

RESEARCH PAPER

Survival analysis for AdVerse events with VarYing follow-up times (SAVVY): Rationale and statistical concept of a meta-analytic study

Regina Stegherr¹  | Jan Beyersmann¹  | Valentine Jehl² | Kaspar Rufibach³  |
Friedhelm Leverkus⁴ | Claudia Schmoor⁵  | Tim Friede⁶ 

¹ Institut für Statistik, Universität Ulm, Ulm, Germany

² Novartis Pharma AG, Basel, Switzerland

³ Methods, Collaboration, and Outreach Group (MCO), Department of Biostatistics, Hoffmann-La Roche Ltd., Basel, Switzerland

⁴ Department of HTA & OR, Pfizer, Berlin, Germany

⁵ Clinical Trials Unit, Faculty of Medicine and Medical Center, University of Freiburg, Freiburg im Breisgau, Germany

⁶ Institut für Medizinische Statistik, Universitätsmedizin Göttingen, Göttingen, Germany

Correspondence

Tim Friede, Institut für Medizinische Statistik, Universitätsmedizin Göttingen, Humboldtallee 32, 37073 Göttingen, Germany.
Email: tim.friede@med.uni-goettingen.de



This article has earned an open data badge “Reproducible Research” for making publicly available the code necessary to reproduce the reported results. The results reported in this article could fully be reproduced.

Abstract

The assessment of safety is an important aspect of the evaluation of new therapies in clinical trials, with analyses of adverse events being an essential part of this. Standard methods for the analysis of adverse events such as the incidence proportion, that is the number of patients with a specific adverse event out of all patients in the treatment groups, do not account for both varying follow-up times and competing risks. Alternative approaches such as the Aalen–Johansen estimator of the cumulative incidence function have been suggested. Theoretical arguments and numerical evaluations support the application of these more advanced methodology, but as yet there is to our knowledge only insufficient empirical evidence whether these methods would lead to different conclusions in safety evaluations. The Survival analysis for AdVerse events with VarYing follow-up times (SAVVY) project strives to close this gap in evidence by conducting a meta-analytical study to assess the impact of the methodology on the conclusion of the safety assessment empirically. Here we present the rationale and statistical concept of the empirical study conducted as part of the SAVVY project. The statistical methods are presented in unified notation, and examples of their implementation in R and SAS are provided.

KEYWORDS

clinical trials, cumulative incidence function, drug safety, meta-regression, risk-benefit assessment

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2020 The Authors. *Biometrical Journal* published by Wiley-VCH GmbH.

1 | INTRODUCTION

Time-to-event or survival endpoints are commonly encountered in clinical trials, and some literature searches have found that survival analysis is the most common advanced statistical technique in medical research (Horton & Switzer, 2005; Sato et al., 2017). Censoring is arguably the major reason to use survival methodology, since statistical inference that does not account for censoring will, in general, be biased. Primary efficacy outcomes in time-to-event studies are often comprehensive outcomes such as overall or progression-free survival (Schumacher, Ohneberg, & Beyersmann, 2016), and Kaplan–Meier curves are commonly used to estimate survival probabilities, while the log-rank test and/or the Cox model are used to compare treatment groups (Gosho, Sato, Nagashima, & Takahashi, 2018). Here, a “comprehensive outcome” is a time-to-event outcome that every patient experiences, possibly after study closure which would then result in an administratively censored observation. There are, however, concerns that many published Kaplan–Meier curves are subject to a so-called “competing risk bias” (Schumacher et al., 2016; van Walraven & McAlister, 2016). Such bias arises for non-comprehensive time-to-event outcomes, and Kaplan–Meier will then overestimate cumulative event probabilities.

Safety evaluation is an essential aspect of clinical trials beyond the efficacy evaluation of new therapies (Yang, Wittes, & Pitt, 2019) with primary focus on quantifying the incidence of adverse events (AEs). But the use of sophisticated survival methodology in practice does not translate to AE analyses, and, arguably, the major workhorses to quantify AE incidence are the incidence proportion, that is the number of patients with an observed AE (of a certain type) divided by group size, and the (exposure adjusted) incidence density, which divides by cumulative patient-time at risk. This gap, using time-to-event methodology for efficacy analyses but not for AE analyses, has been criticized by a number of authors for some time already; see Allignol et al. (2016), Bender et al. (2016), O'Neill (1987), and Unkel et al. (2019). The issue is that the incidence proportion estimates the probability that a patient experiences an AE *and* that it is observed before censoring which is less than the absolute AE risk, that is the probability of experiencing the AE. Closely related to this issue are varying follow-up times which further complicate using incidence proportions; see, in particular, Bender et al. (2016).

