Received: 9 June 2020 | Revised: 15 March 2021 | Accepted: 29 April 2021

DOI: 10.1002/pst.2130

MAIN PAPER

WILEY

Estimating and comparing adverse event probabilities in the presence of varying follow-up times and competing events

Regina Stegherr[1] | Claudia Schmoor[2] | Michael Lübbert[3] | Tim Friede[4] | Jan Beyersmann[1]

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

[2]Clinical Trials Unit, Faculty of Medicine and Medical Center—University of Freiburg, Freiburg, Germany

[3]Hematology, Oncology, and Stem-Cell Transplantation, Faculty of Medicine and Medical Center—University of Freiburg, Freiburg, Germany

[4]Institut für Medizinische Statistik, Universitätsmedizin Göttingen, Göttingen, Germany

Correspondence
Regina Stegherr, Institute of Statistics, Ulm University, Helmholtzstrasse 20, 89081 Ulm, Baden-Württemberg, Germany.
Email: regina.stegherr@alumni.uni-ulm.de

Funding information
Bundesministerium für Bildung und Forschung, Grant/Award Number: 01KG0913; Deutsche Forschungsgemeinschaft, Grant/Award Number: BE 4500/3-1; Universitat Ulm

Abstract

Safety analyses of adverse events (AEs) are important in assessing benefit–risk of therapies but are often rather simplistic compared to efficacy analyses. AE probabilities are typically estimated by incidence proportions, sometimes incidence densities or Kaplan–Meier estimation are proposed. These analyses either do not account for censoring, rely on a too restrictive parametric model, or ignore competing events. With the non-parametric Aalen-Johansen estimator as the “gold standard”, that is, reference estimator, potential sources of bias are investigated in an example from oncology and in simulations, for both one-sample and two-sample scenarios. The Aalen-Johansen estimator serves as a reference, because it is the proper non-parametric generalization of the Kaplan–Meier estimator to multiple outcomes. Because of potential large variances at the end of follow-up, comparisons also consider further quantiles of the observed times. To date, consequences for safety comparisons have hardly been investigated, the impact of using different estimators for group comparisons being unclear. For example, the ratio of two both underestimating or overestimating estimators may not be comparable to the ratio of the reference, and our investigation also considers the ratio of AE probabilities. We find that ignoring competing events is more of a problem than falsely assuming constant hazards by the use of the incidence density and that the choice of the AE probability estimator is crucial for group comparisons.

KEYWORDS

Aalen-Johansen, acute myeloid leukemia, adverse events, competing events, safety

1 | INTRODUCTION

In clinical trials, safety analyses in terms of adverse events (AEs) are a key aspect of benefit–risk assessments of therapies. Inappropriate analysis methods may result in misleading conclusions about a therapy's safety. This may lead to severe consequences for patients within the trial when the analyses were conducted for safety monitoring or patients outside the trial when the therapy is used more widely following the trial.[1] In practice, the probability of an AE of a specific type is most often estimated by the incidence proportion given by the number of patients experiencing the AE out of all patients in the respective treatment group.[2] As the incidence proportion does not account for the time a patient is under observation, it ignores censoring and may lead to an underestimation of the adverse event probability.[1,3] To consider time under observation, survival methods in form of the incidence density which divides by patient-time-at-risk are suggested.[4] In contrast to the incidence proportion, the incidence density, also called incidence rate, estimates a hazard and not a probability. The incidence density can be transformed to the probability scale as we will see later in Section 2.2. This could be interpreted as a parametric version of one minus the non-parametric Kaplan–Meier estimator, but imposing the assumption of constant hazards over time, which is a typical point of criticism.[5]

Additionally to censoring, competing events (CEs) such as death or premature treatment-related discontinuation of participation in the study can occur preventing the observation of the AE.[3] The Kaplan–Meier estimator and the probability transform of the incidence density treat the CEs as censored observations and therefore do not adequately account for CEs. As a result, they overestimate the AE probability.[6,7]

An estimator for the AE probability accounting for all three potential sources of bias, namely censoring, non-constant hazards, and CEs, is the non-parametric Aalen-Johansen estimator.[8] Therefore, it is considered the "gold standard" or reference. In other words, the Aalen-Johansen estimator is the proper generalization of the Kaplan–Meier estimator to multiple outcomes. The Aalen-Johansen estimator coincides with the Kaplan–Meier estimator in the absence of CEs, it coincides with the incidence proportion (as a function of time or as an empirical subdistribution function) in the absence of censoring and it asymptotically coincides with an estimator of the AE probability based on incidence densities, provided that the latter accounts for competing risks. In the presence of censoring, non-constant hazards, and CEs, the Aalen-Johansen estimator is a consistent estimator of the AE probability.

Although the need to take CEs into account by the use of the Aalen-Johansen estimator is acknowledged in the literature on the theory of event history analysis, in areas of applied research including the analyses of AEs, CEs are still often neglected. To illustrate, recent literature searches have found that published Kaplan–Meier estimates have frequently been calculated in the presence of competing events and consequently overestimate cumulative event probabilities.[9,10] For AE analyses, not only Kaplan–Meier estimates, but also incidence proportions and incidence densities are commonly used.[3,4] Under the assumption of constant hazards for both the AE and for the CE, a parametric version of the Aalen-Johansen estimator of the AE probability can be constructed from these two hazards, which are estimated using incidence densities. This estimator is called the probability transform of the incidence density accounting for CEs.

The problems related to biased estimation of the AE probability have been previously described.[1,3,4,11,12] Here we extend the discussion by investigating the following three questions: (i) What is the impact of choosing different estimators of the AE probabilities on group comparisons in terms of bias and precision? (ii) Is the impact of ignoring CEs possibly worse than falsely assuming constant hazards? (iii) How does the time point of analysis influence the previous two questions, especially for the incidence proportion? These questions are still open, even if one agrees on the Aalen-Johansen estimator as a reference for the reasons given above.

Regarding question (i), one aspect of the benefit–risk assessment of safety analyses is treatment comparisons in terms of the relative risk. The relative risk compares two treatments by taking the ratio of the estimators of the experimental treatment and the control treatment. Even if the used probability estimator may underestimate or overestimate the AE probability, the ratio of two probability estimates, obtained with one of the biased estimators, might be comparable to the ratio of the probability estimates obtained with the reference, and an analogous question arises if one considers risk differences rather than ratios. In this paper additionally to the AE probability estimators also the variances of the AE probability estimators are investigated as misspecification of the hazard in the form of falsely assuming constant hazards may also influence the variances of the parametric estimators. A non-parametric bootstrap is suggested as a suitable alternative to obtain the variance estimates under model misspecification.[13] These bootstrapped variances are compared to the asymptotic, model-based estimators to see whether the assumption of constant hazards also impacts the variances.

Question (ii) aims to investigate whether a misspecified incidence density analysis, that is, an analysis assuming constant hazards where the hazards might be time-varying, but that does account for CEs may be useful provided that variance estimation accounts for misspecification.

Question (iii) is a consequence of the problem that the incidence proportion is usually only calculated at the end of follow-up in each group of the two treatment groups. This leads to different evaluation time points in the relative risk of the incidence proportion drawing the interpretation into question.[14] To solve this issue all comparisons are not only conducted at the end of follow-up but also at the shorter of the two observed maximum follow-up times in the two groups, which is in line with, for example, common logrank test comparisons. There is also the concern that a low number of observations under study at the end of follow-up leads to increased variances.[15] Hence, the comparisons of the AE probability estimators are also investigated at two different quantiles of the observed times.

Throughout, our estimands will use the intention-to-treat population when comparing groups, based on estimators of the cumulative AE probability. In practice, treatment effects will be investigated using either treatment policy or while on treatment estimands, depending on the type of CEs and the available data, see Section 2.1.

We also note that Bender and Beckmann[16] have recently investigated whether the ratio of incidence densities may serve as an estimator of the hazard ratio even under misspecification. Their investigation was also motivated by AE analyses, also considered variances (via confidence intervals), and found that results depend on the baseline cumulative AE probability. However, these authors did neither consider competing events (which impacts probabilities) nor bootstrapping variances (which may lead to larger confidence intervals).

The paper is organized as follows, Section 2 introduces the AE probability estimand and estimators with corresponding variance estimators. Section 3 presents the comparisons of the estimators at the different follow-up times based on data from an oncology trial. In Section 4 a simulation study addresses the three questions posed above. The paper concludes with a discussion in Section 5.

2 | ESTIMATORS AND THEIR VARIANCES OF EVENT PROBABILITIES AND THEIR RATIOS

2.1 | Competing risks model and estimand

In the following, we consider data from a two-arm randomized controlled trial with each group following a competing risks model displayed in Figure 1. Every patient starts in the initial state 0 at study entry, that is, at time 0. The event time at which a patient i moves from state 0 to either state 1 or 2, whatever occurs first, is denoted by T_i and the event type is denoted by ϵ_i as $\epsilon_i = 1$ in case of an AE and $\epsilon_i = 2$ in case of a CE in a time-to-first-event and type-of-first-event setting. Observation of patient i's data (T_i, ϵ_i) is subject to right-censoring at time C_i if $C_i < T_i$. Only the minimum of the censoring time C_i or the event time T_i can be observed and (T_i, ϵ_i) remain unobserved if $C_i < T_i$. Therefore, the observable data consists of *i.i.d.* replicates of $(\min(T_i, C_i), 1(T_i \leq C_i) \cdot \epsilon_i)$. Furthermore, (T_i, ϵ_i) and C_i are assumed to be independent.

Suppressing index i in the notation, the AE hazard is defined as $\lambda(t) = \lim_{\Delta t \searrow 0} P(T \in [t, t+\Delta t), \epsilon = 1 \mid T \geq t)/\Delta t$ and the hazard of the CE as $\bar{\lambda}(t) = \lim_{\Delta t \searrow 0} P(T \in [t, t+\Delta t), \epsilon = 2 \mid T \geq t)/\Delta t$, respectively.

Within groups, our main estimand will be the cumulative probability of a type 1 event,

$$P(T \leq \tau, \epsilon = 1) = \int_0^\tau P(T \geq u) \lambda(u) \, \mathrm{d}u,$$

for times τ as defined below. Estimands of group comparison will contrast this estimand between groups. These estimands follow the intention-to-treat principle in the sense that they are based on the intention-to-treat population. If death is the only CE that may possibly occur before an AE, the treatment-policy estimand is adressed.[17] However, in safety analyses as part of the marketing authorization process, regulatory agencies typically consider the effect of the

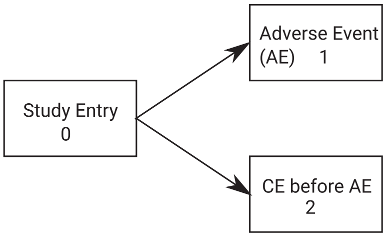

FIGURE 1 Competing risks setting

intention-to-treat on the while on treatment estimand, thus not completely following the intention-to-treat principle. One reason for this choice is that switch or discontinuation of treatment may result in diluting and perhaps anti-conservative effects when assessing safety. This is reflected by the fact that AE follow-up is often shorter than follow-up for primary endpoints in oncology such as death. For instance, in oncology, another CE could be progression, in which case we would consider the intention-to-treat during progression-free survival, corresponding to the while-on-treatment estimand when progression leads to end of treatment.[18] These considerations also have a technical aspect, at least for non-parametric survival analyses. Here, a typical technical requirement is that the asymptotic probability of being at-risk is bounded away from zero.[1]

All estimators are evaluated at time τ. Later τ will take the values $\tau_{\max}^{(A)} = \max\{\min(T_i, C_i) | i \,\text{in group A}\}$ and $\tau_{\max}^{(B)} = \max\{\min(T_i, C_i) | i \,\text{in group B}\}$ which are the maximum follow-up times in the experimental group and control group, respectively. In general, $\tau_{\max}^{(A)} \neq \tau_{\max}^{(B)}$, but comparing both groups by evaluating estimators at the respective maximum follow-up time is what is commonly done in the analyses of adverse events. But to refrain from comparisons being impacted by observing one group much longer than the other, we also consider $\tau_{\max} = \min\left(\tau_{\max}^{(A)}, \tau_{\max}^{(B)}\right)$, $\tau_{P90} = \min\left(\tau_{P90}^{(A)}, \tau_{P90}^{(B)}\right)$, with $\tau_{P90}^{(A)}$ and $\tau_{P90}^{(B)}$ the empirical 90% quantiles of $\min(T_i, C_i)$ in group A and B, respectively, and τ_{P60} defined in the same way.

2.2 | Estimators of event probabilities and their variances

The following five different estimators of the AE probability on (0, τ] are considered and will be compared. The estimators are only displayed for the experimental group A. For the control group B, they are derived analogously.

