the suggested methods. The article concludes with a discussion in Section 6.

2 Theoretical background

We begin by presenting general survival multistate models with a focus on external categorical time-dependent covariates and the estimation of transition probabilities (Section 2.1.). Section 2.2 gives a short overview of some aspects of causal inference we need to create the link between multistate modeling and the identification of causal effects.

2.1 Survival multistate models and time-dependent covariates

In contrast to the standard survival model, a multistate model facilitates the analysis of complex survival data with any finite number of states and any transition between these states.[10] If no transitions out of a state are modeled, the state is called *absorbing*, and *transient* otherwise.

A multistate model could be interpreted as a joint model for time-dependent categorical covariates and the time-to-event endpoint: covariates are included through transitions from one transient state to another and the time-to-event endpoint through the time until the multistate process enters an absorbing state. Figure 1 shows an illness-death model with absorbing state death and transient state progression of disease (PD) that jointly models the oncology endpoints OS and progression-free survival (PFS).[11] The model also reflects the situation of the START study if the CH had not occurred.

Basically, Kalbfleisch and Prentice[12] distinguish between two categories of time-dependent covariates. To put it briefly, in contrast to *internal* covariates, the existence of *external* covariates does not depend on the individual under study.
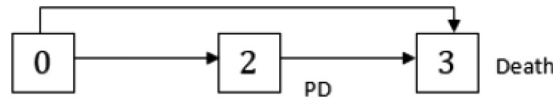
Figure 1. Multistate model with progression of disease (PD) as intermediate state.

A simple example of an external covariate is air pollution in a study of asthma events. The level of air pollution may influence the hazard on asthma events, but asthma events of individuals have no impact on the level of air pollution. An example of an internal covariate is blood measurements at study visits. The PD state in Figure 1 represents the internal time-dependent covariate progression status. Kalbfleisch and Prentice[12] give also a formal definition of external covariates, which is introduced below. Let T_i denote the event time of the ith individual under study and $w_i(t)$ its left-continuous covariate vector at time t. $W_i(t) := \{w_i(u); 0 \le u \le t\}$ includes the covariate history up to time t. Then, for an external covariate it holds:

$$P(T \in [u, u + du]|W(u), T \ge u) = P(T \in [u, u + du]|W(t), T \ge u) \tag{1}$$

for all u, t, such that $0 < u \le t$. An equivalent condition is:

$$P(W(t)|W(u), T \ge u) = P(W(t)|W(u), T = u), \ 0 < u \le t \tag{2}$$

In words, the occurrence of an event at time u does not affect the future path of the external covariate $w(\cdot)$. Moreover, as the external covariate is an output of an external stochastic process, the values of an external covariate do not depend on individual past covariate paths.

As stated above, time-dependent and typically internal covariates with finite range could be incorporated in a multistate model. Including an external time-dependent covariate into a multistate model converts it to an internal covariate to a certain extent, as the covariable then reflects the individual experience of the patient under study.[13] The event of the CH can be described by an external time-dependent covariate, but, as we will discuss later, by adding the CH to the multistate model, we no longer consider it as an external event but its internal effect of treatment interruption.

Let $(M_t)_{t \ge 0}$ be a multistate process with finite state space $\{0, 1, 2, \ldots, \mathcal{K}\}$ denoting the state where an individual is in at time t and fulfilling the time-inhomogeneous Markov assumption. M_t is adapted to its self-exciting filtration $\mathcal{F}_t := (\sigma(M_s : 0 \le s \le t))_{t \in [0, \tau]}$ that can be seen as the history of the multistate process up to time t. In terms of $(M_t)_{t \ge 0}$ the transition hazards $\alpha_{lm}(t)$ from state l to state m are defined via:

$$\alpha_{lm}(t)dt = P(M_{(t+dt)} = m|M_{t-} = l), \ l \ne m. \tag{3}$$

The Markov assumption guarantees that the future course of an individual only depends on the state currently occupied and on time t. Identification of causal effects requires model assumptions and the Markov assumption is an essential condition as we will discuss in Section 4.2. The matrix of transition probabilities $\mathbf{P}(s, t) = (P(M_t = m|M_s = l))_{l, m \in \mathcal{K}}, s \le t \in [0, \tau]$, can be estimated by the Aalen-Johansen estimator:

$$\hat{\mathbf{P}}(s, t) = \prod_{s < u \le t} (\mathbf{I} + \Delta \hat{\mathbf{A}}(u)), \tag{4}$$

where $\mathbf{I}$ is the $(K + 1) \times (K + 1)$ identity matrix and $\prod$ is a finite product over all unique observed transition times in $(s, t]$. The matrix $\Delta \hat{\mathbf{A}}(t)$ has non-diagonal entries containing the increments of the Nelson-Aalen estimators of the cumulative transition hazards. Its diagonal entries are such that the sum of each row equals zero. The increments of the Nelson-Aalen estimator are given by:

$$\Delta \hat{A}_{lm}(t) = \frac{\# \text{ observed } l \ \rightarrow \ m \text{ transitions at } t}{\# \text{ observed in state } l \text{ just prior to } t} \tag{5}$$

Assuming one common initial state of all individuals, the probabilities of being in a certain state at time t, that is the state occupation probabilities $P(M_t = m)$, $m \in \{0, 1, \ldots, \mathcal{K}\}$, $t \in [0, \tau]$, are given by the first row of the Aalen-Johansen estimator (4). The time-point τ is chosen such that identifiability is guaranteed. We are interested in the endpoint OS. Looking at Figure 1, the state occupation probability of the absorbing state death describes the probability of already being dead at a given time t and can be used to quantify the endpoint of OS.

Furthermore, we introduce the following counting process notation. We consider n replicates assumed to be independent with individual process M_t^i, $i = 1, 2, n$. The *at-risk* process denoting the number of patients in state l and under observation just before time t is then given by:

$$Y_l(t) := \sum_{i=1}^{n} \mathbf{1}(M_{t_-}^i = l, t \leq C_i).\tag{6}$$

C_i is defined as the right-censoring time of individual i. The number of i's observed *direct* $l \to m$ transitions in $[0, t]$ is denoted by the counting process $N_{lm}^i(t)$ and $N_{lm}(t) := \sum_{i=1}^{n} N_{lm}^i(t)$, $l \neq m$. We write $\Delta N_{lm}(s)$ for the increments between time s and the previous time-point of a jump of $N_{lm}(s)$. So,

$$\Delta \hat{A}_{lm}(t) = \frac{\Delta N_{lm}(t)}{Y_l(t)}.\tag{7}$$

In the survival analysis literature, different types of right-censoring have been introduced.[5,14] We focus on the concept of independent censoring,[15] which is a much weaker assumption than simple random censoring. Independent censoring still allows valid statistical inference of the hazards, as it preserves the multiplicative structure of the intensity model. Observing a counting process via a filter implies that the jumps of that process are only observed and known when a suitable indicator process is switched on,[15,16] that is, we do not know whether an event occurred or not during a filtered interval. An in-depth discussion on independent censoring and independent filtering is included in the Supplemental Materials.

2.2 Key aspects of causality

Causal analysis goes one step further as the standard statistical analysis and does distinguish between causal effects and other sources of association. Many relevant publications dealing with causality from a statistical point of view use different causal notations and frameworks.[17–19]

Within this section, we focus mainly on the general requirements to identify a desired causal effect from observational data without going into details.

Pearl[18] has introduced a notation that describes the situation where a single variable X_1 is forced to take a specific value $\tilde{x}_1$ by some intervention for the complete population. This means $P(T|do(X_1 = \tilde{x}_1))$ refers to the distribution of T under the situation, where X has been forced to take value $\tilde{x}_1$ by some intervention. In contrast to ordinary conditioning, that is, $P(T|X = \tilde{x}_1)$, which corresponds to the distribution which we could passively observe when $X_1 = \tilde{x}_1$, $P(T|do(X_1 = \tilde{x}_1))$ could be interpreted as the causal effect of X on T by comparing different interventional values of X. We are interested in the initial treatment effect on OS which would have been observed in the absence of the CH, that is, under do(no CH), and not just on the treatment effect that is observed in the patients not affected by the CH.

A common graphical tool for displaying causal relations are causal DAGs (directed acyclic graphs). A detailed introduction to DAGs can be found for example in Greenland et al.[20] or Maathuis et al.[21] There are graphical rules like the back-door criterion which help us to decide whether we could identify a causal effect under the assumed causal model. Random variables are displayed as nodes and directed edges (arrows) convey the causal directionality. The lack of an arrow from one node to another can be interpreted as the absence of a direct causal effect regarding those nodes. Figure 2 shows an example DAG that reflects the situation of the START trial at a fixed time t. We will discuss this DAG in detail in Section 4.1. Let $V = \{V_1, \ldots, V_n\}$ be a set of discrete random variables. Parents of a random variable V_i denoted by $pa(V_i)$ are a set of nodes from which there is a direct arrow into V_i. If there is a sequence of nodes which connects V_i and V_j following the direction indicated by the edges and starting at V_i, we denote V_j a descendant of V_i. A set of variables $V^\star \subset V$ satisfies the back-door criterion relative to the ordered pair (V_i, V_j) if no node in $V^\star$ is a descendant of V_i and $V^\star$ blocks every path between V_i and V_j that contains an arrow pointing into V_i.[22] Then, the intervention distribution is given by:

$$P(V_j = v_j | do(V_i = v_i)) = \sum_{v^\star} P(V_i = v_i | V_j = v_j, V^\star = v^\star) \cdot P(V^\star = v^\star)\tag{8}$$

A variable that has no parent is called exogenous or root node and is determined only by factors outside of the graph. Otherwise, a variable is endogenous.

To quantify causal relations by analyzing observational data, we need some stronger assumptions than for standard statistical analyses. Hernán and Robins[17] describe three identifying assumptions for estimating averaged causal effects.