The incidence density, on the other hand, accounts for both censoring and varying follow-up times by considering patient-time at risk in the denominator rather than the simpler number of patients. The incidence density is not an estimator of absolute AE risk, but rather of the AE hazard under a constant hazard assumption. This rather restrictive parametric assumption has been repeatedly criticized (Bender & Beckmann, 2019; Kraemer, 2009). The Kaplan–Meier curve has been considered as a non-parametric alternative on the probability scale (Crowe et al., 2009; Siddiqui, 2009), but is subject to the aforementioned competing risk bias. This would also be true for its parametric counterpart, when translating the incidence density onto the probability scale.

The nature of competing risk bias is that one minus a Kaplan–Meier curve approximates an empirical distribution function, where the approximation is due to incompletely observed data as a consequence of censoring. Empirical distribution functions eventually reach 100%. For estimating absolute AE risk by the Kaplan–Meier method, the consequence is that one implicitly assumes that every patient would experience the AE under consideration eventually. However, this does not hold true for patients who die before the AE, and the absolute AE risk is consequently overestimated. Here, “death before AE” acts as a competing event (or “competing risk”) and must be accounted for in the statistical analysis. For instance, Allignol et al. (2016) demonstrate in a real data analysis that a simple parametric AE analysis using incidence densities but accounting for competing events may outperform a Kaplan–Meier analysis.

So, the concern is that quantifying absolute AE risk may be either underestimated (incidence proportions) or overestimated (Kaplan–Meier, incidence densities). The issues are mainly censoring, varying follow-up times, competing events, and, in the case of incidence densities, a possibly too restrictive parametric model. Here, a non-parametric benchmark method is provided by the Aalen–Johansen estimator (Aalen & Johansen, 1978) which generalizes the Kaplan–Meier estimator to multiple event types; see Beyersmann and Schmoor (2019) for a recent textbook treatment in the context of AE. The magnitude of such bias in practice on, for example, AE frequency categories will, however, depend on the frequency of competing events, the magnitude of censoring, or the difference in follow-up. In terms of between-group comparisons, the impact of, say, dividing two metrics that both underestimate or both overestimate is also unclear.

The SAVVY (Survival analysis for AdVerse events with VarYing follow-up times) project strives to close this gap in evidence by conducting a meta-analytical study. In this project, participating sponsor companies and organizations (in the following called sponsors) select randomized controlled clinical trials of special interest, particularly those with varying follow-up times between patients and possibly between treatment arms. This includes studies in different therapeutic areas. Within studies, one or more AEs of interest will be selected. Here we present the rationale and statistical concept.

The statistical methods are presented in unified notation, and the implementation in R and SAS is described adding some interest to this work beyond presenting concepts of the main study. While basic methodological considerations on AE analyses, censoring, varying follow-up times, and competing risks have been discussed elsewhere; see, for example, Unkel et al. (2019) and Beyersmann and Schmoor (2019), this paper offers additional detailed insights by considering practically relevant questions such as which kind of event should be viewed as competing and which one as a standard censoring event. For instance, Unkel et al. (2019) briefly touch upon the question whether diagnosed progression is a competing event or rather a censoring event, and if the latter whether such censoring is informative, and, finally, what the impact on AE analyses is. Here, Section 2.2 offers further guidance, also distinguishing between a “hard” competing event definition and a more encompassing one. Furthermore, an important meta-analytical issue is that estimates are typically weighted by the inverse of an estimated variance which is not straightforward in the setting here, since the aim is to compare different methodological approaches performed within one trial. As a consequence, the variance of the difference measure between, for example, Kaplan–Meier and incidence proportion is required. This difficulty has, for example, also been faced (but not solved) by Lacny et al. (2015, 2018), and Section 4.4 will explain how to bootstrap these variances.

