UNIVERSITÄT AUGSBURG
Universitätsbibliothek

OPEN ACCESS
OPUS AUGSBURG

General independent censoring in event driven trials with staggered entry

Jasmin Rühl, Jan Beyersmann, Sarah Friedrich

Angaben zur Veröffentlichung / Publication details:

Rühl, Jasmin, Jan Beyersmann, and Sarah Friedrich. 2023. "General independent censoring in event-driven trials with staggered entry." *Biometrics* 79 (3): 1737–48. https://doi.org/10.1111/biom.13710.

Nutzungsbedingungen / Terms of use:

CC BY-NC-ND 4.0

Dieses Dokument wird unter folgenden Bedingungen zur Verfügung gestellt: / This document is made available under these conditions:
CC-BY-NC-ND 4.0: Creative Commons: Namensnennung - Nicht kommerziell - Keine Bearbeitung
Weitere Informationen finden Sie unter: / For more information see:
https://creativecommons.org/licenses/by-nc-nd/4.0/deed.de

Check for updates

Received: 13 December 2021 | Accepted: 15 June 2022

DOI: 10.1111/biom.13710

BIOMETRIC METHODOLOGY

Biometrics A JOURNAL OF THE INTERNATIONAL BIOMETRIC SOCIETY WILEY

General independent censoring in event-driven trials with staggered entry

Jasmin Rühl[1] | **Jan Beyersmann**[2] | **Sarah Friedrich**[1]

[1]Department of Mathematical Statistics and Artificial Intelligence in Medicine, University of Augsburg, Augsburg, Germany

[2]Institute of Statistics, Ulm University, Ulm, Germany

Correspondence
Jasmin Rühl, Department of Mathematical Statistics and Artificial Intelligence in Medicine, University of Augsburg, Augsburg, Germany.
Email: jasmin.ruehl@math.uni-augsburg.de

Funding information
Deutsche Forschungsgemeinschaft (DFG), Grant FR 4121/2-1

Abstract

Randomized clinical trials with time-to-event endpoints are frequently stopped after a prespecified number of events has been observed. This practice leads to dependent data and nonrandom censoring, which can in general not be solved by conditioning on the underlying baseline information. In case of staggered study entry, matters are complicated substantially. The present paper demonstrates that the study design at hand entails general independent censoring in the counting process sense, provided that the analysis is based on study time information only. To illustrate that the filtrations must not use abundant information, we simulated data of event-driven trials and evaluated them by means of Cox regression models with covariates for the calendar times. The Breslow curves of the cumulative baseline hazard showed considerable deviations, which implies that the analysis is disturbed by conditioning on the calendar time variables. A second simulation study further revealed that Efron's classical bootstrap, unlike the (martingale-based) wild bootstrap, may lead to biased results in the given setting, as the assumption of random censoring is violated. This is exemplified by an analysis of data on immunotherapy in patients with advanced, previously treated nonsmall cell lung cancer.

KEYWORDS

counting process, event-driven trial, independent censoring, staggered study entry

1 | INTRODUCTION

In contrast to other clinical studies, time-to-event studies are driven by the effective sample size, that is, the number of observed events, rather than by the number of recruited subjects. If the number of events is planned in advance, the respective study is called "event-driven." Many examples of such trials can be found in the literature, see, for example, Elisei et al. (2013) in the field of oncology, Sitbon et al. (2015) and McLaughlin et al. (2015) in the pulmonary vascular area, Husain et al. (2019) addressing diabetes and cardiovascular disease, and, in light of the ongoing COVID-19 pandemic, Baden et al. (2021).

The difference between effective and actual sample size is usually caused by right-censoring, which masks the time between the end of the observation period and the event of interest. Supposing simultaneous entry times, simple type I censoring is a consequence of a prespecified time point at which a trial ends, whereas simple type II censoring occurs in case that the follow-up period is stopped after a fixed number of events has been observed (Andersen, 2005). For the analysis of a study that is subject

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.
© 2022 The Authors. *Biometrics* published by Wiley Periodicals LLC on behalf of International Biometric Society.

to the latter, it is important to note that the acquired data include dependencies, since the time point of the last observed event determines the information on every subsequent occurrence: It is only known that censored events must have happened after this time point (or not at all, depending on the event of interest). There are no further consequences for the analysis if all subjects enter the study at the same time: The common time-to-event methods are valid under the assumption of independent censoring in the counting process sense (Andersen et al., 1993, p. 139), which is a substantially less restrictive requirement than that of random censoring. Loosely speaking, independent censoring applies if the intensity of the observable counting process corresponds to the intensity in the (hypothetical) model without censoring, except for a (predictable) at-risk indicator. It is easy to see that this condition is fulfilled in event-driven trials with simultaneous entry, that is, trials subject to simple type II censoring (Aalen et al., 2008, p. 59). Settings that additionally involve external left-truncation (or delayed entry) are likewise covered by the respective considerations (cf. Aalen et al., 2008, p. 32) and are not of interest here.

In case of staggered study entry, the situation is not as clear, though. This scenario is more relevant in the practice of clinical trials and of higher complexity than the situation with one common entry time: Due to the projection of the event times onto the study time scale, an additional source of randomness is introduced, and the previous considerations are no longer justified. The specific case of event-driven trials with staggered entry is not regarded in any of the literature sources we reviewed, and even though there seems to be no formal proof showing that the common time-to-event methods are valid, they are still often applied. What is more, care needs to be taken when searching for literature on the topic, as some authors consider a different definition of independent censoring than the one we adhere to in the following (cf. eg, O'Quigley, 2008, p. 122; Kleinbaum and Klein, 2012, p. 38). An overview of the different concepts of censoring according to Andersen et al. (1993) is given in Andersen (2005). It is further worth mentioning that independent right-censoring is still a relevant subject in the literature, see, for instance, Overgaard and Hansen (2021) for a recent discussion on the topic.

In this paper, we make use of the counting process representation of time-to-event data to demonstrate that event-driven trials with staggered entry entail independent censoring in the counting process sense. An essential condition is that all calendar time information is excluded from the analysis, which ensures that the sequence of the events is concealed. By means of simulations, we reveal two potential issues with the analysis of studies conducted in the described setting in order to make statisticians aware of biased outcomes and ways to avoid them. First, we illustrate the adverse effect on the analysis if calendar times are included in spite of the condition mentioned earlier. Besides, a second simulation study shows that for small sample sizes, techniques which require random censoring can cause substantial bias in the context of event-driven trials with staggered entry. Even though the study population in a clinical trial is typically larger than the samples for which our simulations revealed deviations, the subsets considered in interim analyses may very well be affected by the mentioned bias: Interim analyses are often also event-driven, while taking only a fraction of the final sample size into account. As an example, we consider the OAK study (NCT02008227; Rittmeyer et al., 2017), a randomized phase III study investigating cancer immunotherapy in patients with advanced, previously treated nonsmall cell lung cancer.

The remainder of this manuscript is structured as follows: In Section 2, we introduce the notation while summarizing the counting process representation of time-to-event data and clarifying the formal definition of independent censoring. Section 3 outlines the proof that demonstrates independent censoring in event-driven trials with staggered entry. The simulation studies are presented in Sections 4 and 5, and in Section 6, we analyze a data subset of the OAK study. Eventually, Section 7 comprises a discussion of the consequences for practical applications.

2 | COUNTING PROCESSES AND INDEPENDENT CENSORING

Throughout this paper, we adhere to the notation used by Andersen et al. (1993).