- Positivity: the combination of values possible under the intervention must also be possible under the observational regime. That means the relevant combinations under the observational regime requires a positive possibility. In our example, we observe non-progressive and progressive patients with death event that are not affected by the CH.
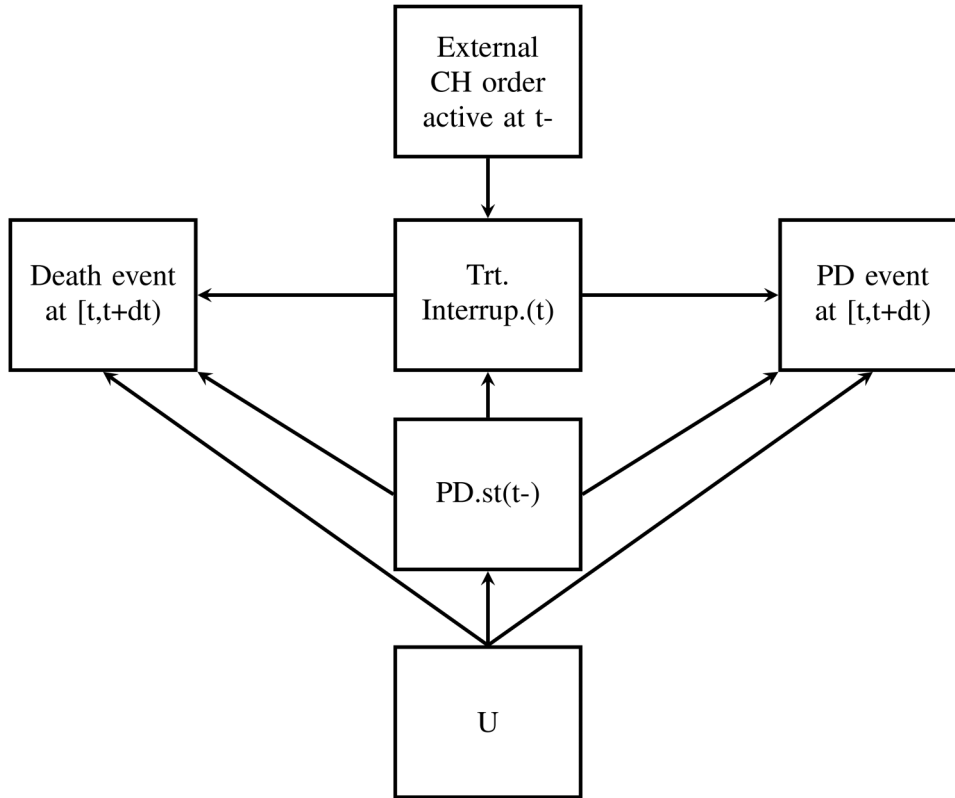
Figure 2. DAG for CH situation at fixed *t*. Note: The DAG describes causal relations for patients still alive at time *t*. DAG: directed acyclic graph; CH: clinical hold.

- Exchangeability: The individuals with intervention would have experienced the same average outcome as the individuals without intervention if they had not been subject to intervention. Randomization in clinical trials is expected to produce exchangeability between the treated and the untreated. Conditional exchangeability means that exchangeability is guaranteed within the levels of a covariate. Consequently, this assumption implies that there are no unmeasured confounders that are a common cause of both the exposure subject to the intervention and the outcome. In our example, the question is which patients can be assumed to experience the same average treatment effect as patients affected by CH would have if there were no CH.
- Consistency: the intervention must be well-defined involving that actual and counterfactual survival times coincides when the actual observed exposure is equal to the intervention value. We will define a concrete intervention representing "no CH" in later sections and we will see that there is more than one option.

A causal DAG $\mathcal{G}$ encodes the identifying assumptions. Thus, the main task is to decide whether the assumptions represented in any given graphical model are reasonable and sufficient to assess causal effects from the observed data.

Furthermore, the graphical model $\mathcal{G}$ with finite set of discrete random variables V ($|V| = n$) is called Markov, if any joint distribution generated by the model can be factorized as[22]:

$$P(v_1, \ldots, v_n) = \prod_{i=1}^{n} P(v_i|pa(v_i)) \qquad (9)$$

The truncated factorization formula,[22] also known as g-computation formula[3] or manipulation theorem,[23] enables us to determine the joint distribution generated by multiple interventions on a set of random variables for any Markovian model.

For any intervention on $X = \{X_1, \ldots, X_m\} \subset V$, $|X| = m$, $m < n$, the joint distribution is given by:

$$P(v|do(x_1, \ldots x_m)) = \prod_{i \in T} P(t_i|pa(t_i)) \prod_{j \in X} \mathbf{1}(v_j = x_j) \qquad (10)$$

where $T = V \setminus X$. In (10) factors of manipulated variables are removed. If v is consistent with the intervention then the remaining factors stay the same. If v is inconsistent with the intervention then the post-intervention distribution is equal to 0.

According to the definition of a causal DAG by Didelez,[24] the DAG is causal, implying that all causal effects are identifiable, if (9) and (10) hold.

3 Multistate modeling of clinical hold

Section 3.1 describes the motivating data example, the START trial sponsored by Merck KGaA. Section 3.2 recapitulates how the CH is incorporated in a multistate model as proposed in Nießl et al.[2] Section 3.3 suggests two methods to account for CH impact extending the censoring approach proposed by Nießl et al.[2]

3.1 Motivating data example: The START trial

The START trial was a phase III, 2:1 randomized, placebo-controlled, and double-blind group-sequential trial. The aim of START was to investigate whether the MUC1 antigen-specific cancer immunotherapy tecemotide given as maintenance therapy after chemoradiation improves OS duration in patients with unresectable stage III non-small-cell lung cancer.[1] Treatment was administrated until PD.

At the time of the CH, accrual was nearly completed with 1182 of 1322 planned subjects, blinded treatment was suspended in 531 patients and 180 patients did not restart treatment after a median of 135 days of treatment interruption when the CH was lifted.

The sponsor decided to continue the trial after increasing the overall sample size and to exclude those patients in a modified intention-to-treat (mITT) analysis believed to be most affected by the treatment interruption, that is, all 274 patients randomly assigned within the six months before the CH were excluded from primary analysis. Overall, 1513 patients were enrolled, with 1006 treated with tecemotide and 507 assigned to placebo. In the mITT subset 1239 patients remained, 829 receiving tecemotide and 410 receiving placebo. Results of the trial showed no significant prolongation of OS duration in the mITT analysis population used for primary analyses. Thus, the question arises if the mITT analysis had in principle compensated for potential implications of the CH.

Nießl et al.[2] showed that the mITT analysis was a reasonable proceeding providing reliable estimates of the initial treatment effect for OS. Moreover, they suggested a more flexible censoring approach for compensating the impact of the CH. More details about the START trial are available in Butts et al.[1]

3.2 Multistate models for treatment and control groups

To evaluate the impact of the CH on the treatment effect, we consider two separate models for treatment and control groups as proposed in Nießl et al.[2] Both models include PD as a transient state, as treatment changed upon PD for the treatment group as well as for the placebo group. It has to be noted that both treatment decisions and PD state in our multistate model refer to the time of the diagnosis of PD. To describe the event of the CH in the treatment group we add two transient states representing the start and end of the CH to our model (see Figure 4). Assuming that an induced treatment interruption of active treatment is the major consequence of the CH, we do not model the CH in the control group (see Figure 3). In the START trial, blinded treatment was discontinued with PD. As a consequence, patients with PD or death prior to CH did not experience a treatment interruption due to the CH. We model the CH as an individual event of a patient's course of disease and treatment, as only patients actually affected by the CH make a transition into the state "CH on." Thus, we have to distinguish between two types of covariates describing the CH, on the one hand, the CH as an external covariate that occurs at one point in calendar time for the entire population, and on the other hand, the CH as an internal covariate incorporated in our multistate model that induces a treatment interruption.[13]

In Nießl et al.[2] the multistate model has been used to generate a better understanding of the impact of the CH on the treatment effect in the START ITT population. We briefly recap the main results: The Nelson-Aalen estimates of the direct death hazards (i.e. death without prior progression) are rather low and no clear difference can be observed between treatment groups. Also, no clear treatment effect could be observed for the time from the PD state to death ($2 \rightarrow 3$). However, the Nelson-Aalen estimates of the cumulative hazards into the PD state indicate a protective effect of treatment on the time until PD, which cannot be observed during treatment suspension due to CH, but is restored with the resumption of treatment. In other words, when comparing the Nelson-Aalen estimates of the "$0 \rightarrow 2$" and "$4 \rightarrow 2$" cumulative transition hazards to the "$0 \rightarrow 2$" cumulative transition hazard of the control group, a treatment effect is observed that does not occur when comparing the estimates of the "$1 \rightarrow 2$" cumulative transition hazard to the "$0 \rightarrow 2$" cumulative transition hazard of the control group. We refer to Nießl et al.[2] for a detailed discussion of the Nelson-Aalen estimates and a graphical illustration incorporating incidence rates into Figures 3 and 4 for easier interpretation.

Figure 3. Multistate model with progression of disease (PD) as intermediate state for the control group.
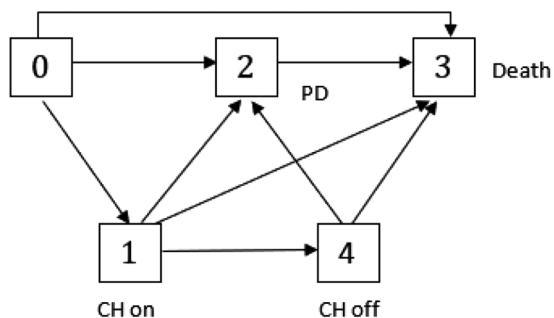

Figure 4. Multistate model $(X_t)_{t\geq 0}$ with PD and CH as intermediate states for the treatment group. PD: progression of disease; CH: clinical hold.

Our aim is to identify the initial treatment effect on OS in a hypothetical situation where the CH has not occurred. With regard to our multistate models OS is defined as the time until the absorbing state "death" is reached.

3.3 Compensating for the impact of the CH using censoring and filtering approaches

Nießl et al.[2] compared the mITT analysis as applied to the START data and the censoring approach to the naïve approach which simply ignores the CH. They concluded that the mITT analysis was a meaningful approach that compensated for the impact of the CH. However, the censoring approach is a simple alternative that provides convincing simulation results and needs less information about the mode of action of the treatment than the mITT analysis where an exclusion window has to be determined. Moreover, Nießl et al.[2] pointed out that the censoring approach has a causal interpretation with regard to a treatment effect that had been observed in the absence of the CH. Thus, we want to extend the censoring concept and will also have a closer look at the connection to causal inference to enable the application of causal censoring in more general settings.