We also note the recent discussions about which estimand to use in the context of safety analyses (Unkel et al., 2019; Yang et al., 2019). The current standard of regulatory agencies in the marketing authorization process is to address the *while on treatment* estimand, as potentially diluting effects of treatment discontinuation or switch may be regarded as anticonservative for comparative safety assessments. Health Technology Assessment (HTA) bodies are most interested in the *treatment policy* estimand, which aims at comparisons of treatment groups with regard to AEs on the entire follow-up period irrespective of intercurrent events as treatment discontinuation. Decisions about market authorization and reimbursement have to consider different aspects of treatment effects (Unkel et al., 2019). When planning a clinical trial and its statistical analysis the clinical question will drive the choice of the estimand and, as a consequence, the data collection process and the methods used. In our situation of a secondary data analysis of preexisting clinical trials, we have to take the data as they were collected by the designs of the individual trials. Although we are interested in a safety comparison of treatments over the entire follow-up period of patients, that is addressing the treatment policy estimand, we are faced with the situation that up to now in many clinical trials AEs are only observed until some prespecified point in time after the last dose of treatment leading to estimates addressing in fact rather the while on treatment estimand. Our interest does not lie in precisely quantifying the risk of an AE of a certain type for a specific drug or in a specific therapeutic field. Our investigation aims at a methodological comparison of estimators of absolute AE risk in the presence of competing events. Given that currently little attention is given to competing events in the analysis of adverse events, we believe that our work is important as it might provide empirical evidence that the choice of estimator and handling of competing events can impact the interpretation of adverse event risks. We believe that this kind of evidence has the potential to change clinical trial practice with regard to safety analyses and risk–benefit assessments. So, our methodological investigation is relevant for the analysis of AE irrespective of the duration of AE recording. We agree that the recording of AEs after treatment discontinuation is important for a complete safety assessment and advice investigators to record safety data in the same way as efficacy data for the entire interesting follow-up period of patients. But, nevertheless, in the analysis of AEs in clinical trials, we are confronted with the situation that the follow-up period for AEs often ends prematurely due to, for example, progression and treatment discontinuation in oncology. This is further elaborated in Section 2.3. We add to the discussion by addressing the question if these treatment-related ends of follow-up should be considered as a competing event or as censored in Section 2.2.

The remainder of the paper is organized as follows. Section 2 discusses in detail the definition of AEs and of competing events and will also briefly consider composite events. The latter will be included to investigate the impact of ignoring censoring without the complication of competing events. Section 3 explains the organization of the data analyses within SAVVY and may serve as a template for future investigations of related questions. Here, an important aspect is that trial-level analyses will be run at sponsors’ sites, but meta-analyses will be run centrally by the academic project collaborators. The statistical methods on trial level are collected in Section 4. This section, in particular, explains how SAVVY will quantify and account for different lengths of follow-up and makes a connection between the different methods of estimation. For instance, the incidence proportion will equal the Aalen–Johansen estimator evaluated at the largest observed time in the absence of censoring. Details of the meta-analyses to be performed are in Section 5. This section, in particular, addresses the multitude of comparisons to be considered as well as assessment of bias and heterogeneity. An illustration of the methodology outlined in this paper is given in Section 6. A brief discussion is given in Section 7, also addressing the question of recurrent AEs, and software code is provided in the online supplement.

2 | DEFINITION OF EVENTS

2.1 | Adverse events

According to the Good Clinical Practice (GCP) guideline, an AE is defined as “any untoward medical occurrence in a patient or clinical investigation subject administered a pharmaceutical product and which does not necessarily have a causal relationship with this treatment” (Committee for Human Medicinal Products, 2016). In this meta-analytic study, the choice of AEs within the selected clinical trials is left to the sponsor. These may be defined as AEs of special interest, as belonging to a specific Medical Dictionary for Regulatory Activities (MedDRA) system organ class or preferred term, as being severe according to a toxicity grading, as being related to the investigational product, as serious, or as a combination of these characteristics.

Often sponsors will select as AEs the adverse drug reactions presented in the core data sheet for a first submission for drug approval, and select the studies that supported the frequency derivation.

These choices may result in a range of frequency, from common AEs to rare ones. It is expected that differences between the methodological approaches will be less marked with very rare AEs or very common AEs. So, grouping of different rare AE types into one AE category would be permissible for studying methodological differences. Making use of this frequency range (from rare to more common per selected trial) allows to further investigate the impact of the frequency on the differences between statistical analysis methods. Ideally, AEs of different frequencies per trial should be chosen, for example, around 90%, around 60%, around 30%, around 10%, and around 1%.

The investigation is restricted to the analysis of the occurrence of the first AE of a specific type and will not consider the analysis of recurrent events. However, the relevance of the present investigation for recurrent AEs will also be considered in Section 7.