Let $\mathcal{T} = [0, \tau]$ be the study time interval, where τ is a given terminal time. Consider a probability space $(\Omega, \mathcal{F}, P)$ and let the (common) hazard rate $\alpha(t)$ determine each participant's absolutely continuous survival time T_i $(i = 1, \dots, n)$, that is, the duration of the time period between study entry and the occurrence of the event of interest. In addition, let C_i be the censoring time of the ith subject, which is also defined on the study time scale.

We first focus on a hypothetical model where the data are not censored, as indicated by the superscript "c" (for "complete"). This model provides the framework to define the true parameters of interest. We consider the right-continuous counting process $\{N^c(t),\ t \in \mathcal{T}\}$, which jumps at the times of the events and is constant otherwise, or more formally, $N^c(t) = \sum_{i=1}^{n} \mathbb{1}\{T_i \le t\}$, where $\mathbb{1}\{\cdot\}$ denotes the indicator function, $N^c(0) = 0$ and $N^c(t) < \infty$ for all $t \in \mathcal{T}$. By definition, N^c is adapted to the filtration $(\mathcal{F}_t^c)_{t\in\mathcal{T}}$ that is generated by the counting process itself, such that the past $\mathcal{F}_{t-}^c$ includes all the information available just prior to time t. The (predictable) intensity process $\lambda^{\mathcal{F}^c}$ is

15410420, 2023, 3, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/biom.13710 by Universitaetsbibl Augsburg, Wiley Online Library on [10/01/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

further defined by $\lambda^{\mathcal{F}^c}(t)\,dt = P(dN^c(t) = 1 \mid \mathcal{F}^c_{t-})$, where $dN^c(t)$ refers to the increment of N^c over the infinitesimal interval $[t, t + dt)$. As Aalen (1978) showed with his "multiplicative intensity model," $\lambda^{\mathcal{F}^c}(t)$ can be expressed as the product of the at-risk process $Y^c(t) = \sum_{i=1}^n \mathbb{1}\{T_i \geq t\}$ and the hazard rate $\alpha(t)$.

The next step is to extend the described model such that censored data are accommodated. The required concepts are those actually used for the analysis of time-to-event data. Suppose that we do not know the individual values of the event and censoring times T_i and C_i, but only observe their minimum as well as the value of the indicator $\mathbb{1}\{T_i \leq C_i\}$ $(i = 1, \ldots, n)$. Thus, it is necessary to define an adjusted càdlàg counting process $\{N(t),\ t \in \mathcal{T}\}$ that jumps at the times of observed events only, that is, $N(t) = \sum_{i=1}^n \mathbb{1}\{T_i \leq t,\ T_i \leq C_i\}$. As before, we assume that $N(0) = 0$, $N(t) < \infty$ for all $t \in \mathcal{T}$, and that N is adapted to a filtration $(\mathcal{F}_t)_{t\in\mathcal{T}}$ with corresponding past $\mathcal{F}_{t-}$. The latter includes merely observable information that has been obtained before time t. One may further extend the filtration by baseline information. Analogously to the definition above, the intensity process associated with N is characterized by $\lambda^{\mathcal{F}}(t)\,dt = P(dN(t) = 1 \mid \mathcal{F}_{t-})$. If we assume random censoring, that is, (stochastic) independence between the survival and censoring times, Aalen's multiplicative intensity model is still valid with respect to the adapted at-risk process $Y(t) = \sum_{i=1}^n \mathbb{1}\{T_i \geq t,\ C_i \geq t\}$. However, random censoring is not fulfilled for event-driven censoring, which is why we consider a less restrictive notion that still permits sound inference for time-to-event data.

To establish this notion, a third filtration $(\mathcal{G}_t)_{t\in\mathcal{T}}$ has to be defined: $\mathcal{G}_{t-}$ complements the past $\mathcal{F}^c_{t-}$ with details about the censoring process, such that both the uncensored and the observable representation of the data are considered. $(\mathcal{F}_t)_{t\in\mathcal{T}}$ is therefore nested within $(\mathcal{G}_t)_{t\in\mathcal{T}}$ and the counting process N is adapted to $(\mathcal{F}_t)_{t\in\mathcal{T}}$ as well as $(\mathcal{G}_t)_{t\in\mathcal{T}}$. By the law of total expectation, it can be shown that the intensities relative to $(\mathcal{F}_t)_{t\in\mathcal{T}}$ and $(\mathcal{G}_t)_{t\in\mathcal{T}}$ are generally not equal (Aalen et al., 2008, pp. 56–57).

Independent censoring defines a very general condition that still allows to obtain unbiased results. A censoring process is independent in the sense of Andersen et al. (1993) if it does not provide any information that alters the intensity of a subject at risk, or in more formal terms, if for all $i \in \{1, \ldots, n\}$ and for all $t \in \mathcal{T}$,

$$\lambda_i^{\mathcal{G}}(t) = \lambda_i^{\mathcal{F}^c}(t), \tag{1}$$

where $\lambda_i^{\mathcal{G}}(t)\,dt = P(T_i \in [t, t + dt) \mid \mathcal{G}_{t-})$ and $\lambda_i^{\mathcal{F}^c}(t)\,dt = P(T_i \in [t, t + dt) \mid \mathcal{F}^c_{t-})$ (Andersen et al., 1993, p. 139). This condition has very useful implications: Assuming that Equation (1) holds, and that the hazard $\alpha(t)$ is $(\mathcal{F}_t)$-predictable for $t \in \mathcal{T}$, it is easy to see that the multiplicative intensity model is valid for the observable intensity $\lambda^{\mathcal{F}}$ (Aalen et al., 2008, p. 60).

3 | INDEPENDENT CENSORING IN EVENT-DRIVEN TRIALS WITH STAGGERED ENTRY

In the further course of this paper, we will focus on the specific case of type II censoring with staggered patient entry. To the best of our knowledge, there are no sources that provide evidence for the validity of independent censoring in this context. Most literature sources that address incomplete observations merely cover random censoring, and elsewhere, the definition of independent censoring is not consistent with the one considered here (cf. the discussion in Martinussen and Scheike, 2006, pp. 52–57).

Before dealing with the proof that Equation (1) is indeed satisfied in event-driven trials with staggered entry, one should be aware of the need to switch between two different time scales and the challenges that come along with it: On the one hand, censoring times are determined by the chronological sequence of the events in case of type II censoring, and as subjects enter successively, we need to work on the calendar time scale in order to characterize the data in the study setting at hand. However, for the analysis of time-to-event data, the study time scale of Section 2 has to be considered. The issue here is that it is not clear what consequences follow from the dependence structure of the data after the switch to the new time scale.