In Nießl et al.[2] all patients—actually affected by the CH or not—are censored at the beginning of the CH exploiting the fact that CH is an external mechanism independent of the individual patient. Let $(X_t)_{t\geq 0}$ be the multistate process according to the multistate model for the treatment group introduced in the previous section (cf. Figure 4). The model of the treatment group reflects the individual treatment interruption due to the CH, but includes no information about the CH order for patients not actually affected by the CH. Thus, the censoring approach cannot be described in terms of the multistate process X_t. Let us consider the multistate process $(Z_t)_{t\geq 0}$ representing the multistate model shown in Figure 5. That model includes a transient state "CH on" representing the individual start of the external CH. That means, a patient makes a transition into state 1 "CH on" as soon as the CH occurred irrespective of whether a treatment interruption is induced or not. Therefore, an important difference between models of Figures 4 and 5 is that the model of Figure 5 includes a PD → CH, that is, $2 \to 1$, transition. For ease of presentation, the end of CH is not modeled. Let $T_i = \inf\{t|Z_t^i = 3\}$ the time of death of individual i and CH.$t_i = \inf\{t|Z_t^i = 1\}$ the time of the start of the CH for individual i on its study time scale. It holds $T_i = \inf\{t|Z_t^i = 3\} = \inf\{t|X_t^i = 3\}$. Not all individuals in the study experienced the CH. Thus, we define $\inf\{\} = \infty$. An individual is censored according to the "censoring by CH" approach as soon as it enters state 1 in Figure 5, that is, if CH.$t_i < T_i$. Then, state 1 is an absorbing observational state and only the transitions illustrated by the solid black arrows are considered as "observed." It is quite obvious that T_i and CH.t_i, that is, the death and censoring times, are not stochastically independent, as the CH leads to treatment discontinuation for progression-free patients. However, censoring by CH is independent censoring with regard to the OS hazard prior to CH and at the same time lead to causal estimates of a treatment effect in a hypothetical world where the CH never happened. We will deal with the question of when independent censoring not only leads to hazard estimates undisturbed by censoring but also causal probability statements in Section 4.

A possibility to reduce the censoring rate is to censor only patients which are actually affected by the CH, that is, who had to suspend the experimental treatment. With regard to our multistate model (cf. Figures 4 or 6), an individual is
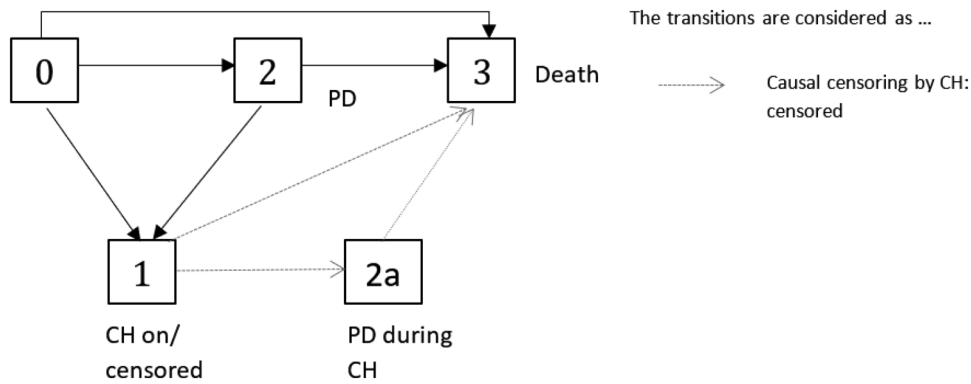
Figure 5. Multistate model $(Z_t)_{t\geq0}$ illustrating "censoring by CH." Note the $2 \to 1$ rather than $1 \to 2$ transition. Solid lines represent observed transitions. Note that state 1 is an absorbing state under causal censoring by clinical hold (CH).
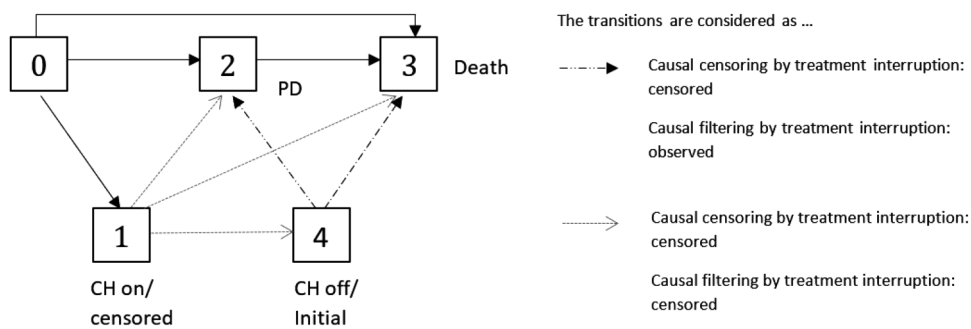

Figure 6. Multistate model $(X_t)_{t\geq0}$ illustrating "censoring and filtering by treatment interruption." Solid lines represent observed transitions.

censored as soon as it enters the state "CH on." An illustration of the concept of "censoring by treatment interruption" is given in Figure 6. With this censoring, state 1 ("CH on") is an absorbing state and only the solid black arrows are considered as "observed." Let $\text{PD.st}_i = \mathbf{1}(X_t^i = 2)$ denote whether an individual is progressive or not. Then, the relation between the two censoring concepts—by CH and by treatment interruption—could be described in the following way: an individual is censored by CH if $\text{CH.t}_i < T_i$ and if additional $\text{PD.st}_i(\text{CH.t}_i -) = 0$ then it is censored by treatment interruption. In summary, we differ the two censoring concepts:

$$\text{Censoring by CH} \Leftrightarrow \text{CH.t}_i \leq T \tag{11}$$

$$\text{Censoring by treatment interruption} \Leftrightarrow \text{CH.t}_i \leq T \text{ and } \text{PD.st}_i(\text{CH.t}_i -) = 0 \tag{12}$$

In addition, we consider the left-continuous time-dependent covariate $\overline{\text{CH}}(t)_i = \mathbf{1}(\exists u < t : X_u^i = 1)$ indicating whether a treatment interruption had been induced due to the CH in $(0, t \wedge T_i)$. The censoring by treatment interruption (12) is independent with regard to the $0 \to 2$ and $0 \to 3$ transition hazards, in the sense that those hazards are not disturbed by censoring the $0 \to 1$ transition. But, in contrast to censoring by CH (11), the censoring is not necessarily independent with regard to the OS times, because progressive patients, which might have a higher risk to die than the non-progressive patients, are not subject to censoring. This underscores the importance of carefully defining what is meant by independent censoring. Since the "censoring by treatment interruption" depends now on individual time-dependent covariate values, it is less intuitive whether this censoring concept is still "causal," that is, provides valid inference for a treatment effect in the hypothetical situation of had no CH occurred.

We will show in Section 4.2 that we can use the partial empirical transition matrix, that is the Aalen-Johansen estimator applied to the censored data (here: censoring by treatment interruption (12)), for causal inference in the hypothetical situation.

The "censoring by treatment interruption" does not use any information collected after the resumption of treatment. We propose a further approach which might be preferable in settings where restoring of the treatment effect can be assumed,

e.g. in chronic, non-life threatening disease. It should be noted that the multistate framework is a suitable tool to evaluate whether a treatment effect is restored after a treatment interruption. Our suggestion is to censor the patients as described before, but to re-include them to the analysis/risk sets after the end of treatment interruption due to CH with the end of CH as entry time, if they are still under observation at the end of their treatment interruption. As illustrated in Figure 6, according to this approach transitions out of state 1 are not considered in the analysis but, in contrast to the "censoring by treatment interruption" approach, transitions out of state 4, that is now treated as the initial state after "observation has been switched on," and "$2 \rightarrow 3$" transitions after CH are considered. It is important to note that it is not possible for a patient to contribute to the risk set twice at the same time, as re-entry does not mean that a patient's study time is reset. This approach corresponds to filtering, as it does not consider any death or PD events during CH. Let us define $CH(t)_i = \mathbf{1}(X_t^i = 1)$ indicating whether an individual is currently affected by the CH. With regard to our multistate model in Figure 6, $CH(t)_i = 1$ when entering state "CH on" and $CH(t)_i = 0$ when leaving that state. Consequently, when filtering we do not "observe" the counting process of death events while $CH(t)_i = 1$. According to the definition of independent filtering as explained in the Supplemental Materials, the filtering is independent with regard to the $0 \rightarrow 2$ and $0 \rightarrow 3$ transition hazards.

In the following we summarize the censoring and filtering concepts introduced in this section. The concepts have in common that they provide the basis for valid causal estimates of the treatment effect that would had been observed in the absence of the CH, but the way the intervention "no CH" is implemented differs. In other words, under certain assumptions they all could provide inference for $P(T \leq t|do(\text{no CH}))$. We will discuss the implications and causal assumptions in more detail in Section 4.

- Censoring by CH: A patient is censored if $CH.t_i < T_i$. In words, a patient is censored at the individual start of the CH order irrespective of whether a treatment interruption is induced or not. This approach has been considered by Nießl et al.[2] and is illustrated in Figure 5. This approach involves the largest reduction of observed events. However, it also allows a causal interpretation when estimating the treatment effect on the OS hazard (e.g. by the Kaplan-Meier estimator).
- Censoring by treatment interruption: A patient is censored if $CH.t_i < T_i$ and $PD.st_i = 0$. In words, an individual is censored as soon as it has to suspend treatment due to the CH, that is, only progression-free patients of the treatment group are potentially censored. This approach will be considered in detail within this paper and is illustrated in Figure 6. It involves a medium reduction of observed events and allows a causal interpretation of the Aalen-Johansen estimator.
- Filtering by treatment interruption: We observe the patient's course of disease via the filter $C_i(t) := \mathbf{1}(CH(t)_i = 0)$. That means, progression-free patients of the treatment group are censored at the beginning of the CH and observation is restarted after the end of the CH. This approach will be considered in detail within this paper and is illustrated in Figure 6. It involves the smallest reduction of observed events, but for a causal interpretation of the Aalen-Johansen estimator the additional assumption of a restored treatment effect after the CH is required.

As illustrated in Figures 5 and 6, the crucial difference between "censoring by CH" and "censoring by treatment interruption" is that in Figure 5 a progressive patient could make a transition into the CH state (i.e. will be censored), in contrast to Figure 6.

It has to be kept in mind that the CH per se is an external event that could be described via an external time-dependent covariate (see Section 2.1.). However, our defined covariates $\overline{CH}(t)$ and $CH(t)$ are not external as they depend on the individual progression status.

Table 1 summarizes part of the introduced notation, which we will need later on.

4 Causal estimation of treatment effects in the presence of treatment interruptions

Before showing that the estimation of the transition matrix when censoring by CH coincides with the estimation of a causal treatment effect under the hypothetical intervention "no CH occurred" (cf. Section 4.2.), we present a DAG to illustrate the underlying data generating mechanism (cf. Section 4.1.).

4.1 Causal DAG

We use a causal DAG (see Figure 2) to represent the CH study situation and to discuss it from a causal point of view. In contrast to a multistate model, where the arrows illustrate the potential subsequent occurrences of events, an arrow in the DAG indicates causal influences (cf. Section 2.2.).