- *Incidence proportion*: Let $d_A(u)$ denote the number of observed AEs at time u in group A and let n_A be the total number of patients in group A, then the incidence proportion is defined as

$$\widehat{\text{IP}}_A(\tau) = \frac{\sum\limits_{u \in (0,\tau]} d_A(u)}{n_A},$$

where the sum is over all observed, unique event times u. The incidence proportion estimates P (AE in (0, τ], AE observed). The corresponding model-based variance estimator is

$$\widehat{\text{var}}\left(\widehat{\text{IP}}_A(\tau)\right) = \frac{\widehat{\text{IP}}_A(\tau)\left(1 - \widehat{\text{IP}}_A(\tau)\right)}{n_A}.$$

- *Probability transform incidence density*: We first define the incidence density as

$$\widehat{\text{ID}}_A(\tau) = \frac{\sum\limits_{u \in (0,\tau]} d_A(u)}{\sum\limits_{i=1}^{n_A} \min(t_i, \tau)},$$

where the denominator is the population time at risk restricted by τ with $T_i = t_i$ the (realization of the) time of the first event irrespective of the event type of patient i. The model-based variance estimator of the incidence density is

$$\widehat{\mathrm{var}}\left(\widehat{\mathrm{ID}}_{\mathrm{A}}(\tau)\right)=\frac{\sum\limits_{u\in(0,\tau]}d_{\mathrm{A}}(u)}{\left(\sum\limits_{i=1}^{n_{\mathrm{A}}}\min(t_i,\tau)\right)^2}.$$

The incidence density is an estimator of the hazard λ_{A} and as a consequence not directly comparable to the incidence proportion. But by the assumption of constant hazards and the connection to the exponential distribution it can be transformed to the probability scale by $1-\exp\left(-\widehat{\mathrm{ID}}_{\mathrm{A}}(\tau)\cdot\tau\right)$. If the assumption of constant hazards holds, $\lambda_{\mathrm{A}}(t)=\lambda_{\mathrm{A}}\forall t$, the incidence density is an unbiased estimator of the hazard. The model-based variance estimator of the probability transform is $\widehat{s}_{\mathrm{A}}^2=\tau^2\cdot\exp\left(-\tau\cdot\widehat{\mathrm{ID}}_{\mathrm{A}}(\tau)\right)^2\cdot\widehat{\mathrm{var}}\left(\widehat{\mathrm{ID}}_{\mathrm{A}}(\tau)\right)$, which can easily be derived using the delta-method.

- *1-Kaplan–Meier*: Let $\Delta\widehat{\Lambda}_{\mathrm{A}}(u)$ be the increment of the Nelson-Aalen estimator of the cumulative AE hazard, that is, $\Delta\widehat{\Lambda}_{\mathrm{A}}(u)=\widehat{\Lambda}_{\mathrm{A}}(u)-\widehat{\Lambda}_{\mathrm{A}}(u-)$, where $u-$ denotes the event time just before u, and the Nelson-Aalen estimator is given by $\widehat{\Lambda}_{\mathrm{A}}(u)=\sum_{t\in(0,u]}d_{\mathrm{A}}(t)/Y_{\mathrm{A}}(t)$, where $Y_{\mathrm{A}}(t)$ denotes the number of individuals at risk in group A just prior to t. Therefore, the increment of the Nelson-Aalen estimator is closely related to $\widehat{\mathrm{ID}}_{\mathrm{A}}(\tau)$ and can be interpreted as an estimator of the current hazard value times the time increment. Then the 1-Kaplan–Meier estimator is defined as

$$1-\widehat{S}_{\mathrm{A}}(\tau)=1-\prod_{u\in(0,\tau]}\left(1-\Delta\widehat{\Lambda}_{\mathrm{A}}(u)\right).$$

The Kaplan–Meier estimator estimates the "event-specific survival function" $S_{\mathrm{A}}(\tau)=\exp\left(-\int_0^\tau\lambda_{\mathrm{A}}(u)\,\mathrm{d}u\right)$ treating CEs as censoring the follow-up time. Its variance can be estimated using the Greenwood variance estimator.[19] Note that S_{A} does *not* have a proper probability interpretation as a consequence of competing events, see below.

- *Aalen-Johansen estimator*: As reference we consider here the Aalen-Johansen estimator which is given by

$$\widehat{AJ}_{\mathrm{A}}(\tau)=\sum_{u\in(0,\tau]}\left\{\prod_{v\in(0,u)}\left(1-\Delta\widehat{\Lambda}_{\mathrm{A}}(v)-\Delta\widehat{\overline{\Lambda}}_{\mathrm{A}}(v)\right)\right\}\Delta\widehat{\Lambda}_{\mathrm{A}}(u)$$

where $\Delta\widehat{\overline{\Lambda}}_{\mathrm{A}}(v)$ is the increment of the Nelson-Aalen estimator of the CE. The model-based variance of the Aalen-Johansen estimator can be estimated using a Greenwood-type estimator.[20] The Aalen-Johansen estimator estimates the cumulative incidence function, that is, the cumulative probability of a type 1 event $P(T\leq\tau,\epsilon=1|\,\mathrm{group\,A})=\int_0^\tau P(T\geq u|\,\mathrm{group\,A})\lambda_{\mathrm{A}}(u)\,\mathrm{d}u$. Here, the Kaplan–Meier estimator in the curly braces of the previous display estimates $P(T\geq u \mid \text{group A})$ and the increment of the Nelson-Aalen estimator estimates $\lambda_{\mathrm{A}}(u)\,\mathrm{d}u$. Comparing 1 minus the event-specific survival function and the cumulative incidence function it can be easily shown than 1 minus the event-specific survival function is greater than the cumulative incidence function as long as CEs are present (see Appendix A for details), and the same inequality holds for the 1-Kaplan–Meier estimator and the Aalen-Johansen estimator. It can be shown that in absence of censoring or if censoring is only observed after the last AE, the Aalen-Johansen estimator is equal to the incidence proportion and that else the incidence proportion is smaller than the Aalen-Johansen estimator (see Appendix B for details).

- *Probability transform incidence density accounting for CEs*:

$$\frac{\widehat{\mathrm{ID}}_{\mathrm{A}}(\tau)}{\widehat{\mathrm{ID}}_{\mathrm{A}}(\tau)+\widehat{\overline{\mathrm{ID}}}_{\mathrm{A}}(\tau)}\left(1-\exp\left(-\tau\left[\widehat{\mathrm{ID}}_{\mathrm{A}}(\tau)+\widehat{\overline{\mathrm{ID}}}_{\mathrm{A}}(\tau)\right]\right)\right),$$

where $\widehat{\overline{\mathrm{ID}}}_{\mathrm{A}}(\tau) = \sum\limits_{u\in(0,\tau]} \overline{d}_{\mathrm{A}}(u) / \sum\limits_{i=1}^{n_{\mathrm{A}}} \min(t_i, \tau)$ with $\overline{d}_{\mathrm{A}}(u)$ the number of observed CEs at time u in group A, is the incidence density of the CE and, hence, $\widehat{\overline{\mathrm{ID}}}_{\mathrm{A}}(\tau)\cdot\tau$ is the parametric analog of $\widehat{\overline{\Lambda}}_{\mathrm{A}}(\tau)$. Using the incidence density of the CE, the connection between the incidence density and the incidence proportion is $\widehat{\mathrm{ID}}_{\mathrm{A}}(\tau) / \left(\widehat{\mathrm{ID}}_{\mathrm{A}}(\tau) + \widehat{\overline{\mathrm{ID}}}_{\mathrm{A}}(\tau)\right) = \sum_{u\in(0,\tau]} d_{\mathrm{A}}(u)/n_{\mathrm{A}}$ in the absence of censoring.[21] In the presence of censoring and under a constant hazards assumption, the second factor of this probability transform, $\left(1 - \exp\left(-\tau\left[\widehat{\mathrm{ID}}_{\mathrm{A}}(\tau) + \widehat{\overline{\mathrm{ID}}}_{\mathrm{A}}(\tau)\right]\right)\right)$, is the estimated probability of experiencing an event of either type until time τ, whereas the first factor, $\widehat{\mathrm{ID}}_{\mathrm{A}}(\tau) / \left(\widehat{\mathrm{ID}}_{\mathrm{A}}(\tau) + \widehat{\overline{\mathrm{ID}}}_{\mathrm{A}}(\tau)\right)$, is the estimated probability of this event being an AE. Moreover, a model-based variance estimator of the probability transform of the incidence density accounting for CEs can be derived with the delta-method[22] and is provided in Appendix C. In the following, we will also call the probability transform of the incidence density accounting for CEs, somewhat loosely, parametric counterpart of the Aalen-Johansen estimator as the two estimators estimate the same quantity under the parametric assumption of constant hazards.

Another way to estimate the variances of the estimators is a non-parametric bootstrap accounting for model misspecifications that may also influence the model-based variances.[13] To be more precise, the bootstrapped variance estimates are obtained by calculating the variance of the estimators of 1000 bootstrap datasets, which are generated by sampling observations of the original dataset with replacement until a dataset of the same size is obtained. The variances of the parametric estimators given above assume that the hazards are constant. This assumption is not made by the non-parametric bootstrap.

Below, we will compare non-parametric bootstrap variance estimators with the closed formula variance estimators from provided above, denoting the latter as "model-based."

2.3 | Between group comparisons

The comparison of the two treatment groups can be done in terms of the relative risk

$$\hat{RR}(\tau) = \frac{\hat{p}_{\mathrm{A}}(\tau)}{\hat{p}_{\mathrm{B}}(\tau)},$$

where $\hat{p}_{\mathrm{A}}(\tau)$ and $\hat{p}_{\mathrm{B}}(\tau)$ are one of the five estimators of the AE probability in the interval (0, τ] in group A and group B which are introduced above. The variance estimator of the relative risk can be constructed based on a log-transformation

$$\hat{\mathrm{var}}\left(\log \hat{RR}(\tau)\right) = \frac{1}{\hat{p}_{\mathrm{A}}(\tau)^2}\cdot\hat{\mathrm{var}}_{\mathrm{A}}(\tau)^2 + \frac{1}{\hat{p}_{\mathrm{B}}(\tau)^2}\cdot\hat{\mathrm{var}}_{\mathrm{B}}(\tau)^2,$$

where $\hat{\mathrm{var}}_{\mathrm{A}}(\tau)^2$ and $\hat{\mathrm{var}}_{\mathrm{B}}(\tau)^2$ are one of the two suggested variance estimators of $\hat{p}_{\mathrm{A}}(\tau)$ and $\hat{p}_{\mathrm{B}}(\tau)$ evaluated at time τ.

3 | AN EXAMPLE: THE DECIDER TRIAL IN ACUTE MYELOID LEUKEMIA

As an example, we use data from the DECIDER trial (DECItabine, DEacetylase inhibition, Retinoic acid; ClinicalTrials.gov identifier: NCT00867672).[23,24] This randomized, multicenter trial had the objective to investigate the efficacy and safety of valproate (VPA) and all-trans retinoic acid (ATRA) in combination with decitabine in 200 older and nonfit patients with acute myeloid leukemia. The trial had a 2 × 2 design, in which patients were randomly assigned to one of four treatment arms: decitabine, decitabine + VPA, decitabine + ATRA, or decitabine + VPA + ATRA. The primary endpoint of this phase II trial was objective response, and the trial was planned to show a difference between the combined ATRA and VPA treatment arms and the combined no ATRA and no VPA arms (control) at one-sided level alpha of 0.1. Under the assumption of treatment effects of 40% versus 25% a sample size of 200 was calculated for a power of 80%. The trial showed that the objective response rate as well as the overall survival rate was increased by ATRA. Here, we consider the comparison of the combined ATRA treatment arms (called group A, $n_A = 96$) with the combined no

ATRA treatment arms (called group B, $n_B = 104$) with respect to the adverse event severe thrombocytopenia, Common Terminology Criteria for Adverse Events (CTCAE) grade 3–5. This AE was observed in 35 patients in group A and 32 patients in group B. As AEs were recorded only until 28 days after end of treatment, death and end of treatment plus 28 days with no observed AE during that time were considered as CEs. As ATRA prolonged overall survival time (median of 8.2 months in group A and 5.1 months in group B),[24] a CE was experienced by 56 patients in group A and by 69 patients in group B, leading on average to a longer follow-up time for AEs in group A (mean 137 days) as compared to group B (mean 125 days). So in this data set, there is hardly any censoring as only five patients in group A and three patients in group B were censored. The time was measured in days and analyses were performed at $\tau_{\max}^{(A)} = 802$, respectively $\tau_{\max}^{(B)} = 980$, the maximum follow-up times in the two groups, at $\tau_{\max} = 802$ the minimum of the two maximum follow-up times, at $\tau_{P90} = 353$ and at $\tau_{P60} = 67$, both chosen according to the quantiles of the observed times.