Theorem 1. *Consider a study with n subjects and suppose that their event times are independent. Let s_i and t_i denote the calendar times of study entry and the event of interest for subject i, such that $s_i < t_i$ $(i = 1, \ldots, n)$. Let further $t_{(1)} < t_{(2)} < \cdots < t_{(n)}$ be the ordered event times (in calendar time scale), assuming no ties. We suppose that the trial is event-driven, that is, the observation period is specified to end at time $t_{(m)}$ for a fixed value of m with $0 < m < n$. Then, it holds for all $i \in \{1, \ldots, n\}$ and for all $t \in \mathcal{T}$ that $\lambda_i^{\mathcal{G}}(t) = \lambda_i^{\mathcal{F}^c}(t)$.*

Figure 1 shows the possible scenarios in the case where $n = 2$ and $m = 1$. With respect to the study time scale, the survival and censoring times are given by $T_i = t_i - s_i$ and $C_i = \min(t_1, t_2) - s_i$ $(i = 1, 2)$, so that the observable information under these conditions can be represented as

$$\begin{aligned} &\Big(\min(T_1, C_1),\ \mathbb{1}\{T_1 \leq C_1\}\Big), \quad \Big(\min(T_2, C_2),\ \mathbb{1}\{T_2 \leq C_2\}\Big), \\ = &\Big(\min(t_1, t_2) - s_1,\ \mathbb{1}\{t_1 \leq t_2\}\Big), \quad \Big(\min(t_1, t_2) - s_2,\ \mathbb{1}\{t_2 \leq t_1\}\Big). \end{aligned}$$

This illustrates that the data of both subjects are determined by each other, and thus, we refer to this particular setting as a "maximal dependence case" (for $n = 2$).

15410420, 2023, 3, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/biom.13710 by Universitaetsbibl Augsburg, Wiley Online Library on [10/01/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

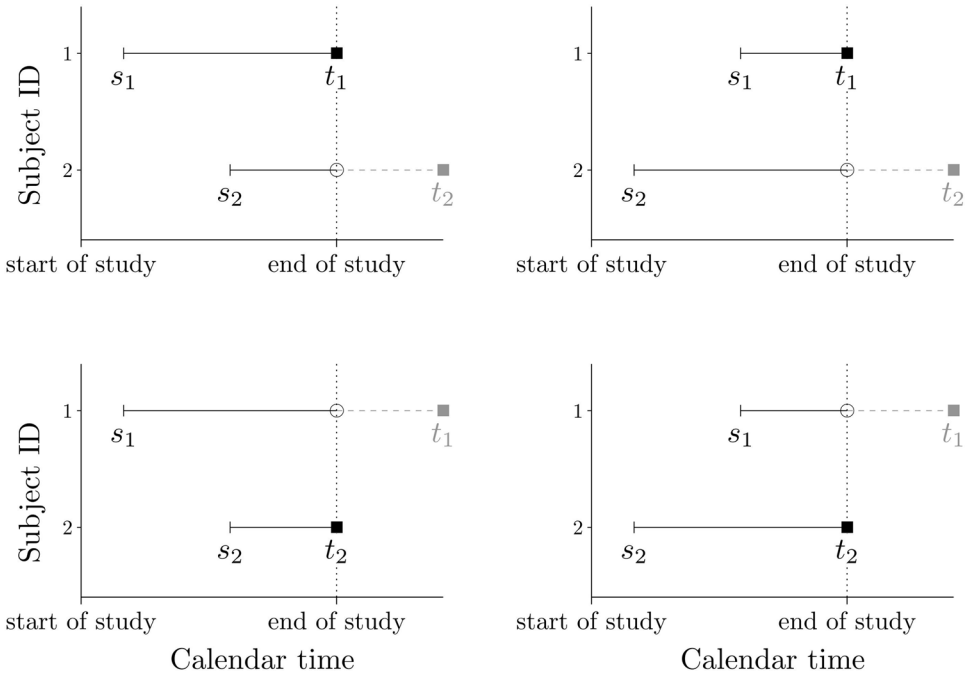

FIGURE 1 Possible study scenarios in the maximal dependence case ($n = 2$, $m = 1$). Squares represent events, while circles indicate censoring

Proof. To prove independent censoring in the general context of event-driven trials with staggered entry, we consider the intensity processes relative to $(\mathcal{F}_t^c)_{t\in\mathcal{T}}$ and $(\mathcal{G}_t)_{t\in\mathcal{T}}$, which are based on the study time scale. The former intensity is defined by

$$\begin{aligned}\lambda_i^{\mathcal{F}^c}(u)\,du &= P\Big(T_i \in [u, u+du) \,\Big|\, \Big\{\big(\mathbb{1}\{T_j \le v\}\big)_{j=1,\dots,n} : v < u\Big\}\Big)\\ &= \mathbb{1}\{T_i \ge u\}\,P(T_i \in [u, u+du) \mid T_i \ge u),\end{aligned}$$

where the last equivalence results from the independence of the uncensored event times. Moreover, the second intensity process

$$\begin{aligned}\lambda_i^{\mathcal{G}}(u)\,du = P\,(&T_i \in [u, u+du) \mid\\ &\Big\{\big(\mathbb{1}\{T_j \le v\},\, \mathbb{1}\{C_j \ge v\}\big)_{j=1,\dots,n} : v < u\Big\}\Big)\end{aligned}$$

can also be expressed as the product of $\mathbb{1}\{T_i \ge u\}$ and

$$\begin{aligned}&P(T_i \in [u, u+du) \mid T_i \ge u,\\ &\{\big(\mathbb{1}\{T_j \le v\},\, \mathbb{1}\{C_k \ge v\}\big)_{\substack{j=1,\dots,i-1,i+1,\dots,n\\ k=1,\dots,n}} : v < u\}).\end{aligned}$$

Without loss of generality, set $i = 1$. To prove that $\lambda_1^{\mathcal{G}}(u)\,du = \lambda_1^{\mathcal{F}^c}(u)\,du$, the data are represented using calendar times, that is, $T_j = t_j - s_j$ and $C_j = t_{(m)} - s_j$ ($j = 1, \dots, n$). Keep in mind, however, that the values of t_j and s_j are not used in the analysis. We therefore need to show that $P(t_1 - s_1 \in [u, u+du) \mid t_1 - s_1 \ge u)$ is equal to

$$\begin{aligned}&P(t_1 - s_1 \in [u, u+du) \mid t_1 - s_1 \ge u,\\ &\{\big(\mathbb{1}\{t_j - s_j \le v\},\, \mathbb{1}\{t_{(m)} - s_k \ge v\}\big)_{\substack{j=2,\dots,n\\ k=1,\dots,n}} : v < u\}),\end{aligned} \tag{2}$$

or equivalently, that the condition on the set

$$(*) = \{\big(\mathbb{1}\{t_j - s_j \le v\},\, \mathbb{1}\{t_{(m)} - s_k \ge v\}\big)_{\substack{j=2,\dots,n\\ k=1,\dots,n}} : v < u\}$$

does not imply further information on $t_1 - s_1$ in both the cases where $t_{(m)} = t_1$ and $t_{(m)} \ne t_1$. The subsequent proof is divided into the two respective scenarios, accordingly.

Case 1. First suppose that $t_{(m)} = t_1$. This means that the indicator $\mathbb{1}\{t_{(m)} - s_1 \ge v\} = \mathbb{1}\{t_1 - s_1 \ge v\}$, with $v < u$, becomes redundant given the condition $t_1 - s_1 \ge u$ in (2). The set $(*)$ can therefore be reduced to $\{(\mathbb{1}\{t_j - s_j \le v\},\, \mathbb{1}\{t_1 - s_j \ge v\})_{j=2,\dots,n} :\ v < u\}$.