2.2 | Competing events definition

Competing events are events that preclude the occurrence of the AE of interest in a time-to-first-event setting. For instance, if in a clinical trial the focus is on estimating the probability of headache, patients who die without prior headache report will never report headache. It is obvious that death is a competing event with respect to the occurrence of headache. However, defining a general rule as to which events should be treated as a competing event, without specific insight into the specific event of interest and trial at hand, is challenging.

As a rule of thumb, any event that both (a) would be viewed from a patient perspective as an event of his/her course of disease or treatment and (b) would stop the recording of the AE of interest should be viewed as a competing event. This situation would typically occur when a patient discontinues the treatment due to another AE judged by the investigator as too severe to continue treatment, or when the patient discontinues treatment or study due to progression/lack of efficacy, and, as a consequence, the recording of AEs ends. Hence, if end of follow-up for AEs, withdrawal of consent or discontinuation is disease or treatment related, this would be handled as a competing event. See Lawless and Cook, 2019, for related considerations.

In contrast to a competing event, the time-to-event is censored if the patient reaches the designated end of follow-up without having had the AE of interest or a competing event as defined above. This situation is present with administrative censoring due to the regular end of the trial or the end of follow-up for AEs due to the planned end of treatment (often end of treatment plus an additional fixed time interval, e.g., 30 days) not triggered by the course of disease.

In the analysis, the different competing events will be combined into one composite competing event, as the aim is to compare different methods to quantify the risk of an interesting AE and not the risk of a competing event of a specific type.

There is much debate among statisticians on the question which events should be analyzed as competing events and which events should lead to censoring the time-to-event. In practice, there is no discussion that death without the previous occurrence of the interesting AE acts as a competing event with respect to AE occurrence. The reason is that after death the AE can definitely not occur any more, and in this sense “death without prior AE” is a so-called “hard” competing event. However, the other events mentioned above as loss to follow-up, withdrawal of consent, or treatment discontinuation can be regarded as so-called “soft” competing events in the sense that, thereafter, the interesting AE in principle still could occur, but cannot be observed due to end of follow-up. Discussions therefore arise around the question whether to treat these soft competing events as censoring or as competing event. One extra concern here is that, for example, treatment

discontinuation will likely alter the AE hazard. See also Unkel et al., 2019, on how these aspects connect to the current debate on estimands as well as the following Subsection 2.3.

Therefore, two different approaches will be compared in this project. In a first approach (called “all-events approach” in the following), all competing events mentioned above will be combined and analyzed as a single competing event and only patients reaching the designated end-of-follow-up with neither former AE of interest nor competing event as defined above will contribute as censored observation. In a second approach (referred to as “death-only approach” in the following), only the hard competing event (i.e., death without prior AE of interest) will be analyzed as competing event, and the soft competing events will be analyzed as censored observations.

2.3 | Adverse events which occur after a competing event

The previous discussion in Section 2.2 on the definition of competing events has made the distinction between the “hard” competing event death without prior AE of interest and other competing events that, inter alia, lead to stopping the recording of AEs, after which, however, AEs can still occur. This discussion relates to the question which estimand is addressed.

As mentioned above, decisions about market authorization and reimbursement consider different aspects of treatment effects, leading to the situation that regulatory agencies in the marketing authorization process often address the while on treatment estimand, whereas HTA bodies are most interested in the treatment policy estimand which aims at comparisons of treatment groups with regard to AEs on the entire follow-up period irrespective of intercurrent events as treatment discontinuation (Unkel et al., 2019). We agree that the recording of AEs after treatment discontinuation is important for a complete safety assessment and advice investigators to record safety data in the same way as efficacy data for the entire interesting follow-up period of patients. It is not our aim to compare estimators without AE data after disease progression or treatment discontinuation. So, in the planned statistical analyses of the SAVVY project we compare the estimators for the treatment policy estimand on the time interval of overall survival whenever possible. We do not delete any AE after treatment discontinuation from the analysis. All AEs recorded in the included clinical trials will be analyzed. But we have to take the data as they were collected in the individual clinical trials included in our investigation. We are not able to analyze AEs which were not recorded. We therefore ask the investigators of the included clinical trials to indicate the time at which documentation of AEs ended due to disease-related events in order to be able to perform the correct statistical analyses. If documentation of AEs ended after treatment discontinuation, this analysis addresses the while on treatment estimand.