3.1 | Estimating the AE probability

Figure 2 displays the five different AE probability estimators for the four different follow-up times. In the original analysis of the trial, the AE probabilities were estimated at $\tau_{\max}^{(A)}$, $\tau_{\max}^{(B)}$ using the incidence proportion, leading to an estimated probability of severe thrombocytopenia of 36.5% in group A and 30.8% in group B.[24] This difference was not regarded as relevant taking into account the superiority of group A with respect to overall survival. Now considering the other estimators, the probability transform of the incidence density results in a higher estimated AE probability than the Aalen-Johansen estimator at all follow-up times. The difference is more pronounced for longer follow-up times as the CEs are observed later in time. The 1-Kaplan–Meier estimator also always obtains a higher estimated AE probability than the Aalen-Johansen estimator but less pronounced than the probability transform of the incidence density. For the latter, the estimated cumulative hazards in Figure 3 illustrate departures from the constant hazard assumption. Moreover, Figure 3 shows that CEs tend to occur later in time than AEs. This is the reason why the difference between the probability transform of the incidence density or the 1-Kaplan–Meier estimator and the Aalen-Johansen estimator is larger at later follow-up times.

The other three estimators differ only slightly. The probability transform of the incidence density accounting for CEs is comparable to the Aalen-Johansen estimator. Even though the assumption of constant hazards is arguably not valid, the difference between the two estimators that account for the CE is nearly negligible. The incidence proportion

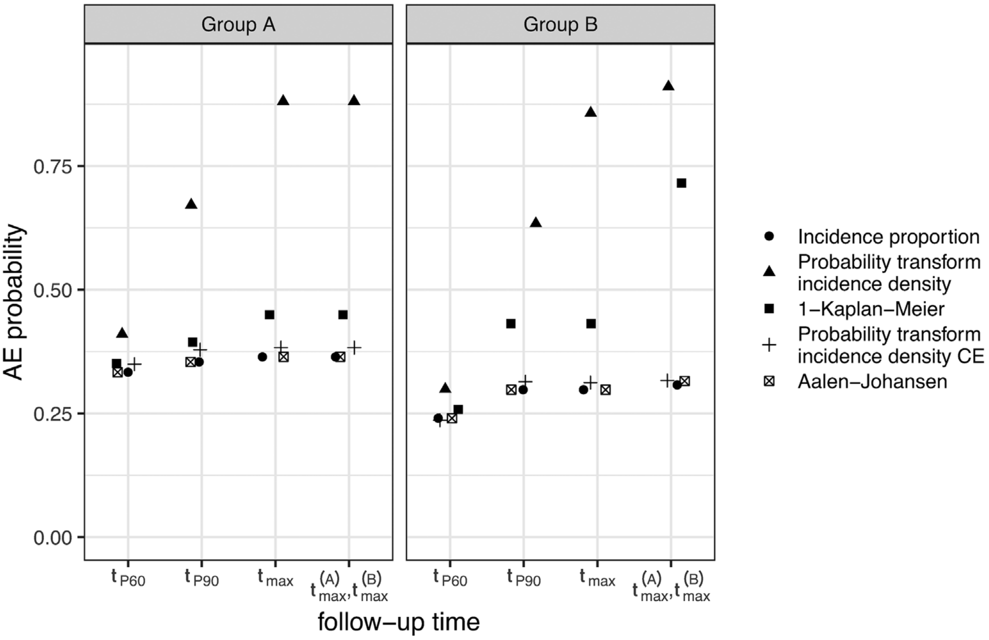

FIGURE 2 Adverse event probability estimators applied to the DECIDER trial at the different follow-up times. The symbols have been jittered for better readability

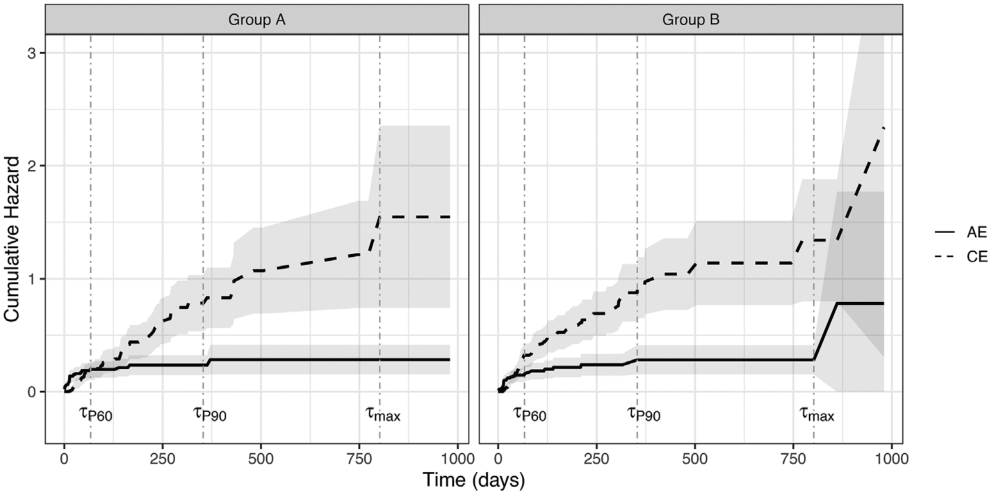

FIGURE 3 Nelson-Aalen estimates of the cumulative hazard with 95% pointwise confidence intervals (95% CI) of the adverse event and the competing event in both treatment groups. The vertical lines display the analysis times

takes the same value as the Aalen-Johansen estimator as long as there are no censored observations in the data.[21] In these data, there is hardly any censoring. In group A, all censoring is observed after the last AE yielding equality of the Aalen-Johansen estimator and the incidence proportion. In group B, however, there is one AE after all censoring resulting in a slightly lower incidence proportion compared to the Aalen-Johansen estimator. When estimating the probabilities at earlier follow-up times this difference is less pronounced as in these situations all AEs are observed before one observation is censored.

3.2 | Variance estimation

The variances of the probability estimators are calculated in two ways, first by using the formulas displayed in Section 2 and second by a non-parametric bootstrap.

Table 1 displays the estimated variances of all estimators for both groups evaluated at the different follow-up times. At the maximal follow-up time in each group, two major points can be observed. Firstly, the variance of the 1-Kaplan–Meier estimator is much larger than all other variances. This verifies the expectation that at later follow-up times the variance of the Kaplan–Meier estimator is increased due to a small number of patients still at risk and this is the motivation for comparing also at the other follow-up times.[15] Also for shorter follow-up, except for τ_{P60}, this increased variance can be observed. At τ_{P60} all variances are virtually the same.

Secondly, for the probability transform of the incidence density, the model-based variance is smaller than the bootstrapped variance. This is likely due to the model misspecification as the constant hazard assumption might not hold for the AE hazard as Figure 3 suggests. Again this effect is more pronounced for longer follow-up times than for τ_{P60}.

The two variance estimators of the Aalen-Johansen estimator and of the probability transform of the incidence density accounting for CEs are almost identical as, like the incidence proportion, they aim to estimate the binomial probability $P(\text{AE in } (0, t])$.

3.3 | Group comparisons

We investigate how the ratio of two biased estimators of the probabilities compares to the ratio of two unbiased estimators. Table 2 displays the relative risks calculated from the estimators and time points shown in Figure 2. Additionally, the 95% confidence intervals of the relative risk calculated with both variance estimators are displayed.

TABLE 1 Estimated variances of the AE probabilities using the analytically derived model-based variances and bootstrapped variances in the DECIDER trial[24]

		Group A		Group B	
Follow-up time	Estimator	Model-based	Bootstrap	Model-based	Bootstrap
$\tau_{\max}^{(A)}, \tau_{\max}^{(B)}$	Incidence proportion	0.0024	0.0025	0.0020	0.0019
$\tau_{\max}^{(A)}, \tau_{\max}^{(B)}$	Probability transform incidence density	0.0018	0.0036	0.0014	0.0034
$\tau_{\max}^{(A)}, \tau_{\max}^{(B)}$	1-Kaplan–Meier	0.0054	0.0060	0.0419	0.0509
$\tau_{\max}^{(A)}, \tau_{\max}^{(B)}$	Aalen-Johansen	0.0024	0.0025	0.0022	0.0020
$\tau_{\max}^{(A)}, \tau_{\max}^{(B)}$	Probability transform incidence density CE	0.0026	0.0026	0.0021	0.0020
$\tau_{\max}$	Incidence proportion	0.0024	0.0025	0.0020	0.0019
$\tau_{\max}$	Probability transform incidence density	0.0018	0.0041	0.0025	0.0045
$\tau_{\max}$	1-Kaplan–Meier	0.0054	0.0060	0.0062	0.0067
$\tau_{\max}$	Aalen-Johansen	0.0024	0.0025	0.0020	0.0019
$\tau_{\max}$	Probability transform incidence density CE	0.0026	0.0027	0.0022	0.0021
τ_{P90}	Incidence proportion	0.0024	0.0024	0.0020	0.0019
τ_{P90}	Probability transform incidence density	0.0039	0.0059	0.0044	0.0049
τ_{P90}	1-Kaplan–Meier	0.0031	0.0036	0.0062	0.0047
τ_{P90}	Aalen-Johansen	0.0024	0.0024	0.0020	0.0019
τ_{P90}	Probability transform incidence density CE	0.0026	0.0028	0.0022	0.0021
τ_{P60}	Incidence proportion	0.0023	0.0022	0.0018	0.0016
τ_{P60}	Probability transform incidence density	0.0030	0.0032	0.0025	0.0023
τ_{P60}	1-Kaplan–Meier	0.0026	0.0024	0.0021	0.0018
τ_{P60}	Aalen-Johansen	0.0023	0.0022	0.0018	0.0016
τ_{P60}	Probability transform incidence density CE	0.0024	0.0027	0.0017	0.0017

Note: The following follow-up times are considered $\tau_{\max}^{(A)}$, $\tau_{\max}^{(B)}$ the maximum follow-up times in group A and group B, respectively, $\tau_{\max}$ the minimum of the two maximal event times, τ_{P90} and τ_{P60} chosen according to the quantiles of the time and account for different lengths of follow-up between group A and group B.

The relative risks calculated by the incidence proportion and the Aalen-Johansen estimator only differ at the maximal follow-up time. This is a direct consequence of the AE after the late censorings in group B leading to a slightly lower AE probability estimated by the incidence proportion than by the Aalen-Johansen estimator.

At all follow-up times, the relative risk estimated with the probability transform of the incidence density and the one estimated with the 1-Kaplan–Meier estimator are smaller than the relative risk obtained with the reference Aalen-Johansen estimator. At the maximum follow-up time, the direction of the estimated relative risks is different. But note that 1 is always included in the confidence interval indicating no statistically significant therapy effects.

The relative risk estimated by the probability transform of the incidence density accounting for CEs and the relative risk estimated by the Aalen-Johansen are similar at the later follow-up times. At τ_{P60} the relative risk estimated by the probability transform of the incidence density accounting for CEs is greater than the one estimated by the Aalen-Johansen estimator.

The differences between the confidence intervals obtained with the model-based and with the bootstrapped variances are due to differences in the estimated variances. The possible impact of this is illustrated by the fact that the relative risk estimated with the Aalen-Johansen estimator is not included in the model-based confidence interval of the probability transform of the incidence density at the maximal follow-up time and at τ_{P90} but it is included in the confidence interval obtained with the bootstrapped variance.