With respect to the second indicator function in the expression above, it holds that

$$\begin{aligned}&t_1 - s_j \ge v\\ \iff\quad &t_1 - s_1 \ge v + s_j - s_1.\end{aligned}$$

This inequality could possibly allow inference on $t_1 - s_1$ in case there is a $j \in \{2, \dots, n\}$ and some $v < u$ such that

15410420, 2023, 3, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/biom.13710 by Universitaetsbibl Augsburg, Wiley Online Library on [10/01/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

$t_1 - s_j \geq v$ as well as $v + s_j - s_1 > u$: One would be able to conclude that $t_1 - s_1 > u$ if the calendar times t_1, s_1, and s_j were available, meaning that independent censoring would no longer apply. As the calendar times are not taken into account, though, the required information is only available if it is conveyed by the study time variables T_i and C_i $(i = 1, \dots, n)$. Through the characterization $s_j - s_1 = t_1 - s_1 - (t_1 - s_j) = T_1 - C_j$, we can deduce that

$$\begin{aligned} & v + s_j - s_1 > u \\ \iff \quad & T_1 > u + C_j - v. \end{aligned}$$

Nevertheless, it is still not possible to confirm whether $T_1 > u$ solely by the knowledge of the past. To see this, note that the inequality $T_1 > u + C_j - v$ can only be verified through $\mathbb{1}\{T_1 \geq w\} = 1$ for an observed time point $w > u + C_j - v$. But we assumed that $C_j = t_1 - s_j \geq v$ earlier, and thus, $w > u$, or in other words, the past does not include w. It follows that the knowledge of the censoring times does not affect the probability in Equation (2), and $(*)$ can be further reduced to $\{(\mathbb{1}\{t_j - s_j \leq v\})_{j=2,\dots,n} : v < u\}$.

Conditioning on the remaining indicator functions will not alter (2) either, since the event times are independent by assumption. As a result, we find that the intensity λ_1 is not affected by the condition on $(*)$ if $t_{(m)} = t_1$.

Case 2. To complete the proof, it remains to show that $(*)$ also does not add information on $t_1 - s_1$ in the case where $t_{(m)} \neq t_1$. The censoring times $t_{(m)} - s_k$ are not related to the event time $t_1 - s_1$ for any $k \in \{2, \dots, n\}$ and thus, we can limit our considerations to $t_{(m)} - s_1$. The associated indicator function $\mathbb{1}\{t_{(m)} - s_1 \geq v\}$ only reveals that $t_1 - s_1 > u$ if the past indicates both $t_{(m)} - s_1 \geq v$ for some $v < u$, and $v + t_1 - t_{(m)} > u$, since

$$\begin{aligned} & t_{(m)} - s_1 \geq v \\ \iff \quad & t_1 - s_1 \geq v + t_1 - t_{(m)}. \end{aligned}$$

Once again, the knowledge of the calendar times might hint that $t_1 - s_1 > u$. The given information on the event times is however restricted to the study time variables T_i and C_i $(i = 1, \dots, n)$. Thus, we rewrite the difference $t_1 - t_{(m)}$ as $t_1 - s_1 - (t_{(m)} - s_1) = T_1 - C_1$. This means

$$\begin{aligned} & v + t_1 - t_{(m)} > u, \\ \iff \quad & T_1 > u + C_1 - v, \end{aligned}$$

and in case the relation above holds true, we have no knowledge thereof unless the past includes time $w > u + C_1 - v$. (All that is known before w is that T_1 exceeds the time point just prior to the currently observed one.) The assumption $C_1 = t_{(m)} - s_1 \geq v$ implies that $w > u$, though, and hence, the past does not suffice to provide the necessary information. In summary, the censoring times are dispensable in $(*)$. The set $\{(\mathbb{1}\{t_j - s_j \leq v\})_{j=2,\dots,n} : v < u\}$ remains instead, and it finally follows by the independence of the event times that the condition on $(*)$ also does not change (2) in the case where $t_{(m)} \neq t_1$.

We conclude that the filtration $\mathcal{G}_{u-}$ does not contribute any additional information on the event times in comparison to $\mathcal{F}^c_{u-}$. Thus, Equation (1) can be confirmed. □

The proof above shows that it is crucial not to condition on s_i and t_i $(i = 1, \dots, n)$ at any point during the analysis of event-driven trials with staggered entry. Otherwise, one might be able to predict the event time of interest, and the relevant arguments for independent censoring do not apply any longer.

4 | SIMULATION STUDY I

In order to illustrate the consequences of our findings, we simulated different scenarios of event-driven trials with staggered patient entry and analyzed the data while taking the respective calendar times into account. As implied by the proof in the previous section, such an approach might distort the intensity of the observable counting process. Our main intent was therefore to detect any effects on the outcomes of the analysis.

The study scenarios we considered for this purpose were based on the parameter combinations $(n, m) \in \{(600, 300), (300, 150), (50, 25), (50, 10), (26, 13)\}$, where n denotes the number of trial participants and m is the number of observed events before censoring was imposed. We expected that potential effects become clearer for rather low values of n and m, since, heuristically, small sample sizes increase the dependence within the data. In each simulated study, the subjects were randomly assigned to one of two equal-sized groups (eg, treatment vs. control), with their event times modeled such that a prespecified hazard ratio was achieved. Our investigations involved both exponentially and Weibull distributed event times, in combination with hazard ratios of 0.8, 1, as well as 1.2. The event times in the control group were generated using a scale parameter of 1 for the exponential settings, and shape and scale parameters of 0.5 and 1 for the Weibull scenarios. Moreover, for the entry times, we considered a uniform distribution over the interval between 0 and the m/n quantile of the respective event time distribution. The rationale behind this boundary was to reduce the probability of subject entries after the end of the observation period.

15410420, 2023, 3, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/biom.13710 by Universitaetsbibl Augsburg, Wiley Online Library on [10/01/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

As a next step, the generated data were fitted to a "standard" Cox regression model that only included a single covariate for the treatment group. This model served as a reference, representing the case where the analysis is conducted under independent censoring, so that our simulations also allowed us to examine the practical application of survival methods to event-driven trials with staggered entry. In order to identify the consequences of conditioning on calendar time variables, we investigated two additional, fairly unconventional models. A practical situation where such conditions become relevant has been described by Meyer et al. (2020), who suggested to subdivide trials that were interrupted by the onset of COVID-19 into stages based on calendar times. It should be emphasized that we do not propose the application of calendar time-based models in the given setting, but rather consider them to examine potential bias that may result from disturbed intensities. The first of the models we examined will be referred to as "Model 1" hereinafter, and apart from a covariate for the group, it also included a second predictor reflecting the calendar time of a subject's entry into the study. The information resulting thereof might not be sufficient to interfere with the analysis, though, which is why we considered a third covariate in "Model 2." The value of this covariate describes the number of subjects already recruited at each new trial admission, and together with the participants' entry times, it provides a more direct approximation of the information that is conveyed by the counting processes and the corresponding calendar times.

Our idea was to illustrate the impact of conditioning on calendar time information by comparing the Breslow estimate of the cumulative baseline hazard that is derived from the Standard Model to the estimates based on Model 1 and Model 2, respectively.

To that end, we simulated each of the mentioned study scenarios 100,000 times and summarized the outcomes by means of the mean and median bias of the Breslow estimates at selected time points as well as their root mean square errors.

There happen to be cases where all of the subjects whose events were observed (or all except for one of them) belonged to the same treatment group. This is due to the small values of the parameter m, and as a consequence, the hazard ratio was estimated to be extremely high or low. To ensure meaningful results, we excluded the respective iterations (see Table 1 for the results in the small sample size settings with Weibull distributed event times and an underlying treatment hazard ratio of 1; the complete outcomes are available in Web Appendix A).