We consider the causal relationships at a fixed time-point t. Consequently, the DAG does not represent the complete study situation. However, it will help us to understand the source of potential bias and to define an adequate Markov multistate model that captures all information to prevent bias introduced by censoring.

Table 1. Notation table (only with main notation).

Notation	Definition
$(Z_t)_{t \geq 0}$	multistate process of Figure 5
$(X_t)_{t \geq 0}$	multistate process of Figure 4
$N_{lm}(t)$	counting process for observed direct $l \rightarrow m$ transitions at t
$\alpha_{lm}(t)$	transition hazard from state l to state m at t
$\hat{\mathbf{A}}(t)$	matrix-valued Nelson-Aalen estimator at t
$\hat{\mathbf{P}}(s, t)$, $\mathbf{P}(s, t)$	(empirical) transition matrix
T_i	$:= \inf \{t \mid Z_t^i = 3\} = \inf \{t \mid X_t^i = 3\}$, that is, individual time of death
$PD.st(t)_i$	$:= I(X_t^i = 2)$, that is, individual progression status at time t
$CH.t_i$	$:= \inf \{t \mid Z_t^i = 1\}$, that is, individual time of start of the CH
$CH(t)$	$:= I(X_t^i = 1)$, that is, $=1$ if the individual is currently affected by a trt. interrupt.
$\overline{CH}(t)_i$	$:= I(\exists u < t : X_u^i = 1)$, that is, $=1$ if a trt.interrupt. had been induced in $(0, t \wedge T_i)$
$C_i(t)$	$:= I(CH(t)_i = 0)$, filtering process

As already stated, the occurrence of the external event CH (not its implications) does not depend on any individual covariate paths. From a causal perspective, the "external" CH is an exogenous variable, that is, a root node with no descendants, as it does not have any causal parents within our model (i.e. within our clinical trial). The CH induces a treatment interruption if a patient is still under risk and progressive-free at the time of the CH. Thus, progression status just before the beginning of the CH is a causal parent of "treatment interruption induced by CH." In other words, the progression status has to be known to understand whether the individual is affected by the CH or not. A treatment interruption might influence the treatment effect on time till death and on time till progression. The DAG must be interpreted locally in time and therefore does not indicate whether a past treatment interruption still influences time to death or PD event after resumption of treatment. The current progression status may affect the time to death as well. Moreover, it can be assumed that progression status just before t and PD or death event at t have common causal parents (denoted by U in Figure 2).

Figure 2 shows the DAG which summarizes the causal relations in our clinical trial within the treatment group at fixed time t. It is important to note that the CH per se is an exogenous variable, but because its impact depends on individual covariates, the induced treatment interruption is endogenous.

4.2 Causal interpretation of the partial empirical transition matrix

In the following section, we discuss the causal interpretation of the partial empirical transition matrix that results from applying our censoring and filtering approaches. We begin by considering g-computation in the setting of time-continuous Markov chains and its connection to our censoring and filtering approaches. Aalen et al.[5] pointed out in their rather brief Section 9.6.2 that estimating a partial transition matrix, that is, the usual estimator of transition probabilities but applied to artificially censored data, for example, because of treatment deviations, gives a valid estimate for the treatment effect in the absence of treatment deviations, that is, under do(no treatment deviation), and that this estimator could be seen as a special case of the g-computation formula.[6] Moreover, Gran et al.[4] present the artificially manipulating transitions in a multistate model for sickness absence at work as a possible method to assess the causal effect of certain interventions. However, the arguments for the causal interpretation of the partial transition matrix after censoring have so far been indicated in the mentioned literature rather than explored and justified in detail. It is important to note that generally censoring in a multistate model framework provides a causal interpretation of the partial empirical transition matrix, but requires certain causal assumptions. Therefore, we will have a closer look at the connection of censoring multistate model data and g-computation and will investigate the assumptions for the identification of a causal treatment effect in the presence of treatment interruptions.

Our objective is to estimate the initial treatment effect on OS which would have been observed in the absence of the CH. Thus, we could consider the state occupation probability of the absorbing state death under do (no CH): $P(X_t = 3 \mid do(\text{no CH})) = P(T \leq t \mid do(\text{no CH}))$, where T is the time of death. We assume no other implications of the CH than treatment interruption. Hence, do(no CH) is principally the same intervention as do(no treatment interruption). Depending on the assumptions about the implications of the treatment interruption, there is more than one option to identify consequences of that intervention. Assuming that the treatment effect is completely restored after the CH, our causal censoring and filtering both represent the same intervention "no CH" and, thus, both estimate a treatment effect with the same causal interpretation. But the two approaches require different modifications of the empirical transition matrices as we will discuss in the following sections.

For ease of presentation, we will not consider event times subject to right-censoring. That means we do not consider "usual" right-censoring additional to the censoring by CH. It is important to note that all derivations within the following sections could easily be extended for independent right-censoring using the well-known standard arguments.

4.2.1 G-computation and censoring by treatment interruption

For the moment, we will ignore the external node U of the causal DAG (cf. Figure 2) which is a common causal parent of both PD and death events, and show that then our censoring approach coincides with the g-computation formula.

Estimating the transition matrix while treating patients affected by the CH as censored results in a partial transition matrix in the sense that less transitions are considered and the state space is reduced compared to the usual transition matrix for our multistate model of Figure 4. We want to show that considering this partial empirical transition matrix corresponds to the g-computation and provides valid estimates with a causal interpretation towards a hypothetical treatment effect without CH. We use the same notation as introduced in Section 3.3. (cf. Table 1).

We drop index i and recall that $\overline{CH}(t) = 1$, if and only if treatment has been interrupted on $(0, t)$. Let us denote

$$t_1, \ldots, t_r \text{ all observed transition times occurred under } \overline{CH}(t_k) = 0 \quad k \in \{1, \ldots, r\} \tag{13}$$

and

$$s_1, \ldots, s_p \text{ all observed transition times occurred under } \overline{CH}(s_k) = 1 \quad k \in \{1, \ldots, p\}, \tag{14}$$

where we assume no ties.

Hence $s_1, \ldots, s_p$ correspond to the transitions during and after the CH including all transitions into and out of the states "CH on" and "CH off" and $t_1, \ldots, t_r$ all other transitions (see Figure 4). The transitions $t_1, \ldots, t_r$ are the observed transitions under "censoring by treatment interruption" that are shown with solid black arrows in Figure 6. The transition times into state 1 are the censoring times.

One way to implement our desired intervention of "no treatment interruption induced by CH" is setting $\overline{CH}(u) = 0$ for all $u \leq t$. From a multistate model perspective, the intervention corresponds to manipulating the $0 \rightarrow 1$ transition, that is, setting $\alpha_{01}(u) = 0$, $u \leq t$ or empirically, setting $\Delta N_{lm}(s_k) = 0$ for all l,m $l \neq m$ and all $k \in \{1, \ldots, p\}$.

According to the g-computation formula (10), we get $\hat{\boldsymbol{P}}(0, t|\text{do(no CH)})$ from the usual Aalen-Johansen estimator (4) by doing the following:

1. We do not consider $\hat{P}(X_{sk}|pa(X_{sk})) = \hat{P}(s_k-, s_k) = I + \Delta\hat{A}(sk)$, in the calculation of the empirical matrix of transition probabilities for all $s_k \leq t$.
2. We set $\overline{CH}(t_k) = 0$ when calculating $\Delta\hat{A}_{lm}(t_k)$, that is

$$\Delta\hat{A}_{lm}(t_k)_{|\overline{CH}(t_k)=0} = \frac{\# \text{ observed } l \rightarrow m \text{ transitions at } t_k \text{ and } \overline{CH}(t_k) = 0}{\# \text{ atrisk in state } l \text{ just prior to } t_k \text{ and } \overline{CH}(t_k -) = 0} \tag{15}$$

Steps 1 and 2 imply that we can leave out those matrices describing transition times $s_1, \ldots, s_p$ completely, we can reduce the dimension of our transition matrix and do not consider individuals with prior treatment interruption in the calculation of $\hat{A}_{23}(u)$. Thus, we get:

$$\hat{P}(0, t|\text{do(no CH)}) = \prod_{t_k \leq t}(\boldsymbol{I} + \Delta\hat{A}(t_k)_{|\overline{CH}(t_k)=0}) \tag{16}$$

The estimator (16) is exactly the estimator of the transition matrix that we obtain when censoring at the start of the CH. The factors $(\boldsymbol{I} + \Delta\hat{A}(t_k))$ in (16) describe the conditional empirical probability of one single transition between states. From a causal point of view, it represents the probability of a single causal node, while all other remain unchanged.

As we have seen in Section 3.3., "censoring by treatment interruption" is independent in the sense that the estimation of the $0 \rightarrow 2$ and $0 \rightarrow 3$ transition hazards is not disturbed by the censoring. Independent censoring ensures the identification of the intensity of an incompletely observed event process. However, this cannot generally be transmitted to the probability scale. That means, in general, independent censoring cannot be interpreted as an intervention that leads to the identifiability of causal effects, that is, probability statements under do(no censoring). In short, the causal DAG helps us to understand the causal relations and find a multistate model that captures all relevant information about stopping treatment such that conditional exchangeability is guaranteed when censoring is applied as an intervention. We will revisit the assumption of exchangeability in Section 4.2.3 and discuss it in more detail. The other two identifying assumptions, positivity and

consistency, are apparently fulfilled as well. Positivity here simply means that we observe non-progressive and progressive patients with death event that are not affected by the CH. Moreover, our intervention is well-defined.

More details including a very simple and constructed data example can be found in the Supplemental Materials.

4.2.2 G-computation and filtering by treatment interruption

Similar considerations to those described in the previous section for causal censoring also apply to our causal filtering approach. As explained in Section 3.3., the difference to the causal censoring approach is that additional to the censoring at the start of the treatment interruption due to CH, patients still under observation at the end of the CH will be analyzed just as they would re-entry the study, that is, transitions out of the state "CH off" are treated as transitions out of the initial state (see Figure 6). Under the assumption that the treatment effect is restored as soon as the treatment is restarted an alternative intervention corresponding to do(no CH) is setting $CH(u) = 0$ for all $u \leq t$. As described in Section 3.3., $CH(u)$ indicates if an individual currently interrupts its treatment. Consequently, $CH(u) = 0$ for patients that are not or no longer affected by the CH.