Furthermore, the model-based confidence intervals of the incidence proportion and the Aalen-Johansen estimator are wider than the confidence interval where the variances are obtained with a bootstrap. For the probability transform of the incidence density and the 1-Kaplan–Meier estimator at $\tau_{\max}^{(A)}$, $\tau_{\max}^{(B)}$ and $\tau_{\max}$, for the probability transform of the

TABLE 2 Relative risk (RR), model-based confidence interval (CI), bootstrap CI and the ratio of the lengths of the two confidence intervals (model-based/bootstrap) calculated from the different AE probability estimators at several follow-up times in the DECIDER trial[24]

			Model-based CI		Bootstrap CI		
Follow-up time	Estimator	RR	Lower	Upper	Lower	Upper	Ratio of the lengths of the CIs
$\tau_{\max}^{(A)}, \tau_{\max}^{(B)}$	Incidence proportion	1.1849	0.8015	1.7517	0.8055	1.7429	1.0138
$\tau_{\max}^{(A)}, \tau_{\max}^{(B)}$	Probability transform incidence density	0.9678	0.8540	1.0968	0.8053	1.1631	0.6786
$\tau_{\max}^{(A)}, \tau_{\max}^{(B)}$	1-Kaplan–Meier	0.6280	0.3294	1.1971	0.3105	1.2703	0.9041
$\tau_{\max}^{(A)}, \tau_{\max}^{(B)}$	Aalen-Johansen	1.1550	0.7805	1.7092	0.7843	1.7008	1.0134
$\tau_{\max}^{(A)}, \tau_{\max}^{(B)}$	Probability transform incidence density CE	1.2097	0.8217	1.7810	0.8259	1.7719	1.0140
$\tau_{\max}$	Incidence proportion	1.2231	0.8233	1.8172	0.8259	1.8114	1.0085
$\tau_{\max}$	Probability transform incidence density	1.0277	0.8857	1.1924	0.8337	1.2668	0.7082
$\tau_{\max}$	1-Kaplan–Meier	1.0415	0.6449	1.6820	0.6296	1.7230	0.9485
$\tau_{\max}$	Aalen-Johansen	1.2231	0.8233	1.8172	0.8259	1.8114	1.0085
$\tau_{\max}$	Probability transform incidence density CE	1.2259	0.8294	1.8120	0.8317	1.8069	1.0075
τ_{P90}	Incidence proportion	1.1882	0.7965	1.7724	0.8036	1.7567	1.0240
τ_{P90}	Probability transform incidence density	1.0594	0.8052	1.3938	0.7756	1.4469	0.8767
τ_{P90}	1-Kaplan–Meier	0.9140	0.5808	1.4384	0.5939	1.4066	1.0553
τ_{P90}	Aalen-Johansen	1.1882	0.7965	1.7724	0.8036	1.7567	1.0240
τ_{P90}	Probability transform incidence density CE	1.2062	0.8159	1.7833	0.8097	1.7970	0.9799
τ_{P60}	Incidence proportion	1.3867	0.8899	2.1608	0.9067	2.1207	1.0468
τ_{P60}	Probability transform incidence density	1.3719	0.9021	2.0863	0.9067	2.0759	1.0128
τ_{P60}	1-Kaplan–Meier	1.3588	0.8683	2.1263	0.8883	2.0785	1.0569
τ_{P60}	Aalen-Johansen	1.3867	0.8899	2.1608	0.9067	2.1207	1.0468
τ_{P60}	Probability transform incidence density CE	1.4800	0.9549	2.2937	0.9445	2.3191	0.9739

incidence density at τ_{P90} and for the probability transform of the incidence density accounting for CEs at τ_{P90} and τ_{P60} the confidence interval based on the bootstrapped variance is longer than the model-based confidence interval. This is also due to the differences in the estimated variances.

In summary, in this data example, estimators that lead to a higher estimate of the AE probability than the Aalen-Johansen estimator result in a smaller relative risk estimate than that based on the Aalen-Johansen estimator. The model-based CIs of the probability transform of the incidence density may be too small in the sense that they do not cover the reference estimate of the RR, but the bootstrapped CIs do not suffer from this shortcoming.

4 | SIMULATION STUDY

In this section, we take a closer look at the three sources of bias, namely censoring, CEs, and non-constant hazards. For this purpose, competing risk data with the event of interest and one CE are simulated for two independent groups A and B.[25,26] The simulation algorithm is described in the Supporting Information. Table 3 describes the simulation

TABLE 3 Summary of the scenarios considered in the simulation study ($N_{\text{REP}} = 10{,}000$). The sample size is equal in both groups $n_A = n_B = n$

Scenario	$\lambda_A(t)$	$\bar{\lambda}_A(t)$	$\lambda_B(t)$	$\bar{\lambda}_B(t)$	n	Censoring
S1 constant	0.00265	0.00424	0.00246	0.00530	200	No
S2 constant ⋆	0.00265	0.00424	0.00246	0.00530	400	No
S3 constant ⋆	0.00265	0.00424	0.00246	0.00530	400	28% in A; 15% in B
S4 time-dependent	$\frac{1}{3}t^2$	$\frac{8}{9}t$	$\frac{1.8}{t+0.5}$	$\frac{8}{9}t$	400	No
S5 time-dependent ⋆	$\frac{1}{3}t^2$	$\frac{8}{9}t$	$\frac{1.8}{t+0.5}$	$\frac{8}{9}t$	400	14% in A; 10% in B
S6 time-dependent	$\frac{1.8}{t+2}$	$\frac{1}{2}t$	$\frac{1.8}{t+2}$	$\frac{1}{8}t$	400	18.5% in A and B
S7 time-dependent	$\frac{1.8}{t+2}$	$\frac{1}{2}t$	$\frac{1.8}{t+2}$	$\frac{1}{8}t$	400	No
S8 time-dependent	$\frac{1}{2}t$	$\frac{1.8}{t+2}$	$\frac{1}{8}t$	$\frac{1.8}{t+2}$	400	No
S9 time-dependent	$\frac{1}{2}t$	$\frac{1.8}{t+2}$	$\frac{1}{8}t$	$\frac{1.8}{t+2}$	400	18.5% in A and B
S10 constant— time-dependent ⋆	0.07	$0.066\ t^{-0.283}$	0.06	$0.042\ t^{-0.283}$	400	1.7% in A; 2.3% in B

TABLE 4 Simulation results: Mean simulated follow-up times

	Scenario			
Follow-up time	S2	S3	S5	S10
$\tau_{\max}^{(A)}$	956.2	472.1	50.2	65.6
$\tau_{\max}^{(B)}$	845.7	462.7	48.6	83.2
$\tau_{\max}$	796.1	453.3	47.6	63.2
τ_{P90}	293.9	218.4	22.6	20.6
τ_{P60}	117.3	90.9	9.4	7.4

Note: The following follow-up times are considered: $\tau_{\max}^{(A)}$, $\tau_{\max}^{(B)}$ the maximum follow-up times in group A and group B, respectively, $\tau_{\max}$ the minimum of the two maximal event times, τ_{P90} and τ_{P60} chosen according to the quantiles of the time and account for different lengths of follow-up between group A and group B ($N_{\text{REP}} = 10{,}000$).

TABLE 5 Simulation results: Mean simulated AE probabilities at the different follow-up times of group A and B

	Group A Scenario				Group B Scenario			
Follow-up time	S2	S3	S5	S10	S2	S3	S5	S10
$\tau_{\max}^{(A)}, \tau_{\max}^{(B)}$	0.3837	0.3696	0.1468	0.6087	0.3162	0.3081	0.8201	0.6928
$\tau_{\max}$	0.3825	0.3674	0.1380	0.6086	0.3161	0.3074	0.8200	0.6875
τ_{P90}	0.3334	0.2989	0.0302	0.5324	0.2843	0.2585	0.7654	0.5565
τ_{P60}	0.2129	0.1788	0.0018	0.3275	0.1891	0.1602	0.5234	0.3119

Note: The following follow-up times are considered: $\tau_{\max}^{(A)}$, $\tau_{\max}^{(B)}$ the maximum follow-up times in group A and group B, respectively, $\tau_{\max}$ the minimum of the two maximal event times, τ_{P90} and τ_{P60} chosen according to the quantiles of the time and account for different lengths of follow-up between group A and group B ($N_{\text{REP}} = 10{,}000$).

scenarios. The simulation scenarios marked with ⋆ are displayed in the following. These represent all main findings of the investigation. The results of the other simulation scenarios can be found in the Supporting Information. For each scenario $N_{\text{REP}} = 10{,}000$ datasets are simulated.

Scenarios S1, S2, and S3 consider constant hazards for both events. The hazards are equal to the incidence densities estimated in the data example presented in Section 3. The hazards $\lambda_A(t)$ and $\lambda_B(t)$ correspond to the

TABLE 6 Simulation results: Mean absolute and mean relative bias

		Group A Scenario				Group B Scenario			
Follow-up time	**Estimator**	**S2**	**S3**	**S5**	**S10**	**S2**	**S3**	**S5**	**S10**
Mean absolute bias (mean of [estimated value − true simulated value])									
$\tau_{\max}^{(A)}, \tau_{\max}^{(B)}$	IP	0.0011	−0.0931	−0.0453	−0.0095	0.0007	−0.0711	0.0274	−0.0175
$\tau_{\max}^{(A)}, \tau_{\max}^{(B)}$	PTID	0.5279	0.3425	0.1130	0.3771	0.5490	0.3697	0.1957	0.2973
$\tau_{\max}^{(A)}, \tau_{\max}^{(B)}$	1-KM	0.5470	0.3450	0.5990	0.3870	0.5707	0.3704	0.1625	0.3049
$\tau_{\max}^{(A)}, \tau_{\max}^{(B)}$	PTID CE	0.0002	−0.0002	−0.0387	0.0004	−0.0001	−0.0001	0.0623	−0.0026
$\tau_{\max}^{(A)}, \tau_{\max}^{(B)}$	AJE	0.0011	0.0006	−0.0232	0.0024	0.0007	0.0005	0.0554	0.0000
$\tau_{\max}$	IP	0.0004	−0.0912	−0.0452	−0.0099	0.0005	−0.0705	0.0275	−0.0184
$\tau_{\max}$	PTID	0.4898	0.3299	0.0570	0.3763	0.5363	0.3630	0.1958	0.2851
$\tau_{\max}$	1-KM	0.4942	0.3314	0.3112	0.3839	0.5502	0.3632	0.1621	0.2864
$\tau_{\max}$	PTID CE	0.0001	−0.0002	−0.0441	0.0003	−0.0000	−0.0001	0.0625	−0.0014
$\tau_{\max}$	AJE	0.0004	0.0002	−0.0262	0.0020	0.0005	0.0003	0.0554	−0.0019
τ_{P90}	IP	0.0003	−0.0495	−0.0312	−0.0050	0.0005	−0.0407	0.0309	−0.0096
τ_{P90}	PTID	0.2070	0.1398	−0.0279	0.2306	0.2293	0.1567	0.1361	0.1514
τ_{P90}	1-KM	0.2072	0.1398	−0.0247	0.2326	0.2302	0.1572	0.0894	0.1514
τ_{P90}	PTID CE	0.0001	−0.0001	−0.0313	0.0105	0.0000	0.0002	0.0737	0.0075
τ_{P90}	AJE	0.0003	0.0000	−0.0297	0.0022	0.0005	0.0006	0.0516	−0.0013
τ_{P60}	IP	0.0002	−0.0147	−0.0179	−0.0007	0.0003	−0.0125	0.0267	−0.0024
τ_{P60}	PTID	0.0542	0.0349	−0.0179	0.0746	0.0610	0.0400	0.0512	0.0441
τ_{P60}	1-KM	0.0537	0.0343	−0.0178	0.0747	0.0608	0.0395	0.0391	0.0436
τ_{P60}	PTID CE	0.0002	−0.0001	−0.0179	0.0076	0.0001	0.0002	0.0429	0.0041
τ_{P60}	AJE	0.0002	−0.0001	−0.0179	0.0014	0.0003	0.0003	0.0339	−0.0004
Mean relative bias (Ratio to true simulated value − 1)									
$\tau_{\max}^{(A)}, \tau_{\max}^{(B)}$	IP	0.0007	−0.2545	−0.2545	−0.0165	−0.0005	−0.2338	0.0719	−0.0258
$\tau_{\max}^{(A)}, \tau_{\max}^{(B)}$	PTID	1.3739	0.9242	0.8283	0.6194	1.7317	1.1960	0.2902	0.4291
$\tau_{\max}^{(A)}, \tau_{\max}^{(B)}$	1-KM	1.4189	0.9115	4.0246	0.6356	1.7872	1.1677	0.2466	0.4401
$\tau_{\max}^{(A)}, \tau_{\max}^{(B)}$	PTID CE	−0.0016	−0.0033	−0.2093	−0.0002	−0.0029	−0.0039	0.1171	−0.0043
$\tau_{\max}^{(A)}, \tau_{\max}^{(B)}$	AJE	0.0007	−0.0023	−0.1028	0.0032	−0.0005	−0.0030	0.1081	−0.0005
$\tau_{\max}$	IP	−0.0010	−0.2508	−0.2652	−0.0171	−0.0013	−0.2325	0.0721	−0.0274
$\tau_{\max}$	PTID	1.2779	0.8950	0.4608	0.6182	1.6931	1.1766	0.2904	0.4145
$\tau_{\max}$	1-KM	1.2842	0.8831	2.1522	0.6307	1.7238	1.1492	0.2463	0.4164
$\tau_{\max}$	PTID CE	−0.0018	−0.0034	−0.2578	−0.0002	−0.0029	−0.0039	0.1175	−0.0026
$\tau_{\max}$	AJE	−0.0010	−0.0030	−0.1278	0.0025	−0.0013	−0.0035	0.1083	−0.0033
τ_{P90}	IP	−0.0017	−0.1688	NA	−0.0101	−0.0013	−0.1611	0.0871	−0.0182
τ_{P90}	PTID	0.6174	0.4631	NA	0.4329	0.8037	0.6015	0.2334	0.2710
τ_{P90}	1-KM	0.6169	0.4614	NA	0.4368	0.8059	0.6017	0.1685	0.2707
τ_{P90}	PTID CE	−0.0020	−0.0039	NA	0.0190	−0.0029	−0.0032	0.1466	0.0125
τ_{P90}	AJE	−0.0017	−0.0035	NA	0.0034	−0.0013	−0.0016	0.1160	−0.0033
τ_{P60}	IP	−0.0035	−0.0874	NA	−0.0028	−0.0022	−0.0834	0.1076	−0.0106
τ_{P60}	PTID	0.2485	0.1882	NA	0.2278	0.3188	0.2436	0.1562	0.1381

TABLE 6 (Continued)

		Group A Scenario				Group B Scenario			
Follow-up time	**Estimator**	**S2**	**S3**	**S5**	**S10**	**S2**	**S3**	**S5**	**S10**
τ_{P60}	1-KM	0.2462	0.1842	NA	0.2282	0.3172	0.2398	0.1323	0.1362
τ_{P60}	PTID CE	−0.0038	−0.0066	NA	0.0226	−0.0034	−0.0044	0.1397	0.0104
τ_{P60}	AJE	−0.0035	−0.0066	NA	0.0035	−0.0022	−0.0037	0.1220	−0.0040

Note: $\tau_{\max}^{(A)}$, $\tau_{\max}^{(B)}$ denotes the maximum follow-up times in group A and group B, respectively, $\tau_{\max}$ the minimum of the two maximal event times, τ_{P90} and τ_{P60} chosen according to the quantiles of the time and account for different lengths of follow-up between group A and group B. Incidence proportion is abbreviated by IP, probability transform incidence density by PTID and Kaplan–Meier by KM ($N_{\text{REP}} = 10{,}000$).