While the mean bias of the Breslow estimates in Model 1 only exceeds that in the Standard Model for particularly small sample sizes, the deviations are multiple times higher in model 2 (with the Monte Carlo standard errors taken into account; see Web Appendix A). We found the differences to be especially large at later time points and for smaller values of n and m. What is most interesting, though, is that throughout all considered scenarios and time points, the extent of the median bias and the root mean square error in the two calendar time models are notably higher as compared to the Standard Model.

The outcomes in the exponential setting with $n = 50$ and $m = 10$ are additionally visualized in Figure 2. As can be seen, the median of the Breslow curves is quite close to the true one in the standard model, except for later time points where the amount of available data decreases. For instance, only 10% of the subjects are followed up over a time span that exceeds 0.3316. In Models 1 and 2, on the other hand, the median curves run clearly below the true cumulative baseline hazard right from the beginning. The mean curve in Model 2 further overestimates the true one by far, as there are some individual scenarios with very extreme slope. The curves have similar properties in the remaining scenarios, but the bias is the most obvious in the settings where $m = 10$ and $n = 50$ (see Figure 3 for the case with Weibull distributed event times; the shadow plots for the other scenarios can be found in Web Appendix A). This demonstrates how conditioning on calendar times affects the outcomes of the analysis.

Besides, we also investigated the partial likelihood estimates of the hazard ratio in the different models directly. The outcomes are summarized by means of the mean and median bias of the estimated treatment log hazard ratios, their root mean square error, and the mean coverage of the individual confidence intervals.

Table 2 shows the results in the small sample size settings where the underlying hazard ratio between the treatment groups equals 1. As implied by the last three columns, the differences between the estimates in the distinct models are rather small overall. Throughout every considered scenario, the median bias hardly deviates from 0, whereas the mean bias tends to increase slightly from the tandard Model over Model 1 to Model 2. The increase is likewise reflected by the root mean square errors, which are more pronounced for the settings where n and m are small. With regard to the coverage, the results are in general quite close to each other. Even though the underlying confidence intervals become wider from model to model, the coverage declines slightly, and once again, the differences are somewhat more marked for smaller values of the parameters n and m. All of the mentioned relations apply regardless of whether iterations with less than two observed events in one group are excluded or not. We also did not notice any differences between exponential and Weibull distributed event times.

Lastly, the outcomes for the covariates that are additionally regarded in Model 1 and Model 2, namely, the calendar

15410420, 2023, 3, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/biom.13710 by Universitaetsbibl Augsburg, Wiley Online Library on [10/01/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 1 Bias of the Breslow estimators at selected time points (Weibull scenarios with hazard ratio 1)

Scenario							
Distribution	***n***	***m***	**Time**	**Measure of bias**	**Standard Model**	**Model 1**	**Model 2**
Weibull	50	25	0.20	Mean bias	0.00719	0.00503	0.03234
				Median bias	−0.01245	−0.04391	−0.04878
				RMSE	0.14982	0.24269	0.41598
			0.45	Mean bias	0.03311	0.03284	0.08215
				Median bias	−0.00461	−0.03989	−0.04422
				RMSE	0.24494	0.36744	0.67557
			0.70	Mean bias	0.03774	0.03928	0.10083
				Median bias	−0.02197	−0.05401	−0.05407
				RMSE	0.32263	0.45241	0.79559
Weibull	50	10	0.03	Mean bias	0.01094	0.00138	0.07112
(1034 excluded iterations[a])				Median bias	−0.00859	−0.04163	−0.04837
				RMSE	0.10663	0.20163	8.42391
			0.05	Mean bias	0.02570	0.01440	0.10420
				Median bias	−0.00620	−0.04226	−0.05025
				RMSE	0.17141	0.26164	9.02352
			0.07	Mean bias	0.03267	0.02099	0.12607
				Median bias	−0.00917	−0.04829	−0.05604
				RMSE	0.23794	0.31989	10.00564
Weibull	26	13	0.20	Mean bias	0.02025	0.03434	0.81749
(4 excluded iterations)[a]				Median bias	−0.02054	−0.08097	−0.08725
				RMSE	0.23050	1.29029	94.21344
			0.45	Mean bias	0.06895	0.09690	1.28769
				Median bias	−0.01100	−0.07965	−0.08390
				RMSE	0.41094	1.44409	113.91960
			0.70	Mean bias	0.07190	0.10951	1.44727
				Median bias	−0.03226	−0.10163	−0.09843
				RMSE	0.50253	1.52040	114.97777

Abbreviation: RMSE, root mean square error.

[a]Iterations with less than two observed events in one treatment group are excluded.

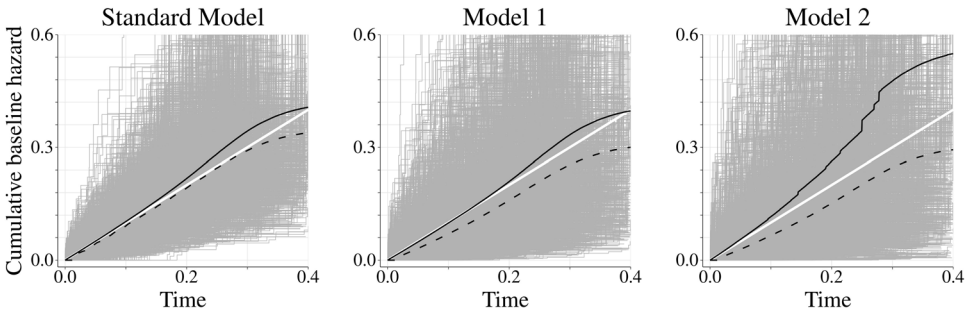

FIGURE 2 Shadow plot of the Breslow estimators in the exponential scenario with hazard ratio 1, $n = 50$, and $m = 10$. The white line shows the true cumulative baseline hazard, the solid and dashed black lines represent the mean and median of the simulated Breslow estimators, respectively. For greater clarity, the shadow lines are restricted to a random sample of size 2000, excluding 1032 iterations with less than two observed events in one treatment group

15410420, 2023, 3, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/biom.13710 by Universitaetsbibl Augsburg, Wiley Online Library on [10/01/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

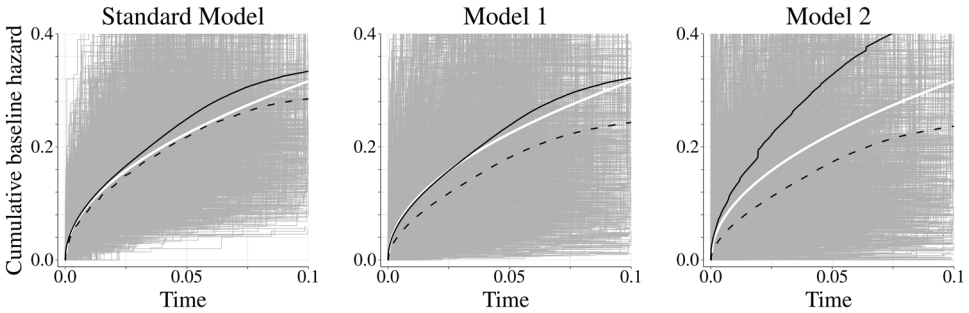

FIGURE 3 Shadow plot of the Breslow estimators in the Weibull scenario with hazard ratio 1, $n = 50$, and $m = 10$. The white line shows the true cumulative baseline hazard, the solid and dashed black lines represent the mean and median of the simulated Breslow estimators, respectively. For greater clarity, the shadow lines are restricted to a random sample of size 2000, excluding 1032 iterations with less than two observed events in one treatment group