The (partial) empirical transition matrix of the filtered process X_t is obtained by coding $0 \rightarrow 1$ transitions as censoring and treat $4 \rightarrow 2/4 \rightarrow 3$ transitions as $0 \rightarrow 2/0 \rightarrow 3$ transitions, respectively. In other words, during CH (that is state 1) observation is switched off, and observation is switched on when leaving that state. As our multistate model distinguishes between the initial state and "non-progressive after CH," it is not perfectly suited for explaining the filtered situation, however, Figure 6 illustrates which observations are still observed under causal filtering.

Using the same reasoning as in Section 4.2.1., but denoting by t_k all observed transition times occurred under $CH(t_k) = 0$ (instead of $\overline{CH}(t_k) = 0$), we obtain $\hat{P}(0, t | \text{do(no CH)}) = \prod_{t_k \leq t} (I + \Delta\hat{A}(t_k))$ according to the g-computation formula. For causal filtering, we require the additional assumption of a restored treatment effect, that is, we assume that $\alpha_{02}(t) = \alpha_{42}(t)$ and $\alpha_{03}(t) = \alpha_{43}(t)$, but with the advantage of including more observed events in the analysis.

More details including a very simple and constructed data example can be found in the Supplemental Materials.

4.2.3 The partial empirical transition matrix and the back-door criterion

We showed that, when ignoring the exogenous causal node U, the partial transition matrix is a direct application of the g-computation formula or truncated factorization formula (10). However, we know that in our setting there are nodes in the causal DAG (Figure 2) with common unobserved exogenous causal parents denoted by U. That means, there are potential unobserved confounders. The argumentation in the previous sections is primarily based on the fact that after application of the g-computation formula all factors $(I + \Delta\hat{A}(t_k))$ that are still included in $\hat{P}(0, t | \text{do(no CH)})$ have a causal interpretation under do(no CH). We will now use the back-door criterion to show that U does not compromise the causal interpretation of the single factors $(I + \Delta\hat{A}(t_k))$.

Looking at the causal DAG (Figure 2) we find that the current progression status (PD.st$(t-)$) blocks every back-door path "trt.interrup.(t) $\leftarrow \cdots \leftarrow$ death event at [t, t+dt]." The same applies to PD events at $[t, t + dt]$. Besides the current progression status, there is only one other causal parent for the treatment interruption, namely the external CH order with exactly one causal arrow pointing into treatment interruption. In other words, the set $Z^\star = $ ("PD.st(t-)," "external CH order active at t-") satisfies the back-door criterion relative to the pair ("trt.interrup.(t)," "death event at [t, t+dt]") and relative to ("trt.interrup.(t)," "PD event at [t, t+dt]"). Thus, according to the back-door criterion, we can identify the causal effect of the intervention "no CH" on $P(X_t = 2 | X_{t-} = 0)$ as:

$$P(X_t = 2 | X_{t-} = 0, \text{do(no CH)}) = \sum_{z^\star} P(X_t = 2 | X_{t-} = 0, \text{no CH}, pa = z^\star) \cdot P(pa = z^\star) \tag{17}$$

where pa denote the causal parents of "no CH" (= no treatment interruption), that are: the current progression status, that is already included in the condition and the external CH order, that is blocked by "no CH." Thus, according to the back-door criterion it holds:

$$P(X_t = 2 | X_{t-} = 0, \text{do(no CH)}) = P(X_t = 2 | X_{t-} = 0). \tag{18}$$

The same reasoning results in:

$$P(X_t = 3 | X_{t-} = 0, \text{do(no CH)}) = P(X_t = 3 | X_{t-} = 0). \tag{19}$$

As progressive patients (i.e. $X_{t-}^i = 2$) are not affected by the CH, it holds:

$$P(X_t | X_{t-} = 2, \text{do(no CH)}) = P(X_t = 3 | X_{t-} = 2). \tag{20}$$

When filtering, the probabilities (18) to (20) are exactly the probabilities that contribute to $\hat{P}(0, t | \text{do(no CH)})$ (cf. Section 4.2.2.).

However, in Section 4.2.1 we do not just omit some factors from $\hat{P}(0, t)$ to obtain $\hat{P}(0, t|\text{do(no CH)})$. Taking into account that earlier treatment interruption could also have an impact on the treatment effect, we adjust the respective risk sets according to the intervention. However, whether this is necessary or not depends on how we defined the multistate model. In our example, adding a further state "disease progression after CH" and applying g-computation with the intervention $\overline{CH}(t) = 0$ would not change the outcome $\hat{P}(0, t|\text{do(no CH)})$. Technically, however, it would not be necessary to adjust any risk sets of the remaining factors. In other words, the solution is to inflate the multistate model in such a way that we just have to omit the respective factors from $\hat{P}(0, t)$. From a practical point of view, the inflation is not necessary as long as the causal implications are fully understood.

This underlines again that the causal graph at a fixed time t is an aid to find the appropriate multistate model leading to a causal analysis, but describes not completely the causal relations of the complex situation.

In summary, we pointed out by using the back-door criterion that the unmeasured parents U can safely be ignored when calculating the partial empirical transition matrix.

As we have discussed in Section 2.2, an essential assumption for the identification of causal effects is the exchangeability assumption. In our setting, this means that we assume that the patients affected by the CH—that are the ones who had to interrupt their treatment and the ones we censor- would experience the same averaged treatment effect as the non-censored. The back-door criterion makes it clear, that exchangeability with regard to the treatment effect holds conditional on the progression history (assuming we have specified the correct causal DAG). Thus, common causal parents of the censoring mechanism and the event of interest have to be included in the multistate model.

5 Simulation studies

We perform simulation studies to evaluate how well our proposed causal censoring and filtering approaches compensate for a negative impact of the CH. We consider different scenarios which are inspired by the original START trial. We use a hazard-based algorithm interpreting the multistate model as a nested series of competing risk experiments to generate data based on the multistate models of Figures 3 and 4.[10] The true treatment effect, that is the treatment effect that would have been observed in the absence of a CH, is determined by simulation for the respective scenarios. For further details, see Nießl et al.[2]

We simulate different scenarios (ID I–IV) where the treatment effect of the OS is manifested primarily in a treatment effect on the time until progression. Compared to the original START trial, we focus on a more considerable treatment effect. We consider a direct progression hazard ratio between treatment and control of 0.6 in all scenarios, except in Scenario II, where we consider an even more pronounced treatment effect of 0.5. For Scenario I.a and Scenario II.a no censoring times are simulated. That means all censored observations are due to the CH. Scenarios III and IV treat the case where the treatment effect is not restored. Transition hazards that are not manipulated within a scenario are parametrically estimated from the START data.

Additionally, we consider a scenario (ID V) that deviates more from the original START setting, but shows a larger negative impact of the CH.

We simulate 1000 studies with a sample size of 1500 individuals assigned in a 2:1 ratio to the treatment or control group. Random censoring times are generated from a Gompertz distribution for all patients, which roughly mimicked the empirical censoring distribution in the original data.

To show whether our causal censoring and filtering methods (i.e. "censoring by treatment interruption" and "filtering by treatment interruption," cf. Section 3.3.) provide causal estimates under "do(no CH)" we compare Aalen-Johansen estimates averaged over 1000 simulation iterations with the true averaged Aalen-Johansen estimator in the absence of CH. As we are interested in the treatment effect of OS we consider the state occupation probability of "death," $P_{03}(0, t)$, and additionally we use those estimates for the calculation of (time-varying) relative risks measuring the treatment difference between treatment and control groups for OS: $\widehat{RR}(t) = \frac{\log(\hat{P}(T<=t|\text{treatment group}))}{\log(\hat{P}(T<=t|\text{control group}))}$. We could not simply calculate the "usual" hazard ratio using the Cox proportional hazards model, because on the one hand, we want to estimate the initial treatment effect that prevents us to include progression into the model, but on the other hand, we need to take into account the time-dependent progression status to avoid selection bias due to our censoring or filtering.

Table 2 shows the bias and the root mean squared error (RMSE) of the Aalen-Johansen estimates at different months for the different treatment effects. Table 3 presents averaged estimates of the time-varying relative risks for the same scenarios and selected time-points. Figures 7 to 9 report the average of the 1000 estimates of $P_{03}(0, t)$ for treatment and control compared to the true state occupation probability as well as simulation based 95% confidence intervals for three different scenarios. As can be seen, the naïve approach overestimates the state occupation probability, that is, it dilutes the treatment effect. The censoring approach compensates well for the negative impact of the CH on the treatment effect, except for later time-points. Here, the number of observed events appears not to be large enough for a valid estimation. For the

Table 2. Causal censoring and filtering approaches compared to naïve approach: Bias and root mean squared error (RMSE) of averaged Aalen-Johansen estimates of $P(T \leq t)$ at different and arbitrary time-points (in months from start of trial) in the treatment group.