AE hazards and the hazards $\bar{\lambda}_{\text{A}}(t)$ and $\bar{\lambda}_{\text{B}}(t)$ to the CE hazards in group A and group B, respectively. In contrast, scenarios S4–S10 investigate the case where at least one hazard is time-dependent. In the scenarios S1, S2, S4, S7, and S8 the data are completely observed, that is, without censoring. In scenarios S3, S5, S6, and S9 between 10 and 28 percent of the observations are censored whereas in scenario S10 similar to the data example only very little censoring is present. The censoring times were generated from a uniform distribution and are independent of the event times.

4.1 | Probability estimators

As in the data example, the comparisons are also conducted at several follow-up times. Table 4 displays the mean simulated follow-up times, that is, the mean simulated values of $\tau_{\max}^{(A)}$, $\tau_{\max}^{(B)}$, $\tau_{\max}$, τ_{P90} and τ_{P60} as they take different values in each simulation run. Therefore, as the true simulated value depends on the follow-up time, it is not the same value over all simulation runs, but different in each simulation run. For example, the true simulated value of the probability of the event of interest in group A is calculated by $P\left(T \leq \tau, \epsilon = 1 | \text{group A}\right) = \int_0^\tau \exp\left(-\int_0^u \lambda_{\text{A}}(t) + \bar{\lambda}_{\text{A}}(t) \mathrm{d}t\right) \lambda_{\text{A}}(u) \mathrm{d}u$. For this reason, we calculate the mean simulated value over all $N_{\text{REP}} = 10{,}000$ simulation runs. The mean simulated AE probabilities at the different follow-up times of group A and B are displayed in Table 5. For more stable results in the calculation of the mean simulated value, the mean was calculated on the logit-transformed results and back-transformed to the probability scale. For the same reason in the calculation of the relative mean bias in Table 6, the mean was calculated on the log ratios and then exp-transformed back to the original scale. The Tables 5 and 6 show three important quantities of the comparison: the mean simulated AE probabilities at the different follow-up times of group A and B, the absolute mean bias and the relative mean bias.

The incidence proportion underestimates the AE probability if censoring is present, and therefore, at all follow-up times in scenarios S3, S5, and S10. Under a large percentage of censoring as in scenario S3 at the maximum follow-up time, the AE probability estimated by the incidence proportion is on average about 9% smaller than the mean simulated AE probability.

The probability transform of the incidence density and the 1-Kaplan–Meier estimator are higher than the simulated value. For the 1-Kaplan–Meier estimator this difference is the most pronounced in scenario S5 in group A as this is the scenario with the most CEs. At the maximum follow-up time in group A the AE probability estimated by the 1-Kaplan–Meier estimator is on average about 400% increased compared to the simulated value. For smaller follow-up times the differences are less pronounced since fewer CEs are present. Furthermore, it is notable that for group A in scenario S5 there is a huge difference between the estimate of the probability transform of the incidence density and the 1-Kaplan–Meier estimator. Due to the quadratic term in the hazard of the event of interest the assumption of constant hazards for the AE hazard is severely violated resulting in quite different estimates of the two estimators that censor CEs.

In the scenarios S2, S3, and S10 when considering the mean absolute and relative bias the Aalen-Johansen estimator and the probability transform of the incidence density accounting for CEs are both unbiased. But if both hazards are non-constant as in scenario S5 the probability transform of the incidence density accounting for CEs is more biased.

In scenario 5 at the maximum follow-up time all estimators are biased. But the Aalen-Johansen estimator obtains the smallest bias in group A and the second smallest bias in group B. Note that in scenario S5 group A the mean relative bias is marked as NA for all estimators at τ_{P90} and τ_{P60} as in most simulation runs no events of interest were observed until then resulting in a ratio of 0.

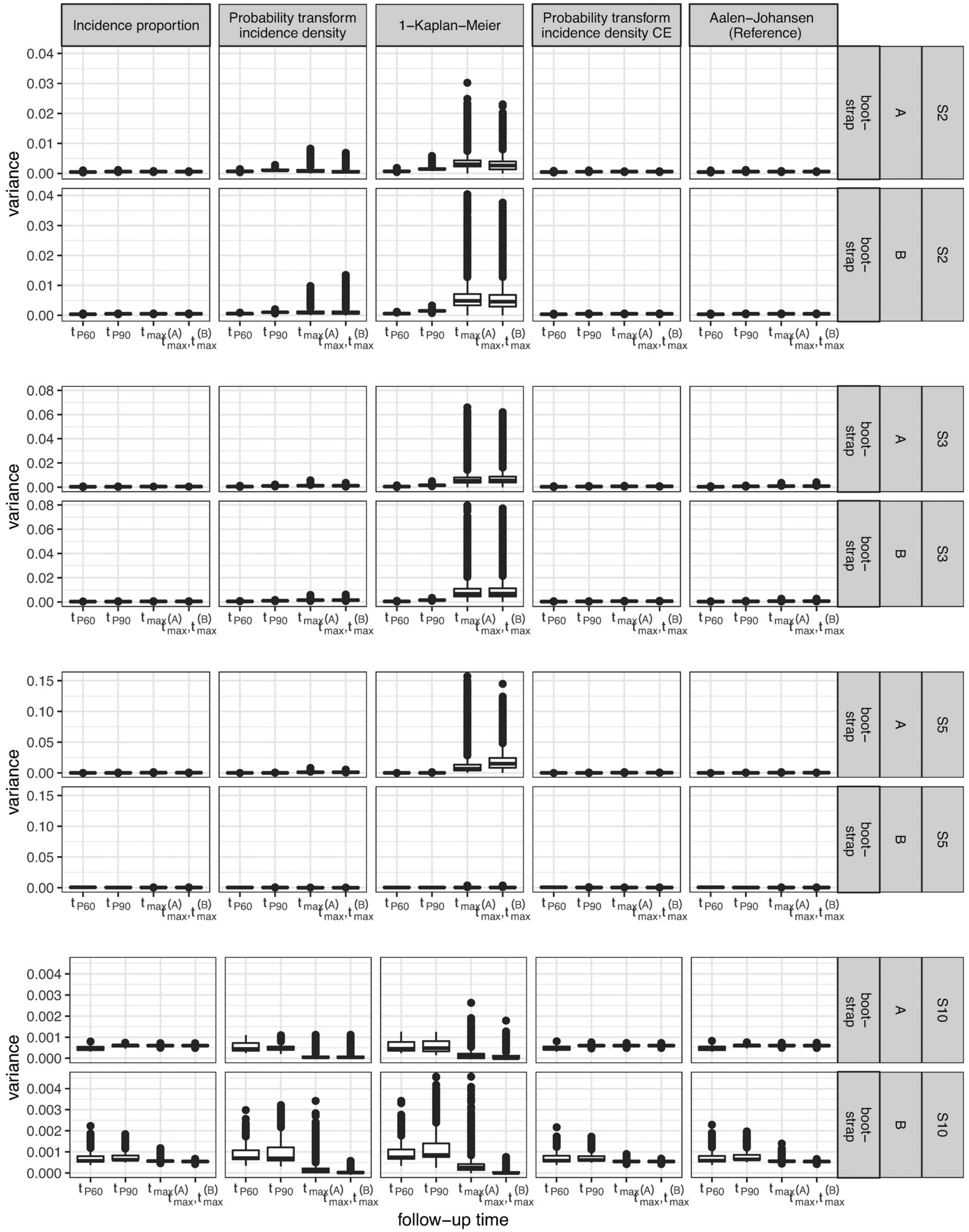

FIGURE 4 Boxplots of the estimated variances for each simulation scenario. The variances are estimated by two different approaches, model-based and bootstrap. The following follow-up times are considered $\tau_{\max}^{(A)}$, $\tau_{\max}^{(B)}$ the maximum follow-up times in group A and group B, respectively, $\tau_{\max}$ the minimum of the two maximal event times, τ_{P90} and τ_{P60} chosen according to the quantiles of the time and account for different lengths of follow-up between group A and group B ($N_{\text{REP}} = 10{,}000$)

TABLE 7 Simulation results: Mean simulated relative risk (RR)

	Scenario			
Follow-up time	**S2**	**S3**	**S5**	**S10**
$\tau_{\max}^{(A)}, \tau_{\max}^{(B)}$	1.2133	1.1996	0.1876	0.8787
τ_{P60}	1.2102	1.1952	0.1762	0.8854
τ_{P90}	1.1729	1.1561	0.0409	0.9570
τ_{P60}	1.1254	1.1158	0.0035	1.0503

Note: The following follow-up times are considered: $\tau_{\max}^{(A)}$, $\tau_{\max}^{(B)}$ the maximum follow-up times in group A and group B, respectively, $\tau_{\max}$ the minimum of the two maximal event times, τ_{P90} and τ_{P60} chosen according to the quantiles of the time and account for different lengths of follow-up between group A and group B ($N_{\text{REP}} = 10{,}000$).

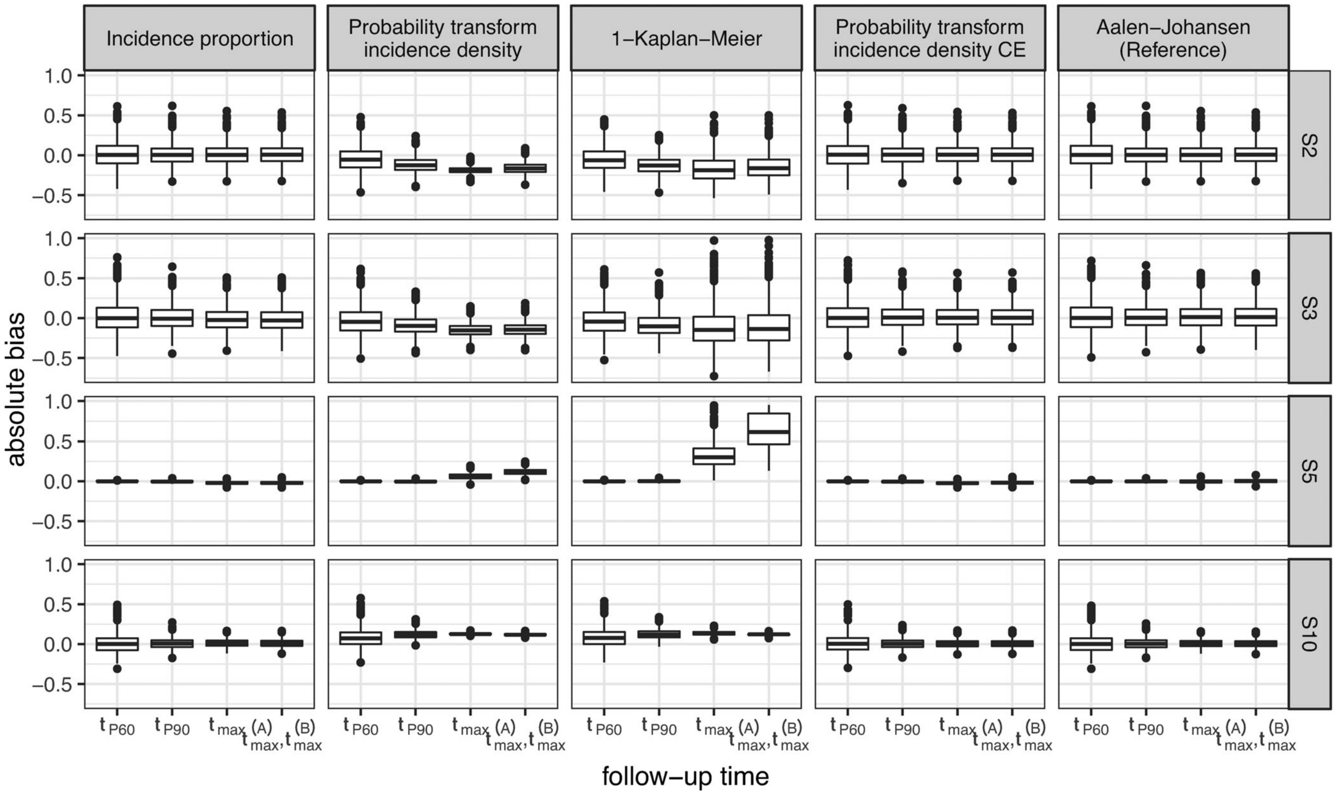

FIGURE 5 Boxplots of the difference between the relative risk of the estimators and the simulated relative risks (mean absolute bias). The following follow-up times are considered $\tau_{\max}^{(A)}$, $\tau_{\max}^{(B)}$ the maximum follow-up times in group A and group B respectively, $\tau_{\max}$ the minimum of the two maximal event times, τ_{P90} and τ_{P60} chosen according to the quantiles of the time and account for different lengths of follow-up between group A and group B ($N_{\text{REP}} = 10{,}000$)

The difference between the 1-Kaplan–Meier estimator and the simulated values is larger than the one between the parametric counterpart of the Aalen-Johansen estimator or the Aalen-Johansen estimator and the simulated values. Therefore, it seems that ignoring CEs may be more harmful than falsely assuming constant hazards.