TABLE 2 Bias of the estimated log hazard ratios (scenarios with hazard ratio 1)

Scenario						
Distribution	***n***	***m***	**Measure of bias**	**Standard Model**	**Model 1**	**Model 2**
Exponential	50	25	Mean bias	0.00130	0.00142	0.00199
			Median bias	0.00137	0.00097	0.00093
			RMSE	0.41187	0.42149	0.43517
			Coverage	0.95511	0.95193	0.94823
Exponential	50	10	Mean bias	0.00665	0.00687	0.00701
(1032 excluded iterations)[a]			Median bias	0.00233	0.00297	0.00450
			RMSE	0.65962	0.67233	0.69416
			Coverage	0.98131	0.97872	0.97418
Exponential	26	13	Mean bias	−0.00299	−0.00279	−0.00200
(4 excluded iterations)[a]			Median bias	−0.00130	−0.00008	−0.00086
			RMSE	0.59677	0.62824	0.67787
			Coverage	0.95929	0.95326	0.94575
Weibull	50	25	Mean bias	0.00262	0.00282	0.00274
			Median bias	0.00139	0.00155	0.00285
			RMSE	0.41109	0.42120	0.43426
			Coverage	0.95354	0.95114	0.94806
Weibull	50	10	Mean bias	0.00235	0.00263	0.00227
(1034 excluded iterations)[a]			Median bias	0.00258	0.00290	0.00337
			RMSE	0.65654	0.66925	0.69106
			Coverage	0.98294	0.97954	0.97413
Weibull	26	13	Mean bias	−0.00238	−0.00256	−0.00235
(4 excluded iterations)[a]			Median bias	−0.00052	0.00042	0.00165
			RMSE	0.58901	0.62311	0.67014
			Coverage	0.96170	0.95446	0.94719

Abbreviation: RMSE, root mean square error.

[a]Iterations with less than two observed events in one treatment group are excluded.

time of study entry as well as the number of recruited subjects, were evaluated, too (see Web Appendix A). The bias we observed for the latter attains only moderate levels, whereas the estimated hazard ratio for the entry times is highly inaccurate in some of the considered scenarios. In the setting with $n = 50$ and $m = 10$, this analysis yields particularly biased results. Besides, the deviation from the true hazard ratio is more pronounced in case that the event

15410420, 2023, 3, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/biom.13710 by Universitaetsbibl Augsburg, Wiley Online Library on [10/01/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

times follow a Weibull distribution. It thus seems like the covariate for the entry times in fact disturbs the analysis.

All presented analyses were furthermore performed for underlying treatment hazard ratios of 0.8 and 1.25, respectively (see Web Appendix A), but the outcomes did not reveal any significant differences compared to what we observed before.

In summary, the results of the simulations are consistent with our expectations: If the calendar times of study entry are used, the estimated hazard ratios diverge from the true ones, even though the bias for the log hazard ratio of group membership is very moderate. The deviation becomes more notable, the smaller one chooses the values of the parameters n and m (ie, for $n \leq 50$), and also if the proportion m/n is reduced. This follows from the heuristically stronger dependence within the data. The number of recruited subjects further provides additional information on the sequence of the events, which is why the bias increases in Model 2. The impact of conditioning on calendar time information is however reflected more distinctly by the estimated cumulative baseline hazard: Our simulations showed that for sample sizes of 50 and below, the Breslow estimators are very prone to bias in those models where the calendar times are taken into account, and the deviation increases with the amount of details on the sequence of the events. As a consequence, predicted survival probabilities will not be reliable if one conditions on calendar times.

To show that the described effects are in fact caused by the interaction between staggered study entry and type II censoring, we repeated the simulations for the exponential scenarios with hazard ratio 1 and parameters $(n, m) \in \{(50, 25), (50, 10), (26, 13)\}$, but instead of implementing type II censoring, we generated random censoring times. Compared to the previous outcomes, the absolute value of the median bias of the Breslow estimators is notably smaller in the models that include calendar time variables, even though the sample sizes are very small. This is also clearly visualized in the shadow plots of the Breslow curves (see Web Appendix A). Our proposition thus seems to be confirmed.

5 | SIMULATION STUDY II

The proof in Section 3 shows that independent censoring is fulfilled in event-driven trials with staggered entry, however, one cannot assume that random censoring holds. A situation where both conditions come into play is bootstrapping: While Efron's nonparametric bootstrap, that is, drawing with replacement from the data, requires random censoring (Efron, 1981), it is sufficient to consider data that fulfill Aalen's multiplicative intensity model when applying the wild bootstrap proposed by Lin, Wei and Ying (1993) (see also Beyersmann et al., 2013). In event-driven trials with staggered entry, the wild bootstrap is therefore expected to perform superior.

We conducted another simulation study to demonstrate to which extent the accuracy of both resampling methods can differ in practice. Preliminary simulations hinted that the effect is more pronounced in the more complicated illness-death-model compared to the classical survival setting. We thus adhered to the simulation scenario described in Nießl et al. (2021) to generate type II censored illness-death data without recovery, except that our simulations additionally involved staggered entry: The calendar times of $n = 100$ study admissions were sampled from a uniform distribution over the interval between 0 and 60. As in Nießl et al. (2021), the waiting times in the initial state followed an exponential distribution with parameter 0.04, and from there, subjects moved to the state of being ill or dead with probabilities 0.25 and 0.75, respectively. The waiting time for the transition from illness to death was further simulated by random numbers generated from an exponential distribution with parameter 0.1. Eventually, type II censoring was imposed at the time when $m = 50$ subjects had moved to the state of death, regardless of their prior state occupation.

We applied both the usual nonparametric bootstrap as well as the wild bootstrap using 1000 samples, respectively, in order to determine 95% confidence intervals for the Nelson–Aalen estimator. To that end, the transition from illness to death was considered. It should be noted that internal left-truncation as a result of the progression into the illness state additionally complicates inference here. The confidence intervals at times 16, 18, and 20 were computed based on the log-transformed formula given in Andersen et al. (1993, p. 208), but with the standard normal quantiles replaced by the 0.975 quantiles of the studentized bootstrap estimates. With respect to Efron's bootstrap, these estimates are defined by

$$\left(\widehat{A}^{(b)}(t) - \widehat{A}(t)\right) / \sqrt{\widehat{\mathrm{var}}\left(\widehat{A}^{*}(t)\right)} \quad (b = 1, \ldots, 1000).$$

The terms $\widehat{A}^{(b)}$ and $\widehat{A}$ in this formula denote the Nelson–Aalen estimators from the bth bootstrap sample and from the original data set, respectively, and $\widehat{\mathrm{var}}\left(\widehat{A}^{*}(\cdot)\right)$ is the empirical variance of the resampled Nelson–Aalen estimates. The quantiles we used for the wild bootstrap were, on the other hand, based on

$$\sum_{u \leq t} \left(\frac{\mathbb{1}\{Y(u) > 0\}}{Y(u)} \cdot \Delta N(u) \cdot G^{(b)}(u) \right) / \sqrt{\widehat{\mathrm{var}}\left(Z^{*}(t)\right)} \quad (b = 1, \ldots, 1000),$$

where Y refers to the number of subjects at risk, ΔN reflects the number of transitions from illness to death, $G^{(b)}$ are

15410420, 2023, 3, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/biom.13710 by Universitaetsbibl Augsburg, Wiley Online Library on [10/01/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 3 Coverage probabilities (in %) and widths of the bootstrapped confidence intervals at selected time points

	EBS		WBS	
Time	Coverage	Width	Coverage	Width
16	83.0	2.217	95.3	3.298
18	84.3	2.298	95.6	3.455
20	82.0	2.349	95.9	3.640

Abbreviations: EBS, Efron's nonparametric bootstrap; WBS, wild bootstrap.