	Bias				RMSE			
Method	M_{25}	M_{35}	M_{45}	M_{55}	M_{25}	M_{35}	M_{45}	M_{55}
Scenario I: $\exp(\beta_{02}) = 0.6$, $\exp(\beta_{12}) = 1.0$, $\exp(\beta_{42}) = 0.6$								
Naïve	0.0140	0.0164	0.0168	0.0173	0.0350	0.0328	0.0298	0.0279
Causal censoring	0.0002	−0.0002	−0.0081	−0.0284	0.0006	−0.0004	−0.0143	−0.0459
Causal filtering	0.0001	0.0001	−0.0000	0.0006	0.0003	0.0002	−0.0001	0.0010
Scenario I.a: $\exp(\beta_{02}) = 0.6$, $\exp(\beta_{12}) = 1.0$, $\exp(\beta_{42}) = 0.6$—no (usual) censoring								
Naïve	0.0125	0.0150	0.0164	0.0165	0.0311	0.0300	0.0290	0.0266
Causal censoring	−0.0013	−0.0006	−0.0016	−0.0163	−0.0033	−0.0012	−0.0029	−0.0264
Causal filtering	−0.0015	−0.0012	−0.0004	−0.0000	−0.0038	−0.0023	−0.0006	−0.0000
Scenario II: $\exp(\beta_{02}) = 0.5$, $\exp(\beta_{12}) = 1.1$, $\exp(\beta_{42}) = 0.5$								
Naïve	0.0210	0.0253	0.0260	0.0248	0.0575	0.0550	0.0499	0.0430
Causal censoring	−0.0013	0.0004	−0.0057	−0.0289	−0.0037	0.0008	−0.0110	−0.0502
Causal filtering	−0.0011	−0.0011	−0.0016	−0.0029	−0.0030	−0.0024	−0.0031	−0.0050
Scenario II.a: $\exp(\beta_{02}) = 0.5$, $\exp(\beta_{12}) = 1.1$, $\exp(\beta_{42}) = 0.5$—no (usual) censoring								
Naïve	0.0228	0.0264	0.0275	0.0276	0.0625	0.0575	0.0529	0.0480
Causal censoring	0.0006	0.0012	−0.0025	−0.0174	0.0017	0.0027	−0.0047	−0.0303
Causal filtering	0.0008	0.0004	0.0003	0.0004	0.0021	0.0009	0.0006	0.0007
Scenario III: $\exp(\beta_{02}) = 0.6$, $\exp(\beta_{12}) = 1.0$, $\exp(\beta_{42}) = 0.7$								
Naïve	0.0168	0.0206	0.0227	0.0239	0.0417	0.0412	0.0403	0.0386
Causal censoring	−0.0008	−0.0012	−0.0067	−0.0272	−0.0019	−0.0023	−0.0118	−0.0440
Causal filtering	0.0033	0.0052	0.0069	0.0084	0.0083	0.0103	0.0122	0.0136
Scenario IV: $\exp(\beta_{02}) = 0.6$, $\exp(\beta_{12}) = 1.0$, $\exp(\beta_{42}) = 0.9$								
Naïve	0.0238	0.0310	0.0358	0.0412	0.0592	0.0618	0.0634	0.0666
Causal censoring	−0.0006	−0.0018	−0.0057	−0.0234	−0.0016	−0.0036	−0.0100	−0.0379
Causal filtering	0.0117	0.0177	0.0226	0.0287	0.0292	0.0353	0.0400	0.0464
Scenario V: $\exp(\beta_{03})/\exp(\beta_{02})/\exp(\beta_{23}) = 0.6$, $\exp(\beta_{12}) = 1$, $\exp(\beta_{42}) = 0.6$								
Naïve	0.1311	0.1169	0.1013	0.0841	0.2331	0.1775	0.1397	0.1053
Causal censoring	0.0002	−0.0266	−0.0633	−0.0947	0.0004	−0.0404	−0.0874	−0.1186
Causal filtering	−0.0004	−0.0012	−0.0006	0.0002	−0.0008	−0.0018	−0.0009	0.0003

M_i: evaluation at month i.

$\exp(\beta_{02}) :=$ direct progression hazard ratio between treatment and control.

$\exp(\beta_{12}) :=$ hazard ratio between treatment and control during the CH.

$\exp(\beta_{42}) :=$ hazard ratio between treatment and control after resumption of treatment.

$\exp(\beta_{03}) :=$ direct death hazard ratio between treatment and control.

$\exp(\beta_{23}) :=$ death hazard ratio after progression between treatment and control.

filtering approach, very small biases are observed for all time-points, except for the case where the treatment effect is not (fully) restored after the end of the CH. Furthermore, filtering leads to much smaller confidence intervals than censoring. Overall, the bias of the treatment effect induced by the treatment interruption due to the CH is rather moderate in the situation of the START trial. Although Scenarios I to IV illustrate well that our proposed analysis methods help to sufficiently compensate for the impact of the CH, we consider also scenarios where the bias induced by the naïve method is much bigger. The description of the additional scenarios and their results are mainly presented in the Supplemental Materials except the results of Scenario V. In Scenario V, more patients are affected by the CH and the duration is longer (see the Supplemental Materials for more details). Moreover, a treatment effect on the time till death without prior progression and on the time till death after progression is added. Figure 10 illustrates the simulation results of Scenario V. Further simulation results could be found in the Supplemental Materials.

6   Discussion

In the present article, we have pointed out that censoring or filtering the empirical transition matrix by treatment interruption induced by an external mechanism has a causal interpretation towards a treatment effect under do(no treatment interruption) using the example of the CH in the phase III START trial with the primary endpoint OS.

Table 3. Causal censoring and filtering approaches compared to naïve approach: Average of estimated time-varying relative risks including bias and root mean squared error (RMSE) at different and arbitrary time-points (in months from start of trial).

Method	Estimated RR			Bias			RMSE		
	M_{25}	M_{35}	M_{45}	M_{25}	M_{35}	M_{45}	M_{25}	M_{35}	M_{45}
Scenario I: $\exp(\beta_{02}) = 0.6$, $\exp(\beta_{12}) = 1.0$, $\exp(\beta_{42}) = 0.6$									
Naïve	0.720	0.747	0.745	0.029	0.037	0.038	0.068	0.081	0.093
Causal censoring	0.693	0.714	0.734	0.003	0.004	0.027	0.061	0.088	0.476
Causal filtering	0.693	0.712	0.706	0.002	0.003	0.000	0.059	0.069	0.080
Scenario I.a: $\exp(\beta_{02}) = 0.6$, $\exp(\beta_{12}) = 1.0$, $\exp(\beta_{42}) = 0.6$—no (usual) censoring									
Naïve	0.716	0.740	0.740	0.024	0.030	0.035	0.064	0.072	0.078
Causal censoring	0.689	0.709	0.741	−0.002	−0.001	0.036	0.060	0.075	0.413
Causal filtering	0.689	0.706	0.702	−0.003	−0.004	−0.003	0.057	0.063	0.066
Scenario II: $\exp(\beta_{02}) = 0.5$, $\exp(\beta_{12}) = 1.1$, $\exp(\beta_{42}) = 0.5$									
Naïve	0.665	0.682	0.675	0.041	0.055	0.060	0.071	0.086	0.097
Causal censoring	0.626	0.638	0.667	0.003	0.011	0.052	0.041	0.084	0.518
Causal filtering	0.627	0.633	0.621	0.003	0.006	0.007	0.054	0.062	0.070
Scenario II.a: $\exp(\beta_{02}) = 0.5$, $\exp(\beta_{12}) = 1.1$, $\exp(\beta_{42}) = 0.5$—no usual censoring									
Naïve	0.666	0.681	0.673	0.042	0.054	0.057	0.067	0.079	0.084
Causal censoring	0.627	0.636	0.645	0.004	0.008	0.030	0.052	0.065	0.358
Delayed study entry	0.627	0.633	0.620	0.004	0.006	0.005	0.049	0.055	0.057
Scenario III: $\exp(\beta_{02}) = 0.6$, $\exp(\beta_{12}) = 1.0$, $\exp(\beta_{42}) = 0.7$									
Naïve	0.724	0.755	0.760	0.034	0.045	0.054	0.070	0.086	0.101
Causal censoring	0.691	0.711	0.745	0.000	0.002	0.039	0.061	0.093	0.577
Causal filtering	0.698	0.721	0.723	0.008	0.012	0.017	0.018	0.060	0.071
Scenario IV: $\exp(\beta_{02}) = 0.6$, $\exp(\beta_{12}) = 1.0$, $\exp(\beta_{42}) = 0.9$									
Naïve	0.739	0.781	0.795	0.049	0.072	0.088	0.078	0.104	0.125
Causal censoring	0.692	0.712	0.758	0.001	0.003	0.052	0.061	0.089	0.542
Causal filtering	0.715	0.751	0.762	0.025	0.042	0.055	0.065	0.084	0.102
Scenario V: $\exp(\beta_{03})/\exp(\beta_{02})/\exp(\beta_{23}) = 0.6$, $\exp(\beta_{12}) = 1$, $\exp(\beta_{42}) = 0.6$									
Naïve	0.730	0.676	0.633	0.272	0.271	0.264	0.284	0.285	0.281
Causal censoring	0.484	0.424	0.341	0.027	0.018	−0.028	0.153	0.223	0.207
Causal filtering	0.465	0.411	0.376	0.007	0.006	0.007	0.062	0.062	0.064

M_i: evaluation at month i.
$\exp(\beta_{02}) :=$ direct progression hazard ratio between treatment and control.
$\exp(\beta_{12}) :=$ hazard ratio between treatment and control during the CH.
$\exp(\beta_{42}) :=$ hazard ratio between treatment and control after resumption of treatment.
$\exp(\beta_{03}) :=$ direct death hazard ratio between treatment and control.
$\exp(\beta_{23}) :=$ death hazard ratio after progression between treatment and control.

We suggested a censoring approach which censors all progression-free patients of the treatment group at the start of the CH and a filtering approach which restarts observation after the end of the CH. We showed that estimating the transition matrix after applying our censoring and filtering approaches can be seen as an implementation of the g-computation formula. We further discussed requirements on the censoring or filtering mechanism for drawing causal inference in the presence of a time-varying treatment. An important finding is that we do not necessarily need random censoring (or filtering), but rather the independent censoring (or filtering) assumption has to be fulfilled and, in addition, the state space of the multistate model has to be rich enough to justify a causal interpretation. In our example, if progression were not in the model, the treatment effect for OS would be subject to selection bias, as patients who are longer non-progressive are more likely to be censored. That means that all covariates that we need to produce conditional exchangeability, which is an essential assumption for the identification of causal effects, must be captured by the multistate model. Censoring by an external covariate ensures exchangeability. The crucial point here is the correct identification of the causal DAG and the specification of a suitable Markov model.