In scenarios in which at least one hazard is constant the Aalen-Johansen estimator and the probability transform of the incidence density accounting for CEs perform similar but in a scenario with both non-constant hazards as in scenario S5 the Aalen-Johansen estimator still has a small bias compared to the other estimators. In real data analyses the true value is unknown and if one estimator needs to be chosen to work in any situation, the Aalen-Johansen estimator is the best choice. This makes the Aalen-Johansen the reference estimator in this context. As long as there is no censoring present the incidence proportion and the Aalen-Johansen estimator coincide and therefore have the same bias (scenario S2). The probability transform of the incidence density and the 1-Kaplan–Meier estimator are more biased than the Aalen-Johansen estimator.

4.2 | Variance estimators

Analogously to the data example, the estimated variances are calculated model-based with the formulas from Section 2.2 and with a non-parametric bootstrap. The boxplots in Figure 4 display the estimated variances in all simulation scenarios for all considered probability estimators.

Considering the boxplots of the variances of the estimators one immediately notices the increased variances of the Kaplan–Meier estimator for the scenarios S2, S3, and group A of S5. The reason for these outliers is that, if the last event is an AE, the Kaplan–Meier estimator equals 0 and as a consequence, the 1-Kaplan–Meier estimate remains unchanged at 1 near the end of follow-up in some of the bootstrap replications. The model-based variance using the Greenwood estimator is slightly smaller but still increased compared to the other estimators. This problem was already mentioned in Pocock et al.[15] and is the motivation why we also consider some earlier follow-up times. Late censored observations and late CEs as present in group B of scenario S5 and scenario S10 prevent the Kaplan–Meier estimator from dropping to 0 and therefore result in a smaller variance.

Furthermore, the parametric counterpart of the Kaplan–Meier estimator is less sensitive to the type of the last event. The variance of this estimator has only a few outliers at the end of follow-up. In this comparison, the parametric estimator has a smaller variance than the non-parametric Kaplan–Meier estimator.

In scenario S10 in group B an increased variance for earlier follow-up times can be detected. The reason is that censoring, increasing the variance, occurs early and the last event is rarely censored.

TABLE 8 Simulation results: Mean relative bias of the relative risk estimators compared to the mean true simulated relative risk in Table 7.

		Scenario			
Follow-up time	**Estimator**	**S2**	**S3**	**S5**	**S10**
$\tau_{\max}^{(A)}, \tau_{\max}^{(B)}$	Incidence proportion	0.0012	−0.0270	−0.3045	0.0096
$\tau_{\max}^{(A)}, \tau_{\max}^{(B)}$	Probability transform incidence density	−0.1310	−0.1237	0.4172	0.1332
$\tau_{\max}^{(A)}, \tau_{\max}^{(B)}$	1-Kaplan–Meier	−0.1321	−0.1182	3.0305	0.1358
$\tau_{\max}^{(A)}, \tau_{\max}^{(B)}$	Probability transform incidence density CE	0.0013	0.0006	−0.2922	0.0041
$\tau_{\max}^{(A)}, \tau_{\max}^{(B)}$	Aalen-Johansen	0.0012	0.0007	−0.1904	0.0038
$\tau_{\max}$	Incidence proportion	0.0002	−0.0239	−0.3146	0.0106
$\tau_{\max}$	Probability transform incidence density	−0.1541	−0.1294	0.1321	0.1440
$\tau_{\max}$	1-Kaplan–Meier	−0.1614	−0.1238	1.5292	0.1513
$\tau_{\max}$	Probability transform incidence density CE	0.0011	0.0005	−0.3359	0.0024
$\tau_{\max}$	Aalen-Johansen	0.0002	0.0004	−0.2131	0.0059
τ_{P90}	Incidence Proportion	−0.0004	−0.0092	NA	0.0082
τ_{P90}	Probability transform incidence density	−0.1033	−0.0864	NA	0.1274
τ_{P90}	1-Kaplan–Meier	−0.1046	−0.0876	NA	0.1307
τ_{P90}	Probability transform incidence density CE	0.0008	−0.0007	NA	0.0063
τ_{P90}	Aalen-Johansen	−0.0004	−0.0019	NA	0.0068
τ_{P60}	Incidence Proportion	−0.0013	−0.0043	NA	0.0079
τ_{P60}	Probability transform incidence density	−0.0533	−0.0446	NA	0.0789
τ_{P60}	1-Kaplan–Meier	−0.0539	−0.0449	NA	0.0810
τ_{P60}	Probability transform incidence density CE	−0.0004	−0.0022	NA	0.0120
τ_{P60}	Aalen-Johansen	−0.0004	−0.0022	NA	0.0120

Note: The following follow-up times are considered $\tau_{\max}^{(A)}, \tau_{\max}^{(B)}$ the maximum follow-up times in group A and group B, respectively, $\tau_{\max}$ the minimum of the two maximal event times, τ_{P90} and τ_{P60} chosen according to the quantiles of the time and account for different lengths of follow-up between group A and group B ($N_{\text{REP}} = 10{,}000$).

However, it is notable that differences between the estimated variances of the non-parametric Aalen-Johansen estimator and its parametric counterpart are small. Furthermore, the bootstrapped variances of the incidence proportion, of the probability transform of the incidence density accounting for CEs and of the Aalen-Johansen estimator are comparable to the model-based ones.

4.3 | Estimating the relative risk

In the data example, estimators that overestimate the AE probability underestimate the relative risk. This is further investigated in the simulations.

Table 7 displays the mean simulated relative risk. The different scenarios consider situations of a larger probability in group A (scenario S2, S3, and S10 at τ_{P60}) and of a smaller probability in group A (scenario S5 and S10 except τ_{P60}).

The boxplots in Figure 5 display the mean absolute bias of the relative risk calculated with the estimators compared to the simulated relative risk. Table 8 displays the corresponding mean relative bias. Using the incidence proportion, the probability transform of the incidence density accounting for CEs or the Aalen-Johansen estimator when comparing two treatment groups in terms of the relative risk induced no bias in the scenarios with constant hazards without censoring and time-dependent hazards. In the scenario with constant hazards and censoring (scenario S3) the incidence proportion results in a negative bias whereas the probability transform of the incidence density accounting for CEs and the Aalen-Johansen estimator are unbiased.

Using either the probability transform of the incidence density or the 1-Kaplan–Meier estimator the relative risk is smaller than the simulated relative risk in scenario S2 and S3 where there is a beneficial effect of B but higher in scenario S5 at $\tau_{\max}^{(A)}$, $\tau_{\max}^{(B)}$ and $\tau_{\max}$ and in scenario S10 at $\tau_{\max}^{(A)}$, $\tau_{\max}^{(B)}$, $\tau_{\max}$ and τ_{P90} where a positive effect of A is simulated. This is a consequence of the bias being greater in group B in scenarios S2 and S3, respectively, the bias being smaller in group B in scenario S5 at $\tau_{\max}^{(A)}$, $\tau_{\max}^{(B)}$ and τ_{max} and in scenario S10 at $\tau_{\max}^{(A)}$, $\tau_{\max}^{(B)}$, τ_{max} and τ_{P90}. Again as the probability of the event of interest in group A is often estimated as 0 in most cases in scenario S5 the mean relative bias is marked as NA.

The estimators that perform well in terms of the mean absolute or relative bias in both groups in Table 6, that is, have a negligible bias, also perform well when estimating the relative risk. The combination of the Tables 6 and 8 show that there is no case in which the AE probability estimates are severely biased but the estimated relative risk is unbiased.

5 | CONCLUSIONS AND DISCUSSION

In this paper, we compared several estimators quantifying the AE probability. The reference estimator in a time-to-event analysis with CEs is the Aalen-Johansen estimator. Simulations and the real data example illustrated that using the probability transform of the incidence density accounting for CEs or the incidence proportion may also provide unbiased estimates as long as the hazards are constant or there is no censoring at the considered follow-up time. In this context, we note that data monitoring committees or data safety monitoring boards will typically face situations characterized by high amounts of censored observations when monitoring adverse events during an ongoing trial.

In AE analyses with many CEs as progression, treatment discontinuation, or death only few observations may be censored in a time-to-first-event (and type-of-first-event) setting. Therefore, the use of the incidence proportion may often be justified if censoring is low and the CEs have been specified appropriately. This emphasizes the need to consider thoroughly how the CEs are defined in the specific trial as this directly impacts the amount of censoring, see Stegherr et al.[18] for guidance on this aspect. As outlined earlier, the type of CEs will also have an impact on the type of estimands, typically either while on treatment or treatment policy, and the type of CEs may be influenced by the design of AE follow-up. Moreover, in the presence of CEs the probability transform of the incidence density and the 1-Kaplan–Meier estimator always overestimate the AE probability. Using the 1-Kaplan–Meier estimator or the probability transform of the incidence density in group comparisons in terms of the relative risk can lead to the opposite conclusion about a therapy's safety as compared to using the Aalen-Johansen estimator.

The incorrect use of the Kaplan–Meier estimator in the presence of CEs also occurs in other disease areas, among others in the context of cardiovascular diseases.[9,10] The assumption of constant hazards has been challenged for events associated with aging, for example, death and some types of cancer, as the hazards increase with time, or for events that occur early, for example, remission, as there the hazards decrease with time.[5]

To summarize, in the literature about the analyses of AEs[16,27] mainly the constant hazards assumption is criticized, but one of our main results is that ignoring CEs and treating them falsely as censoring at the event time in a Kaplan–Meier-type calculation may be worse than misspecifying the model by falsely assuming constant hazards.

In practice a common way to report the AE probability is by the use of frequency categories.[28,29] Thereby, the AE probability is categorized to "very common" if the probability is greater than 10%, "common" if the probability is between 1% and 10%, "uncommon" if the probability is between 0.1% and 1%, "rare" if the probability is between 0.01% and 0.1%, and "very rare" if the probability is smaller than 0.01%. For very common AEs as here in the DECIDER trial and in most scenarios of the simulation study often all estimators derive the same category. But for smaller categories the choice of the estimator plays an important role. An absolute bias of 0.0001 may lead to a different frequency category for a very rare AE.

AEs may also be recurrent,[27] but the issue of CEs will remain as relevant as in time-to-first-event analyses. Recent papers by Charles-Nelson et al.[30] and by Andersen et al.[31] emphasize the importance of considering competing (terminal) events also in analyses of recurrent events. See these references[27,32,33] for suggestions on the analysis of recurrent AEs.

Here we emphasize the need to carefully consider the different follow-up times in safety analyses. One reason is that due to a small number of patients still at risk at the end of follow-up the variances of the estimators may be increased.[15] We investigated this aspect using times $\tau_{\max}$, τ_{P90} and τ_{P60}, defined via quantiles in treatment groups. We note that for times τ_{P90} and τ_{P60} this comes with possibly reduced statistical power, especially if AEs occur late in the study period. Another aspect with different follow-up times between groups is to calculate the relative risk using one common time point in both groups as otherwise, the interpretation may be difficult.[14] We reiterate that also the common logrank test only compares two groups while the risk sets are nonempty in both groups, that is, up to time $\tau_{\max}$. Here, we only considered the relative risk, since it is frequently used and often considered easy to interpret. But similar conclusions can be drawn for the risk difference (results not shown).