(independent and identically distributed) standard normal multipliers, and the empirical variance $\widehat{\text{var}}\left(Z^*(\cdot)\right)$ is based on the resampled values of the sum in the expression above. Other than that, the values of $\widehat{\text{var}}\left(\widehat{A}^*(\cdot)\right)$ and $\widehat{\text{var}}\left(Z^*(\cdot)\right)$ were also used as variance estimates in the respective formulas for the confidence intervals.

In order to limit the Monte Carlo error for the coverage below 1.6%, we repeated the simulations 1000 times, and assessed the coverage of the generated confidence intervals. The true cumulative hazard was approximated numerically: We considered the average of the Nelson–Aalen estimates obtained from 10,000 sampled studies that followed the design described above, but without censoring.

The outcomes are shown in Table 3. As can be seen, the wild bootstrap attains coverage levels that are significantly closer to the intended level of 95% in comparison to the classical bootstrap approach. Similar findings were obtained for other choices of $n \leq 200$ and $m = n/2$ (see Web Appendix B). Hence, we recommend that martingale-based survival methods, such as the wild bootstrap, should be preferred in settings where censoring is independent, but not random.

6 | ANALYSIS OF THE OAK TRIAL

Before cancer immunotherapy was approved, docetaxel had been the standard of care in patients with advanced-stage or metastatic, previously treated nonsmall cell lung cancer. Treatment with docetaxel is however associated with severe toxic effects that limit its beneficial effects (Hanna et al., 2004). The OAK study was a 1:1 randomized, open-label phase III study that compared the efficacy and safety of docetaxel to that of atezolizumab, an immunotherapy agent targeting the programmed death ligand 1 (Rittmeyer et al., 2017). The primary endpoint of the study was overall survival, and in the original analysis, atezolizumab was found to be beneficial in comparison to docetaxel.

We consider the primary efficacy population, which included the first 850 patients recruited at 194 oncology centers across 31 countries between March and November 2014. According to the statistical analysis plan, the data should be evaluated when approximately 595 deaths had been observed (CDER, 2016). The corresponding information is publicly available as a supplementary table with Gandara et al. (2018), and we used these data in order to examine analysis methods based on the assumption of random censoring in practical situations that involve type II censored data with staggered entry times.

Similarly as in Section 5, the classical nonparametric bootstrap and the wild bootstrap were applied to compute pointwise confidence intervals for the cumulative hazard function as well as the survival probability function. The sample size of 850 subjects was too large for the dependencies to have any notable effect, though, such that both resampling approaches produced basically equal results (see Web Appendix C for a graphical representation of the pointwise confidence intervals for the cumulative hazard and the survival probability.)

Against the background of interim analyses, study data are however often evaluated early, when only few of the originally planned events have been observed (cf. eg, DeMets, Furberg and Friedman, 2006, Section 1, Case 20 and Section 4, Case 28; Hughes et al., 2018). This motivated us to consider several small-sample subsets of the primary population. Web Appendix C provides figures showing the bootstrapped confidence intervals for random subsets that involve 75, 50, and 40 observed events, respectively. These results show that differences become more visible with smaller numbers of observed events, which may be more common in interim analyses, but further research is needed, for example, in the analysis of recurrent events.

7 | DISCUSSION

In this paper, we demonstrated that event-driven trials with staggered entry result in independent censoring in the counting process sense, but not in random censoring. This implies that survival methodology which relies on martingale properties allows for valid inference, but violations of the martingale structure potentially bias results. Analyses of data that are obtained from small samples or population subsets are particularly prone to such bias because the underlying dependencies are more pronounced. To this end, note that interim analyses are often also event-driven, and, as indicated by the simulations in Section 5, caution should therefore be exercised when using methods that are based on the assumption of random censoring. More thorough investigations of this observation are part of future research. Moreover, our proof showed that it is essential not to condition on the calendar times of study entry in the analysis. As long as the calendar times are disregarded,

15410420, 2023, 3, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/biom.13710 by Universitaetsbibl Augsburg, Wiley Online Library on [10/01/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

approaches that rely on the counting process framework, including for example accelerated failure time models, remain valid. The potential bias that might arise otherwise has been demonstrated by the (somewhat artificial) simulation design in Section 4. However, the relevance of our findings became apparent recently in the context of trials conducted during the COVID-19 pandemic. A common strategy here is to divide the trial period into pre-, during, and postpandemic phases that are based on calendar times (EMA, 2020; Meyer et al., 2020), and as a consequence, resulting analyses may be biased in time-to-event trials with event-driven censoring.

Finally, it is worth noting that while censoring is inherent to any time-to-event analysis, there is still a lot of discussion and confusion about it in the literature. Many seemingly different assumptions have been proposed and inconsistent definitions have further added to the confusion. In this manuscript, we followed the definition of Andersen et al. (1993), see also Andersen (2005) for an overview. In the context of type I censoring, Overgaard and Hansen (2021) recently investigated and classified different assumptions imposed on a right-censoring mechanism.

ACKNOWLEDGEMENTS

Support by the DFG (Grant FR 4121/2-1) is gratefully acknowledged. The authors also thank the editor, the associate editor, and two anonymous reviewers for their valuable comments and suggestions.

DATA AVAILABILITY STATEMENT

The data from the phase III, open-label, multicenter, randomized study investigating the efficacy and safety of atezolizumab compared with docetaxel in patients with previously treated nonsmall cell lung cancer, which were used in this paper to support our findings, are openly available at https://doi.org/10.1038/s41591-018-0134-3 (Supplementary Table 8).

ORCID

Jasmin Rühl https://orcid.org/0000-0002-1721-0640
Sarah Friedrich https://orcid.org/0000-0003-0291-4378

REFERENCES

Aalen, O.O. (1978) Nonparametric inference for a family of counting processes. *Annals of Statistics*, 6, 701–726.

Aalen, O.O., Borgan, Ø. and Gjessing, H.K. (2008) *Survival and Event History Analysis - A Process Point of View*. New York: Springer.

Andersen, P.K. (2005) Censored data. In: *Encyclopedia of Biostatistics*. New York: Wiley.

Andersen, P.K., Borgan, Ø., Gill, R.D. and Keiding, N. (1993) *Statistical Models Based on Counting Processes*. New York: Springer.

Baden, L.R., El Sahly, H.M., Essink, B., Kotloff, K., Frey, S., Novak, R. et al. (2021) Efficacy and safety of the mRNA-1273 SARS-CoV-2 vaccine. *New England Journal of Medicine*, 384, 403–416.

Beyersmann, J., Di Termini, S. and Pauly, M. (2013) Weak convergence of the wild bootstrap for the Aalen-Johansen estimator of the cumulative incidence function of a competing risk. *Scandinavian Journal of Statistics*, 40, 387–402.