The two concepts independent censoring and exchangeability have in common that they claim that the individuals are representative in a certain way. Independent censoring implies that the observed individuals are "representative" of the incompletely observed individuals in the sense that the intensity of the counting process in the complete data world can be estimated. Exchangeability refers to the intervened individuals that are representative of the non-intervened
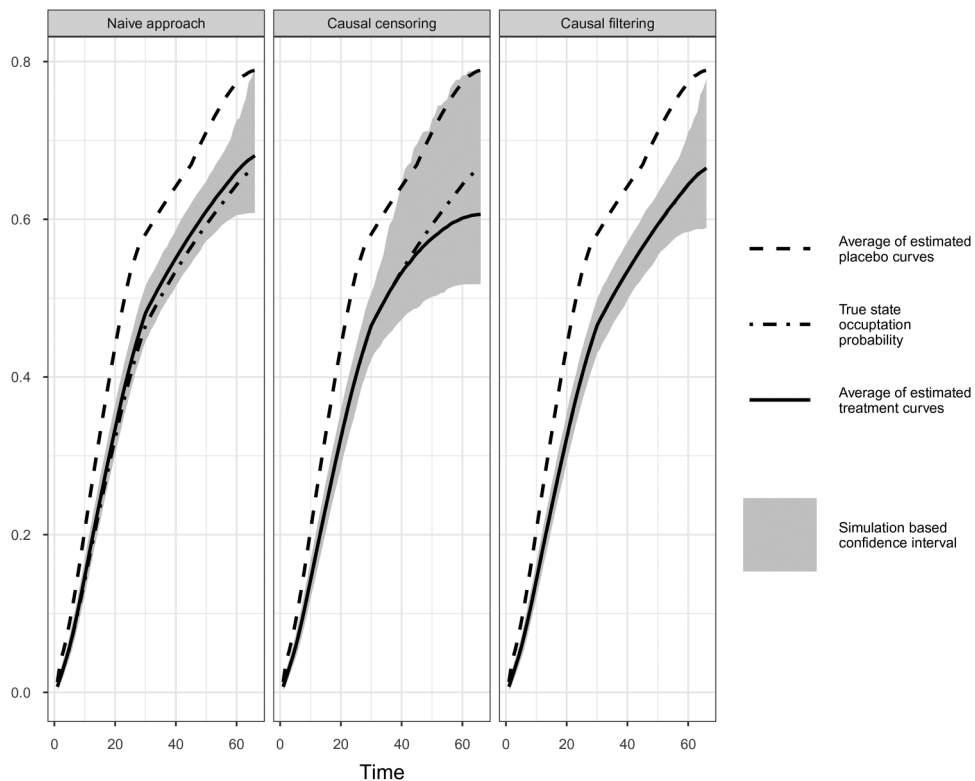
Figure 7. Comparison of simulation results of naïve, causal censoring and causal filtering approach. Simulation scenario I: $\exp(\beta_{02}) = 0.6$, $\exp(\beta_{12}) = 1.0$, $\exp(\beta_{42}) = 0.6$.
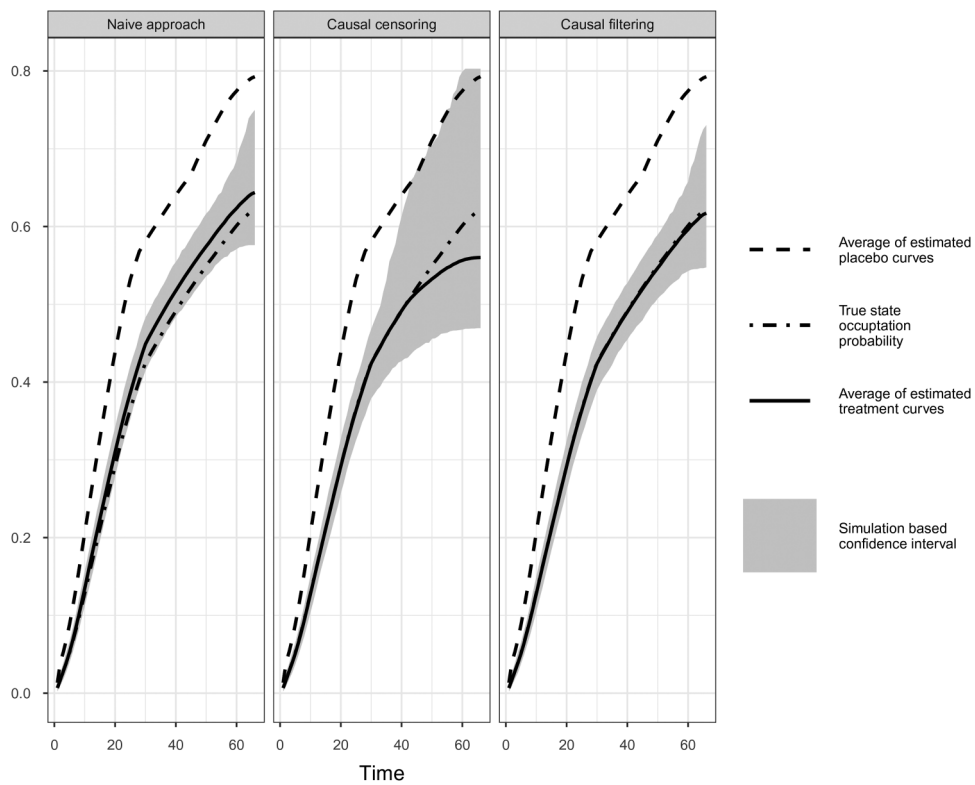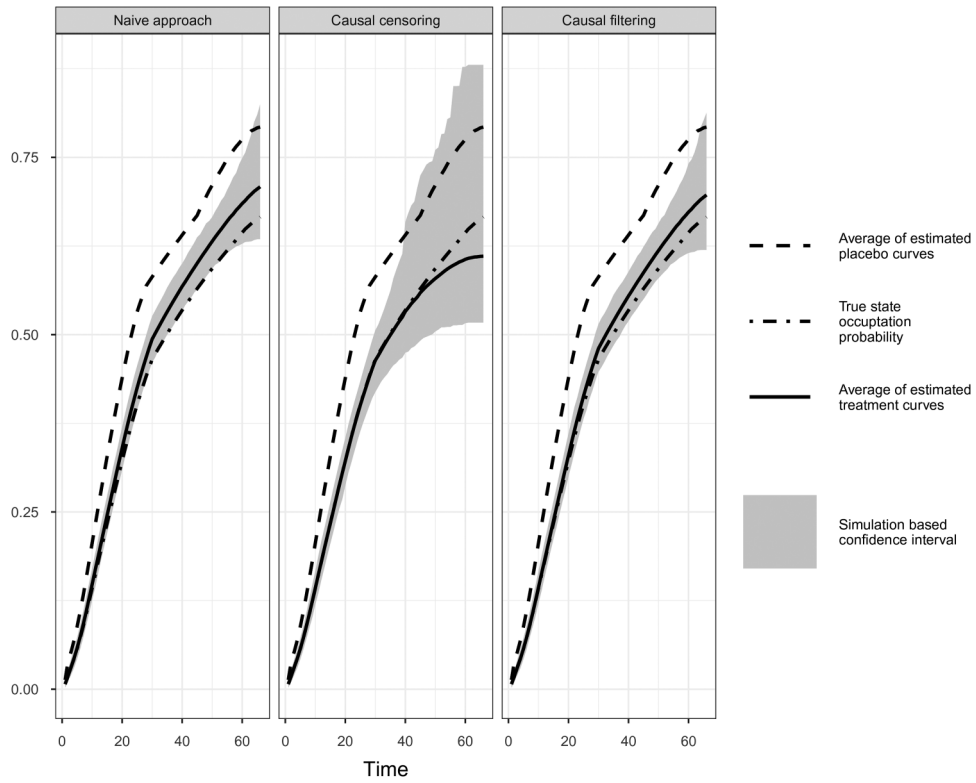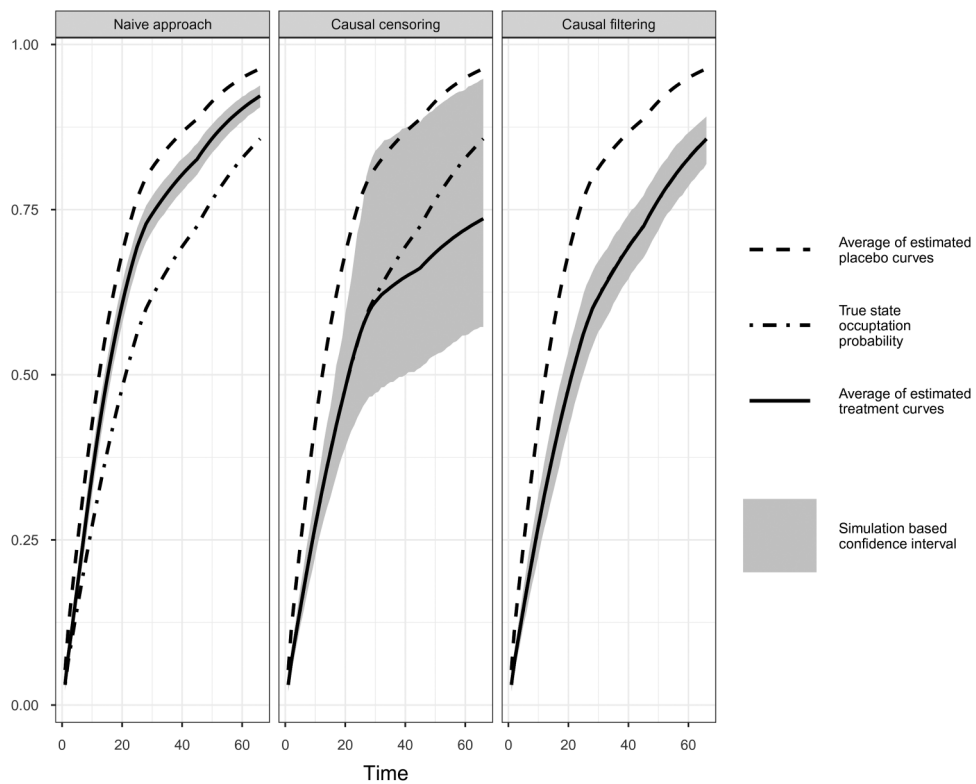

Figure 8. Comparison of simulation results of naïve, causal censoring and causal filtering approach. Simulation Scenario II: $\exp(\beta_{02}) = 0.5$, $\exp(\beta_{12}) = 1.1$, $\exp(\beta_{42}) = 0.5$.

Figure 9. Comparison of simulation results of naïve, causal censoring and causal filtering approach. Simulation Scenario IV: $\exp(\beta_{02}) = 0.6$, $\exp(\beta_{12}) = 1.0$, $\exp(\beta_{42}) = 0.9$.

Figure 10. Comparison of simulation results of naïve, causal censoring and causal filtering approach. Simulation Scenario V.

individuals—had they been intervened—such that probabilities from a hypothetical world, where the intervention is active, can be estimated.