We further compared the model-based and the bootstrapped variance estimates of the estimators. Thereby, for the incidence proportion, the Kaplan–Meier estimator, the probability transform of the incidence density accounting for CEs, and the Aalen-Johansen estimator no relevant differences between the two variance estimators were found. For the probability transform of the incidence density, we found a smaller estimated variance for the model-based approach than for the bootstrapped one. The model-based variance assumes constant hazards whereas the nonparametric bootstrapped variance estimator does not.[13]

Here, we focused on the comparison of probabilities and the ratio of probabilities (relative risk) instead of the comparison of hazards. Interpretation of hazards and hazard ratios is often challenging in particular in the presence of CEs.[34] Arguably, the best way to communicate results is by visualizing the estimated probabilities for the two groups. All estimators introduced in Section 2 can easily be visualized in plots (see Figure 2), although theses plots do not provide direct information about the possible dynamic pattern of the treatment effect.[35] Here, to better understand the differences between the estimated AE probabilities in the data example, we also considered a plot of the cumulative hazards estimated by the Nelson-Aalen estimator in Figure 3. Therefore, as most analyses consider both the probabilities and the hazards one other point of future research is to compare different estimators of the hazard ratio, for example, the hazard ratio obtained by the Cox proportional hazards model, the ratio of the Nelson-Aalen estimators and the ratio of the incidence density of both groups.

To continue the discussion on studying probabilities or hazards in a time-to-AE, hence, in a competing events setting, we note than one major motivation for using the cumulative AE probability as our basic estimand, see Section 2.1, was that both the commonly used incidence proportion and the Kaplan–Meier estimator target the probability scale. This is also the scale that informs AE frequency categories and it motivated conversion of the also commonly used incidence density onto the probability scale, either in the spirit of Kaplan–Meier or accounting for CEs. We note that in this context the presentation of the use of incidence density as opposed to incidence proportions is often somewhat technical, focusing on the different denominators, but hardly on the parameter being estimated, see Beyersmann and Schrade[21] for a discussion in the context of AEs during hospital stay.

However, incidence densities are quite instructive for a discussion on studying probabilities or hazards for time-to-AE. An ideal situation when weighing the benefits and harms of a treatment should be a situation where there is no effect on AEs. Formalizing this as no effect on the AE hazard λ (and assuming constant hazards), recall that the any-time AE probability is

$$\frac{\lambda}{\lambda+\bar{\lambda}},$$

with CE hazard $\bar{\lambda}$. If the treatment is beneficial in that it reduces the CE hazard, the anytime AE probability will be increased. Therefore, the use of the probability scale is often criticized as "biased," but we believe that the effect at hand is just common sense: The time spent in the initial state of Figure 1 is prolonged, because the all-events-hazard $\lambda+\bar{\lambda}$ is reduced. During this prolonged time, an unaltered AE hazards acts, leading to the effect on the AE probability just described. However, this scenario does shed a light, firstly, on the subtle relation between hazards and probabilities when there is more than just one event type, and, secondly, on the important question of how to formalize effects in the presence of CEs. Arguably, "no effect on AE" is perhaps most easily formalized on the hazard scale, but the above simple calculation demonstrates that one must account for CEs. Again, the Kaplan–Meier estimator does not achieve this and would consequently provide a misleading picture on the probability scale.

The analyses of this paper are motivated by the Survival analysis for AdVerse events with VarYing follow-up times project (SAVVY).[18] This is a joint project of academic institutions and pharmaceutical companies with the aim to improve standards for the reporting of incidences of adverse events. Whereas this paper is more focused on a methodological comparison of the estimators under different settings, the SAVVY project includes an empirical study calculating estimators mentioned in this paper in several randomized controlled trials and summarizing the results in a meta-analysis to assess the sources of bias in safety analyses in real clinical trials. In the end, both this paper and the SAVVY project intend to reduce the usage of biased estimators in the analysis of AEs.

To sum up and to return to the three questions raised in the introduction of this paper: The answer to question (i) is that the choice of the estimator is also crucial for group comparisons in terms of the relative risk. The bias in calculating the AE probabilities and variances of the AE probabilities do also directly influence the relative risk and the confidence intervals. Regarding question (ii), since the probability transform of the incidence density accounting for CEs is less biased with respect to the Aalen-Johansen estimator than of 1-Kaplan–Meier estimator, ignoring CEs can be worse than falsely assuming constant hazards. When then considering question (iii), for earlier time points of analysis for most AE probability and relative risk estimators the bias with respect to the Aalen-Johansen estimator is smaller. Especially, for the incidence proportion the bias is almost negligible.

ACKNOWLEDGMENTS

R.S. was partially supported by Grant BE 4500/3-1 of the German Research Foundation (DFG). The DECIDER trial was supported by the German Federal Ministry of Education and Research (BMBF, Clinical Trials Program Grant 01KG0913). Open Access funding enabled and organized by Projekt DEAL.

CONFLICT OF INTEREST

The authors declare no potential conflict of interests.

DATA AVAILABILITY STATEMENT

The data of the clinical trial are not publicly available due to confidentiality restrictions. The simulated data that support the findings of this study are available on request from the corresponding author.

ORCID

Regina Stegherr https://orcid.org/0000-0002-0528-348X
Tim Friede https://orcid.org/0000-0001-5347-7441

REFERENCES

1. Unkel S, Amiri M, Benda N, et al. On estimands and the analysis of adverse events in the presence of varying follow-up times within the benefit assessment of therapies. *Pharm Stat*. 2019;18(2):166-183.
2. O'Neill RT. Statistical analyses of adverse event data from clinical trials: special emphasis on serious events. *Drug Inf J*. 1987;21(1):9-20.
3. Allignol A, Beyersmann J, Schmoor C. Statistical issues in the analysis of adverse events in time-to-event data. *Pharm Stat*. 2016;15(4):297-305.
4. Bender R, Beckmann L, Lange S. Biometrical issues in the analysis of adverse events within the benefit assessment of drugs. *Pharm Stat*. 2016;15(4):292-296.
5. Kraemer HC. Events per person-time (incidence rate): a misleading statistic? *Stat Med*. 2009;28(6):1028-1039.
6. Lacny S, Wilson T, Clement F, et al. Kaplan-Meier survival analysis overestimates the risk of revision arthroplasty: a meta-analysis. *Clin Orthop Relat Res*. 2015;473(11):3431-3442.
7. Lacny S, Wilson T, Clement F, et al. Kaplan–Meier survival analysis overestimates cumulative incidence of health-related events in competing risk settings: a meta-analysis. *J Clin Epidemiol*. 2018;93:25-35.

8. Aalen OO, Søren J. An empirical transition matrix for non-homogeneous Markov chains based on censored observations. *Scand J Stat*. 1978;5(3):141-150.
9. van Walraven C, McAlister FA. Competing risk bias was common in Kaplan-Meier risk estimates published in prominent medical journals. *J Clin Epidemiol*. 2016;69:170-173.
10. Schumacher M, Ohneberg K, Beyersmann J. Competing risk bias was common in a prominent medical journal. *J Clin Epidemiol*. 2016; 80:135-136.
11. Hollaender N, Gonzaleu-Maffe J, Jehl V. Quantitative assessment of adverse events in clinical trials: comparison of methods at an interim and the final analysis. *Biom J*. 2020;62:658-669.
12. Beyersmann J, Schmoor C. Textbook of clinical trials in oncology: a statistical perspective. In: Halabi S, Michiels S, eds. *Chapter the Analysis of Adverse Events in Randomized Clinical Trials*. Boca Raton: Chapman and Hall/CRC; 2019.
13. Hjort NL. On inference in parametric survival data models. *Int Stat Rev*. 1992;60(3):355-387.
14. Kunz LM, Normand SLT, Sedrakyan A. Meta-analysis of rate ratios with differential follow-up by treatment arm: inferring comparative effectiveness of medical devices. *Stat Med*. 2015;34(21):2913-2925.
15. Pocock SJ, Clayton TC, Altman DG. Survival plots of time-to-event outcomes in clinical trials: good practice and pitfalls. *Lancet*. 2002; 359(9318):1686-1689.
16. Bender R, Beckmann L. Limitations of the incidence density ratio as approximation of the hazard ratio. *Trials*. 2019;20(1):485.
17. ICH. ICH E9 (R1): addendum on estimands and sensitivity analysis in clinical trials to the guideline on statistical principles for clinical trials; 2019. https://database.ich.org/sites/default/files/E9-R1_Step4_Guideline_2019_1203.pdf. Accessed October 19, 2020.
18. Stegherr R, Beyersmann J, Jehl V, et al. Survival analysis for AdVerse events with VarYing follow-up times (SAVVY): rationale and statistical concept of a meta-analytic study. *Biom J*. 2021;63:650-670.
19. Andersen PK, Gill RD, Borgan Ø, Keiding N. *Statistical Models Based on Counting Processes*. New York, NY: Springer; 1993.
20. Allignol A, Schumacher M, Beyersmann J. Empirical transition matrix of multi-state models: the etm package. *J Stat Softw*. 2011;38(4):1-15.
21. Beyersmann J, Schrade C. Florence nightingale, William Farr and competing risks. *J Roy Stat Soc Ser A*. 2017;180(1):285-293.
22. Bonofiglio F, Beyersmann J, Schumacher M, Koller M, Schwarzer G. Meta-analysis for aggregated survival data with competing risks: a parametric approach using cumulative incidence functions. *Res Synth Methods*. 2016;7(3):282-293.
23. Grishina O, Schmoor C, Döhner K, et al. DECIDER: prospective randomized multicenter phase II trial of low-dose decitabine (DAC) administered alone or in combination with the histone deacetylase inhibitor valproic acid (VPA) and all-trans retinoic acid (ATRA) in patients >60years with acute myeloid leukemia who are ineligible for induction chemotherapy. *BMC Cancer*. 2015;15(1):430.
24. Lübbert M, Grishina O, Schmoor C, et al. Valproate and retinoic acid in combination with decitabine in elderly nonfit patients with acute myeloid leukemia: results of a multicenter, randomized, 2×2, phase II trial. *J Clin Oncol*. 2020;38(3):257-270.
25. Beyersmann J, Latouche A, Buchholz A, Schumacher M. Simulating competing risks data in survival analysis. *Stat Med*. 2009;28(6): 956-971.
26. Beyersmann J, Allignol A, Schumacher M. *Competing Risks and Multistate Models with R*. New York, NY: Springer; 2012.
27. Siddiqui O. Statistical methods to analyze adverse events data of randomized clinical trials. *J Biopharm Stat*. 2009;19(5):889-899.
28. CIOMS Working Groups III and V. Guidelines for preparing core clinical—Safety information on drugs. Geneva: Council for International Organizations of Medical Sciences; 1999.
29. EMA. A guideline on summary of product characteristics (SmPC); 2009. https://ec.europa.eu/health/sites/health/files/files/eudralex/vol-2/c/smpc_guideline_rev2_en.pdf. Accessed July 3, 2020.
30. Charles-Nelson A, Katsahian S, Schramm C. How to analyze and interpret recurrent events data in the presence of a terminal event: an application on readmission after colorectal cancer surgery. *Stat Med*. 2019;38(18):3476-3502.
31. Andersen PK, Angst J, Ravn H. Modeling marginal features in studies of recurrent events in the presence of a terminal event. *Lifetime Data Anal*. 2019;25(4):681-695.
32. Nelson W. Recurrent events data analysis for product repairs, disease recurrences, and other applications. ASA-SIAM Series on Statistics and Applied Probability. Society for Industrial and Applied Mathematics; 2003.
33. Güttner 'A, Kübler J, Pigeot I. Multivariate time-to-event analysis of multiple adverse events of drugs in integrated analyses. *Stat Med*. 2007;26(7):1518-1531.
34. Beyersmann J, Dettenkofer M, Bretz H, Schumacher M. A competing risks analysis of bloodstream infection after stem-cell transplantation using subdistribution hazards and cause-specific hazards. *Stat Med*. 2007;26(30):5360-5369.
35. Martinussen T, Vansteelandt S, Andersen PK. Subtleties in the interpretation of hazard ratios. *Lifetime Data Anal*. 2020;26(4):833-855.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Stegherr R, Schmoor C, Lübbert M, Friede T, Beyersmann J. Estimating and comparing adverse event probabilities in the presence of varying follow-up times and competing events. *Pharmaceutical Statistics*. 2021;20(6):1125–1146. https://doi.org/10.1002/pst.2130

APPENDIX A.: PROOF: THE ESTIMAND OF THE 1-KAPLAN–MEIER ESTIMATOR IS GREATER THAN THE CUMULATIVE INCIDENCE FUNCTION

For the cumulative incidence function (left-hand side below) that can be estimated by the Aalen-Johansen estimator and the estimand of the 1-Kaplan–Meier estimator (right-hand side) the following inequality holds:

$$
\begin{aligned}
P(T \le t | \epsilon = 1, \text{group A}) &= \int_0^t P(T \ge u - | \text{group A}) \lambda_{\text{A}}(u) \mathrm{d}u \\
&= \int_0^t \exp\left(- \int_0^u \lambda_{\text{A}}(s) + \bar{\lambda}_{\text{A}}(s) \mathrm{d}s \right) \lambda_{\text{A}}(u) \mathrm{d}u \\
&\star \le \int_0^t \exp\left(- \int_0^u \lambda_{\text{A}}(s) \mathrm{d}s \right) \lambda_{\text{A}}(u) \mathrm{d}u = 1 - \exp\left(- \int_0^t \lambda_{\text{A}}(u) \mathrm{d}u \right)
\end{aligned}
$$

where relation $\star$ holds since $\exp\left(- \int_0^u \bar{\lambda}_{\text{A}}(s) \mathrm{d}s \right) \le 1$ with equality if $\bar{\lambda}_{\text{A}}(s) = 0 \forall s$, that is, no CEs are present in the data. Note that relation $\star$ does not postulate the existence of latent event-specific times. As a consequence, $\exp\left(- \int_0^t \lambda_{\text{A}}(s) \mathrm{d}s \right)$ has no proper probability interpretation in settings with CEs.