Center for Drug Evaluation and Research. (2016) Statistical Review: BLA 761,041. Available at: https://www.accessdata.fda.gov/drugsatfda_docs/nda/2016/761041Orig1s000StatR.pdf [Accessed 20 June 2022).

DeMets, D.L., Furberg, C.D. and Friedman, L.M. (2006) *Data Monitoring in Clinical Trials: A Case Studies Approach*. New York: Springer.

Efron, B. (1981) Censored data and the bootstrap. *Journal of the American Statistical Association*, 76, 312–319.

Elisei, R., Schlumberger, M.J., Müller, S.P., Schöffski, P., Brose, M.S., Shah, M.H. et al. (2013) Cabozantinib in progressive medullary thyroid cancer. *Journal of Clinical Oncology*, 31, 3639–3646.

EMA. (2020) Points to Consider on Implications of Coronavirus Disease (COVID-19) on Methodological Aspects of Ongoing Clinical Trials. Avialable at: https://www.ema.europa.eu/en/documents/scientific-guideline/points-consider-implications-coronavirus-disease-covid-19-methodological-aspects-ongoing-clinical_en.pdf (Accessed June 20 2022).

Gandara, D.R., Paul, S.M., Kowanetz, M., Schleifman, E., Zou, W., Li, Y. et al. (2018) Blood-based tumor mutational burden as a predictor of clinical benefit in non-small-cell lung cancer patients treated with atezolizumab. *Nature Medicine*, 24, 1441–1448.

Hanna, N., Shepherd, F.A., Fossella, F.V., Pereira, J.R., De Marinis, F., von Pawel, J. et al. (2004) Randomized phase III trial of pemetrexed versus docetaxel in patients with non-small-cell lung cancer previously treated with chemotherapy. *Journal of Clinical Oncology*, 22, 1589–1597.

Hughes, R., Dalakas, M.C., Merkies, I., Latov, N., Léger, J.M., Nobile-Orazio, E. et al. (2018) Oral fingolimod for chronic inflammatory demyelinating polyradiculoneuropathy (FORCIDP Trial): a double-blind, multicentre, randomised controlled trial. *Lancet Neurology*, 17, 689–698.

Husain, M., Birkenfeld, A.L., Donsmark, M., Dungan, K., Eliaschewitz, F.G., Franco, D.R. et al. (2019) Oral semaglutide and cardiovascular outcomes in patients with type 2 diabetes. *New England Journal of Medicine*, 381, 841–851.

Kleinbaum, D.G. and Klein, M. (2012) *Survival Analysis - A Self-Learning Text*. New York: Springer.

Lin, D.Y., Wei, L.J. and Ying, Z. (1993) Checking the Cox model with cumulative sums of martingale-based residuals. *Biometrika*, 80, 557–572.

Martinussen, T. and Scheike, T.H. (2006) *Dynamic Regression Models for Survival Data*. New York: Springer.

McLaughlin, V., Channick, R.N., Ghofrani, H.A., Lemarié, J.C., Naeije, R., Packer, M. et al. (2015) Bosentan added to sildenafil therapy in patients with pulmonary arterial hypertension. *European Respiratory Journal*, 46, 405–413.

Meyer, R.D., Ratitch, B., Wolbers, M., Marchenko, O., Quan, H., Li, D. et al. (2020) Statistical issues and recommendations for clinical trials conducted during the COVID-19 pandemic. *Statistics in Biopharmaceutical Research*, 12, 399–411.

15410420, 2023, 3, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/biom.13710 by Universitaetsbibl Augsburg, Wiley Online Library on [10/01/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Nießl, A., Allignol, A., Beyersmann, J. and Mueller, C. (2021) 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*. DOI: 10.1016/j.ecosta.2021.09.008.

O'Quigley, J. (2008) *Proportional Hazards Regression*. New York: Springer.

Overgaard, M. and Hansen, S.N. (2021) On the assumption of independent right censoring. *Scandinavian Journal of Statistics*, 48, 1234–1255.

Rittmeyer, A., Barlesi, F., Waterkamp, D., Park, K., Ciardiello, F., von Pawel, J. et al. (2017) Atezolizumab versus docetaxel in patients with previously treated non-small-cell lung cancer (OAK): a phase 3, open-label, multicentre randomised controlled trial. *Lancet*, 389, 255–265.

Sitbon, O., Channick, R.N., Chin, K.M., Frey, A., Gaine, S., Galié, N. et al. (2015) Selexipag for the treatment of pulmonary arterial hypertension. *New England Journal of Medicine*, 373, 2522–2533.

SUPPORTING INFORMATION

Web Appendices A, B, and C, referenced in Sections 4, 5, and 6, as well as the underlying computer code are available with this paper at the Biometrics website on Wiley Online Library

WEB TABLE A.1: Overview of the simulation scenarios

WEB TABLEs A.2–A.7 and WEB FIGUREs A.1–A.8: Results summarizing the bias of the Breslow estimators and the estimated log hazard ratios (event-driven censoring with hazard ratio 1, excluding iterations with extreme hazard ratios)

WEB TABLEs A.8–A.12: Results summarizing the bias of the Breslow estimators and the estimated log hazard ratios (event-driven censoring with hazard ratio 1, complete results)

WEB TABLEs A.13–A.18 and WEB FIGUREs A.9–A.18: Results summarizing the bias of the Breslow estimators and the estimated log hazard ratios (event-driven censoring with hazard ratio 0.8, excluding iterations with extreme hazard ratios)

WEB TABLEs A.19–A.23: Results summarizing the bias of the Breslow estimators and the estimated log hazard ratios (event-driven censoring with hazard ratio 0.8, complete results)

WEB TABLEs A.24–A.29 and WEB FIGUREs A.19–A.28: Results summarizing the bias of the Breslow estimators and the estimated log hazard ratios (event-driven censoring with hazard ratio 1.25, excluding iterations with extreme hazard ratios)

WEB TABLEs A.30–A.34: Results summarizing the bias of the Breslow estimators and the estimated log hazard ratios (event-driven censoring with hazard ratio 1.25, complete results)

WEB TABLEs A.35–A.39 and WEB FIGUREs A.29–A.31: Results summarizing the bias of the Breslow estimators and the estimated log hazard ratios (random censoring, excluding iterations with extreme hazard ratios)

WEB TABLEs A.40–A.44: Results summarizing the bias of the Breslow estimators and the estimated log hazard ratios (random censoring, complete results)

Web Table B.1: Coverage probabilities and widths of the bootstrapped confidence intervals

WEB FIGURE C.1: Pointwise 95% confidence intervals for the cumulative hazard

WEB FIGURE C.2: Pointwise 95% confidence intervals for the survival probability

WEB FIGURE C.3: Pointwise 95% confidence intervals for the cumulative hazard in a random subset with 75 observed events

WEB FIGURE C.4: Pointwise 95% confidence intervals for the cumulative hazard in a random subset with 50 observed events

WEB FIGURE C.5: Pointwise 95% confidence intervals for the cumulative hazard in a random subset with 40 observed events

Data S1

Data S2

Data S3

Data S4

How to cite this article: Rühl, J., Beyersmann, J., & Friedrich, S. (2023) General independent censoring in event-driven trials with staggered entry. *Biometrics*, 79, 1737–1748. https://doi.org/10.1111/biom.13710

15410420, 2023, 3, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/biom.13710 by Universitaetsbibl Augsburg, Wiley Online Library on [10/01/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License