An example where independent censoring does not (generally) provide causal estimates is censoring by competing risks. Censoring by the competing event implies that censoring depends for each individual on the type of event that occurred. This kind of censoring is independent with regard to the identification of the hazard of the cause-specific event of interest. However, there will be (unmeasured/unknown) causal parents that affect both types of event. Consequently, we do not know whether the cause-specific hazards would alter in a world where one could only die from one cause of failure, if the competing risks are different causes of death. Moreover, we could (generally) not define a plausible and realistic intervention to remove the competing event.[25] Consequently, the effect of treatment on the event of interest in a world without competing event could not (generally) be identified from the observed data, as neither the consistency nor the exchangeability assumption is fulfilled. However, there might be situations where the competing event is nothing that inevitably has to happen, and, thus, a realistic intervention could be constructed. Transplantation as competing risk for dialysis patients is such an example.[26]

In simulation studies, we compared the naïve approach, which simply ignores the CH, with a causal censoring and a filtering approach for different scenarios. We found that both approaches provide causal estimates of the treatment effect which would had been observed in the absence of the CH. Under the additional assumption of the treatment effect being restored after lifting the CH, the filtering approach is preferable as it incorporates the fact that most of the patients resume treatment after the CH. Thus, filtering leads to more observed event times and, especially at late time-points, the estimates of the transition probabilities using filtering are more accurate and the simulation-based CIs are smaller provided that the treatment effect is more or less restored after the end of the CH. This assumption can be checked, for instance, by comparing the Nelson-Aalen estimates before and after the treatment interruption due to the CH. In terms of statistical inference, the standard asymptotic results and mathematical arguments for counting processes[15] can be applied to our censoring and filtering approaches, under the assumption of censoring or filtering being independent. Moreover, bootstrap techniques are applicable and are a convenient approach for constructing confidence intervals.[27,28] However, the Aalen-Johansen estimates as well as the relative risks that we have considered in the simulations studies are determined at specified time-points. A line of future research would be to provide a single estimate for the contrast between control and treatment groups including confidence intervals and *p*-values. An option might be to use restricted mean survival time (RMST) analysis, Royston and Parmar, e.g. [29] which has the advantage that it can deal with non-proportional hazards. Estimation of the RMST is straightforward using the Aalen-Johansen estimator and allows the application of re-sampling methods like the wild bootstrap[30] that also applies in non i.i.d. scenarios like event-driven trials.[31,32] Another possibility would be the multistate resampling method proposed in Bluhmki et al.[33] that uses Nelson-Aalen estimates to generate bootstrap data sets from which we could derive bootstrapped hazard ratios. The trick here is that the method of Bluhmki et al.[33] allows to generate synthetic bootstrap data sets after causal censoring or filtering.

In addition to the censoring and filtering approaches, we performed an analysis of a treatment effect in the absence of CH using the popular structural accelerated failure time model (SAFTM). Theoretical considerations and the simulation results are included in the Supplemental Materials. The SAFTM also makes a number of strong assumptions. Besides the assumptions of rank-preservation and "no unmeasured confounders," an equal treatment effect for patients switching to treatment as for those who initially receive the treatment is assumed. In contrast to our non-parametric censoring approach, the SAFTM assumes a parametric and deterministic relationship between the counterfactual event times and the observed event times. Compared to the naïve approach, the SAFTM reduced the bias induced by the treatment interruption considerably in our analyses. As measured by the derived hazard ratios, the results of the SAFTM are comparable, but not better than the results of our proposed censoring and filtering approaches.

The considerations within this paper also considerably extend earlier investigations of Nießl et al.,[2] who proposed a multistate model incorporating the CH, analyzed the impact of the CH, and suggested several methods to account for its impact, including a causal censoring approach that censors all patients at the start of the CH regardless of treatment group and progression status. However, Nießl et al.[2] did not provide a formal and general argumentation. One advantage of the censoring and filtering approaches considered in the present article is that the number of observed events included in the analysis is much higher confirming that the understanding of the causal relations might help to improve the analysis strategy. In the Supplemental Materials, the censoring approach that censors all patients (see "Censoring by CH" in Section 3.3.) and the mITT analysis, as applied to the START study and as described in Nießl et al., are applied to the simulated scenarios I to IV and compared with our censoring and filtering approaches. Overall, we can state that the mITT analysis leads to a small bias for all time points. The censoring method is unbiased for early time points, but shows a large bias for later time points (the bias is larger than for the causal censoring approach), which can be explained by the smaller number of observed events compared to the other methods.

We also applied our new methods to the original START trial (results not shown). Until month 30 we see hardly no differences between the Aalen-Johansen estimates of mITT, naïve and the causal method and only a moderate treatment effect. This is in line with previous analyses.[1,2] After month 30, the causal censoring method approaches more clearly the placebo curve compared to the other methods, which is probably due to the small number of events, as also observed in the simulations.

As Nießl et al.,[2] we assumed that our multistate model fulfills the Markov assumption common in causal reasoning,[18] which is a strong assumption. This implies that we assume that progressive patients have the same risk to die regardless of whether they have experienced the CH or not in the past. An alternative would be to consider a semi-Markov model, in which the post-progression hazard depends on the time since progression. However, censoring on the post-progression time scale will lead to dependent censoring in a non-Markov model and conditioning on an intercurrent event will not lead to a causal analysis. Another alternative is to model the post-progression hazards with time to progression as a covariate. Overall, it is a topic of future research to discuss the possibilities and implications of causal censoring also for non-Markov settings, the latter being an active field of research for 'standard' (not necessarily causal) multistate models.[31]

Another interesting point for future research are causal censoring approaches in the presence of interval-censored data. Interval censoring implies that the occurrence of an event is only known to fall in some time interval, but the exact event time is not known. We refer to Chen et al.[34] for general methods for interval-censored data including an application for estimating a causal effect in the presence of interval-censored data. The connection to our motivating trial is that occurrence of progression is typically interval censored. Hence, our intermediate state has to be interpreted as progression diagnosis, which is not interval censored. It is progression diagnosis that informs treatment interruption or start of a second line treatment.

Our developments allow to apply causal censoring or filtering in general settings with externally induced non-adherence to the study protocol. For example, one possible application of the causal censoring would be to answer the question: What would be the treatment effect of a clinical trial in the absence of COVID-19? The pandemic has caused many direct and indirect effects affecting the planned conduct of a clinical trial. The time of censoring could be a defined anchor date indicating the start of indirect impact of the pandemic.[7]

Declaration of conflicting interests

Funding

ORCID iD

Alexandra Erdmann https://orcid.org/0000-0002-8582-9664

Supplemental material

Supplemental material for this article is available online.

References

1. Butts C, et al. Tecemotide (l-blp25) versus placebo after chemoradiotherapy for stage iii non-small-cell lung cancer (start): a randomised, double-blind, phase 3 trial. *Lancet Oncology* 2014; **15**: 59–68.
2. Nießl A, Beyersmann J and Loos A. Multistate modeling of clinical hold in randomized clinical trials. *Pharm Stat* 2020; **19**: 262–275.
3. Robins J. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Math Model* 1986; **7**: 1393–1512.
4. Gran JM, Lie SA, Øyeflaten I, et al. Causal inference in multi-state models—sickness absence and work for 1145 participants after work rehabilitation. *BMC Public Health* 2015; **15**: 1–16.
5. Aalen O, Borgan O and Gjessing H. *Survival and event history analysis: a process point of view*. Springer Science & Business Media, 2008.
6. Keiding N. Event history analysis and inference from observational epidemiology. *Stat Med* 1999; **18**: 2353–2363.
7. Venkatakrishnan K, Yalkinoglu O, Dong JQ, et al. Challenges in drug development posed by the Covid-19 pandemic: an opportunity for clinical pharmacology. *Clin Pharmacol Ther* 2020; **108**: 699–702.
8. van Geloven N, Swanson SA, Ramspek CL, et al. Prediction meets causal inference: the role of treatment in clinical prediction models. *Eur J Epidemiol* 2020; **35**: 619–630.

9. Ryalen PC, Stensrud MJ and Røysland K. Transforming cumulative hazard estimates. *Biometrika* 2018; **105**: 905–916.
10. Beyersmann J, Allignol A and Schumacher M. *Competing risks and multistate models with R*. New York: Springer Science & Business Media, 2012.
11. Meller M, Beyersmann J and Rufibach K. Joint modeling of progression-free and overall survival and computation of correlation measures. *Stat Med* 2019; **38**: 4270–4289.
12. Kalbfleisch JD and Prentice RL. *The statistical analysis of failure time data*. New York: John Wiley & Sons, 2011.
13. Andersen PK and Keiding N. Multi-state models for event history analysis. *Stat Methods Med Res* 2002; **11**: 91–115.
14. Klein JP and Moeschberger ML. *Survival analysis: techniques for censored and truncated data*. New York: Springer Science & Business Media, 2006.
15. Andersen PK, Borgan O, Gill RD, et al. *Statistical models based on counting processes*. New York: Springer Science & Business Media, 1993.
16. Martinussen T and Scheike TH. *Dynamic regression models for survival data*. New York: Springer Science & Business Media, 2007.
17. Hernán MA and Robins JM. *Causal inference: what if*. Boca Raton: Chapman & Hill/CRC, 2020.
18. Pearl J. *Causality*. Cambridge: Cambridge University Press, 2009.
19. Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. *J Educ Psychol* 1974; **66**: 688–701.
20. Greenland S, Pearl J and Robins JM. Causal diagrams for epidemiologic research. *Epidemiology* 1999; 10: 37–48.
21. Maathuis M, Drton M, Lauritzen S, et al. *Handbook of graphical models*. New York: CRC Press, 2018.
22. Pearl J. An introduction to causal inference. *Int J Biostat* 2010; **6**: 1–62.
23. Spirtes P, Glymour CN, Scheines R, et al. *Causation, prediction, and search*. Cambridge: MIT press, 2000.
24. Didelez V. Causal concepts and graphical models. In: *Handbook of Graphical Models*. New York: CRC press, 2018, pp. 353–380.
25. Young JG, Stensrud MJ, Tchetgen Tchetgen EJ, et al. A causal framework for classical statistical estimands in failure-time settings with competing events. *Stat Med* 2020; **39**: 1199–1236.
26. van Geloven N, le Cessie S, Dekker FW, et al. Transplant as a competing risk in the analysis of dialysis patients. *Nephrol Dial Transplant* 2017; **32**: ii53–ii59.
27. Davison A and Hinkley D. *Bootstrap methods and their application*. Cambridge: Cambridge University Press, 1997.
28. van de Laan M and Robins J. *Unified methods for censored longitudinal data and causality*. New York: Springer, 2013.
29. Royston P and Parmar MK. Restricted mean survival time: an alternative to the hazard ratio for the design and analysis of randomized trials with a time-to-event outcome. *BMC Med Res Methodol* 2013; **13**: 152.
30. Bluhmki T, Schmoor C, Dobler D, et al. A wild bootstrap approach for the Aalen–Johansen estimator. *Biometrics* 2018; **74**: 977–985.
31. Nießl A, Allignol A, Beyersmann J, et al. Statistical inference for state occupation and transition probabilities in non-Markov multi-state models subject to both random left-truncation and right-censoring. *Econometrics and Statistics* 2021. https://doi.org/10.1016/j.ecosta.2021.09.008
32. Rühl J, Beyersmann J and Friedrich S. General independent censoring in event-driven trials with staggered entry. *Biometrics* 2022 https://doi.org/10.1111/biom.13710.
33. Bluhmki T, Putter H, Allignol A, et al. Bootstrapping complex time-to-event data without individual patient data, with a view toward time-dependent exposures. *Stat Med* 2019; **38**: 3747–3763.
34. Chen DGD, Sun J and Peace KE. *Interval-censored time-to-event data: methods and applications*. New York: CRC Press, 2012.