APPENDIX B.: CONNECTION BETWEEN THE INCIDENCE PROPORTION AND THE AALEN-JOHANSEN ESTIMATOR

If we look more closely at the Aalen-Johansen estimator, we see two cases where it is equal to the incidence proportion. For simplicity, we only consider the end of follow-up $\tau_{\max}^{(\text{A})}$. For considering earlier follow-up times the formula can be adapted straightforwardly. We assume n_{A} individuals with m distinct event or censoring times $0 \le t_1 < \ldots < t_m = \tau_{\max}^{(\text{A})}$, $m \le n_{\text{A}}$. Then the Aalen-Johansen estimator can be written as

$$
\begin{aligned}
&\widehat{AJ}_{\text{A}}\left(\tau_{\max}^{(\text{A})}\right) = \sum_{u \in \left(0, \tau_{\max}^{(\text{A})}\right]} \prod_{v \in (0,u)} \left(1 - \Delta\widehat{\Lambda}_{\text{A}}(v) - \Delta\widehat{\bar{\Lambda}}_{\text{A}}(v)\right) \Delta\widehat{\Lambda}_{\text{A}}(u) = \sum_{u \in \left(0, \tau_{\max}^{(\text{A})}\right]} \prod_{v \in (0,u)} \left(1 - \frac{d_{\text{A}}(v) + \bar{d}_{\text{A}}(v)}{Y(v)}\right) \cdot \frac{d_{\text{A}}(u)}{Y(u)} \\
&= \frac{d_{\text{A}}(t_1)}{n_{\text{A}}} + \left(1 - \frac{d_{\text{A}}(t_1) + \bar{d}_{\text{A}}(t_1)}{n_{\text{A}}}\right) \cdot \frac{d_{\text{A}}(t_2)}{n_{\text{A}} - d_{\text{A}}(t_1) - \bar{d}_{\text{A}}(t_1)} + \left(1 - \frac{d_{\text{A}}(t_1) + \bar{d}_{\text{A}}(t_1)}{n_{\text{A}}}\right)\left(1 - \frac{d_{\text{A}}(t_2) + \bar{d}_{\text{A}}(t_2)}{n_{\text{A}} - d_{\text{A}}(t_1) - \bar{d}_{\text{A}}(t_1)}\right) \cdot \frac{d_{\text{A}}(t_3)}{n_{\text{A}} - d_{\text{A}}(t_1) - \bar{d}_{\text{A}}(t_1) - d_{\text{A}}(t_2) - \bar{d}_{\text{A}}(t_2)} \\
&+ \ldots + \left(1 - \frac{d_{\text{A}}(t_1) + \bar{d}_{\text{A}}(t_1)}{n_{\text{A}}}\right)\left(1 - \frac{d_{\text{A}}(t_2) + \bar{d}_{\text{A}}(t_2)}{n_{\text{A}} - d_{\text{A}}(t_1) - \bar{d}_{\text{A}}(t_1)}\right) \cdot \ldots \cdot \left(1 - \frac{d_{\text{A}}(t_{m-1}) + \bar{d}_{\text{A}}(t_{m-1})}{n_{\text{A}} - \left(\sum_{k=1}^{m-2} d_{\text{A}}(t_k) + \bar{d}_{\text{A}}(t_k)\right)}\right) \cdot \frac{d_{\text{A}}(t_m)}{n_{\text{A}} - \left(\sum_{k=1}^{m-1} d_{\text{A}}(t_k) + \bar{d}_{\text{A}}(t_k)\right)} \\
&= \frac{d_{\text{A}}(t_1)}{n_{\text{A}}} + \left(\frac{n_{\text{A}} - d_{\text{A}}(t_1) - \bar{d}_{\text{A}}(t_1)}{n_{\text{A}}}\right) \cdot \frac{d_{\text{A}}(t_2)}{n_{\text{A}} - d_{\text{A}}(t_1) - \bar{d}_{\text{A}}(t_1)} + \left(\frac{n_{\text{A}} - d_{\text{A}}(t_1) - \bar{d}_{\text{A}}(t_1)}{n_{\text{A}}}\right)\left(\frac{n_{\text{A}} - d_{\text{A}}(t_1) - \bar{d}_{\text{A}}(t_1) - d_{\text{A}}(t_2) - \bar{d}_{\text{A}}(t_2)}{n_{\text{A}} - d_{\text{A}}(t_1) - \bar{d}_{\text{A}}(t_1)}\right) \cdot \frac{d_{\text{A}}(t_3)}{n_{\text{A}} - d_{\text{A}}(t_1) - \bar{d}_{\text{A}}(t_1) - d_{\text{A}}(t_2) - \bar{d}_{\text{A}}(t_2)} \\
&+ \ldots + \left(\frac{n_{\text{A}} - d_{\text{A}}(t_1) - \bar{d}_{\text{A}}(t_1)}{n_{\text{A}}}\right)\left(\frac{n_{\text{A}} - d_{\text{A}}(t_1) - \bar{d}_{\text{A}}(t_1) - d_{\text{A}}(t_2) - \bar{d}_{\text{A}}(t_2)}{n_{\text{A}} - d_{\text{A}}(t_1) - \bar{d}_{\text{A}}(t_1)}\right) \cdot \ldots \cdot \left(\frac{n_{\text{A}} - \left(\sum_{k=1}^{m-1} d_{\text{A}}(t_k) + \bar{d}_{\text{A}}(t_k)\right)}{n_{\text{A}} - \left(\sum_{k=1}^{m-2} d_{\text{A}}(t_k) + \bar{d}_{\text{A}}(t_k)\right)}\right) \cdot \frac{d_{\text{A}}(t_m)}{n_{\text{A}} - \left(\sum_{k=1}^{m-1} d_{\text{A}}(t_k) + \bar{d}_{\text{A}}(t_k)\right)}
\end{aligned}
\tag{B1}
$$

We can distinguish three cases and in two of them the Aalen-Johansen estimator reduces to the binomial estimator, the incidence proportion.

- *No censoring*: In case of no censoring, there is no $i \in \{1, \ldots, m\}$ such that both $d_{\text{A}}(t_i)$ and $\bar{d}_{\text{A}}(t_i)$ are equal to 0 at t_i. Then, it is easy to see that Equation (B1) is equal to $\sum_{k=1}^{m} d_{\text{A}}(t_k) / n_{\text{A}}$, the incidence proportion.

- *Censoring only after last AE*: Assume the first m_1 observations are either AE or CE and let the last $m - m_1$ observations be either CEs or censored. This is equal to assuming there is no $i \in \{1, \ldots, m_1\}$ such that both $d_{\text{A}}(t_i)$ and $\bar{d}_{\text{A}}(t_i)$ are equal to

0 at t_i. Then Equation (B1) is equal to $\sum_{k=1}^{m_1} d_{\mathrm{A}}(t_k)/n_{\mathrm{A}}$ and as $d_{\mathrm{A}}(t_i) = 0$ for $i = m_1 + 1, \ldots, m$, it is also equal to the incidence proportion $\sum_{k=1}^{m} d_{\mathrm{A}}(t_k)/n_{\mathrm{A}}$.

- *Censoring in between AE and CE*: If there are no AEs or CEs but only censoring observed at time t_i, that is, $d_{\mathrm{A}}(t_i) = \bar{d}_{\mathrm{A}}(t_i) = 0$, for all AE after t_i in $\prod_{v \in (0,u)} \left(1 - \frac{d_{\mathrm{A}}(v) - \bar{d}_{\mathrm{A}}(v)}{Y(v)}\right)$ one (more) factor is equal to 1 instead of something smaller than 1. The incidence proportion assumes all of these factors being smaller than 1 resulting in something smaller being added for each AE after a censoring. The consequence is a smaller estimate compared to the Aalen-Johansen estimator which accounts censoring.

APPENDIX C.: MODEL-BASED VARIANCES OF THE PROBABILITY TRANSFORM OF THE INCIDENCE DENSITY ACCOUNTING FOR COMPETING EVENTS

It is known that under the assumption of constant hazards

$$\sqrt{n}\left(\left(\mathrm{I}\hat{\mathrm{D}}_{\mathrm{A}}, \mathrm{I}\hat{\bar{\mathrm{D}}}_{\mathrm{A}}\right)^T - (\theta_1, \theta_2\,)^T\right) \xrightarrow{d} Z \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \mathrm{var}(\theta_1) & 0 \\ 0 & \mathrm{var}(\theta_2) \end{pmatrix}\right).$$

Asymptotically, the incidence densities are independent. The variance of the CE incidence density can be estimated analogously to the incidence density of the AE:

$$\mathrm{var}\left(\mathrm{I}\hat{\bar{\mathrm{D}}}_{\mathrm{A}}(\tau)\right) = \frac{\sum_{u \in (0,\tau]} \bar{d}_{\mathrm{A}}(u)}{\left(\sum_{i=1}^{n_{\mathrm{A}}} \min(t_i, \tau)\right)^2}.$$

The following mapping is defined to transform the two incidence densities to the estimator accounting for CEs on the probability scale:

$$\Phi(x,y) = \frac{(1 - \exp(-\tau(x+y)))\cdot x}{(x+y)}$$

Then the multivariate delta-method can be applied and the holds:

$$\sqrt{n}\left(\Phi\left(\mathrm{I}\hat{\mathrm{D}}_{\mathrm{A}}, \mathrm{I}\hat{\bar{\mathrm{D}}}_{\mathrm{A}}\right) - \Phi(\theta_1, \theta_2)\right) \xrightarrow{d} \Phi'(\theta_1, \theta_2) Z \sim N\left(0, s_{\mathrm{A}}^2\right)$$

where s_{A}^2 is the variance of the probability transform of the incidence density and can be estimated by

$$\begin{aligned}
\hat{s}_{\mathrm{A}}^2 = &\left(\exp\left(-\tau\cdot\mathrm{I}\hat{\mathrm{D}}_{\bullet\mathrm{A}}(\tau)\right)\cdot\right. \\
&\left.\frac{\mathrm{I}\hat{\bar{\mathrm{D}}}_{\mathrm{A}}(\tau)\cdot\left(\exp\left(\tau\cdot\mathrm{I}\hat{\mathrm{D}}_{\bullet\mathrm{A}}(\tau)\right) - 1\right) + \tau\cdot\mathrm{I}\hat{\mathrm{D}}_{\mathrm{A}}(\tau)\cdot\mathrm{I}\hat{\mathrm{D}}_{\bullet\mathrm{A}}(\tau)}{\mathrm{I}\hat{\mathrm{D}}_{\bullet\mathrm{A}}(\tau)^2}\right)^2 \\
&\cdot\hat{\mathrm{var}}\left(\mathrm{I}\hat{\mathrm{D}}_{\mathrm{A}}(\tau)\right) \\
&+\left(\mathrm{I}\hat{\mathrm{D}}_{\mathrm{A}}(\tau)\cdot\exp\left(-\tau\cdot\mathrm{I}\hat{\mathrm{D}}_{\bullet\mathrm{A}}(\tau)\right)\cdot\frac{\tau\cdot\mathrm{I}\hat{\mathrm{D}}_{\bullet\mathrm{A}}(\tau) - \exp\left(\tau\cdot\mathrm{I}\hat{\mathrm{D}}_{\bullet\mathrm{A}}(\tau)\right) + 1}{\mathrm{I}\hat{\mathrm{D}}_{\bullet\mathrm{A}}(\tau)^2}\right)^2 \\
&\cdot\hat{\mathrm{var}}\left(\mathrm{I}\hat{\bar{\mathrm{D}}}_{\mathrm{A}}(\tau)\right).
\end{aligned}$$

with $\mathrm{I}\hat{\mathrm{D}}_{\bullet\mathrm{A}}(\tau) = \mathrm{I}\hat{\mathrm{D}}_{\mathrm{A}}(\tau) + \mathrm{I}\hat{\bar{\mathrm{D}}}_{\mathrm{A}}(\tau)$.