Our investigation aims at a methodological comparison of estimators of absolute AE risk which is complicated by the presence of competing events. We do not aim at precisely quantifying the risk of an AE of a certain type for a specific drug or in a specific therapeutic field. The latter would be restricted by occurrence of a competing event which, by definition in the previous subsection, leads to stop of AE recording, such that access to such data is not possible. However, our aim is methodological, we do not aim at a subject matter statement. The methodological investigation is relevant for the analysis of AEs irrespective of the duration of AE recording, as AEs are subject to competing events over the entire follow-up interval of overall survival.

2.4 | Composite events

Additionally to the statistical analyses of AEs considering competing events, further analyses encompassing both the AE of interest and the competing event as composite event will be performed, thereby addressing the composite estimand (Rufibach, 2019; Unkel et al., 2019). The time to the composite event will be defined as time to the interesting AE or to the competing event whatever occurs first, and patients with neither the interesting AE nor the competing event will contribute a censored time-to-event. The rationale for the inclusion of this approach is to gauge the impact of using time-to-event methodology to account for varying follow-up times without the methodological complication of competing events. To this end, time-to-event analyses accounting for censoring will be compared to the traditionally used incidence proportion.

We emphasize that our consideration of composites is solely motivated by methodological considerations on investigating the impact of (ignoring) censoring in the absence of competing events. In practice, use of any composite endpoint must consider whether or not combining different event types into one composite is clinically meaningful. As a consequence, also in this project only those composite endpoints will be considered, which will yield meaningful results

(Ferreira-González et al., 2017). For the situation of composite endpoints, we furthermore stress that for an understanding of the complete picture it is important to additionally analyze the separate components of the composite endpoints as it is done in our project.

3 | ORGANIZATION OF THE DATA ANALYSIS

The SAVVY project group consists of the academic project collaborators who planned the statistical analyses (in the following referred to as the “analysis center”) and the participating sponsors, who contribute randomized clinical trial data for analysis. In the SAVVY project, the data analysis involves the following steps:

1. *Preregistration*: Confidential preregistration of the clinical trials selected by the sponsors with the analysis center and allocation of a SAVVY trial identifier (ID) to registered trials by the analysis center
2. *Individual trial analysis*: Analysis of the registered trials at sponsor’s site using code provided by the analysis center and transfer of aggregated trial level results to the analysis center
3. *Meta-analysis*: Meta-analysis of trial level results at the analysis center

In the following, these steps will be considered one by one in more detail.

3.1 | Preregistration

The sponsors select the randomized clinical trials and the AEs they wish to enter into this project. In order to reduce the risk of selection bias, these trials have to be confidentially preregistered with the analysis center before running the analyses. For the identification of the trials, a unique trial ID according to a publicly accessible trial registry, for example, clinicaltrials.gov or the German Clinical Trials Register, has to be provided together with some characteristics of the trial and the selected AEs (see Table 1). The identification of the trials included will not be disclosed otherwise. In publications or presentations, it will only be reported that studies have been identified to the analysis center. The analysis center will handle any information related to trial-level data in a confidential manner. Also, the exchange of trial-related information between the sponsor and the analysis center will take place in a secured manner following the sponsors’ individual policies regarding the secure exchange of confidential information.

To register the trials and the AEs for the SAVVY project, a spreadsheet is filled in by the sponsor containing one row per AE of interest with the main AE characteristics, for example, seriousness, severity, MedDRA system organ class (SOC), MedDRA preferred term (PT). If a sponsor does not want to provide a particular information or if the information is not relevant due to the particular grouping applied, “NA” for “not applicable” can be used. Table 1 shows the characteristics of the selected trials and AEs that will be captured.

After receipt of the registration sheet, a SAVVY trial ID will be allocated to the trial by the analysis center. The SAVVY ID will be entered into the trial characteristics spreadsheet and returned to the sponsor.

3.2 | Individual trial analysis

The analyses of the individual clinical trials will be done at sponsor’s site using SAS or R code provided by the analysis center. Therefore, it is not required to release any individual patient data to the analysis center. Only aggregated data, summarizing the results of the analyses, will be shared with the analysis center. The analysis center will not share the aggregated data of one sponsor with any other sponsor. A manual is provided by the analysis center to the sponsor describing what needs to be done after receipt of the SAVVY trial ID and the program code. As a prerequisite, at the sponsors’ sites, the individual clinical trials datasets must be brought into a format which allows the application of the provided SAS or R code. The required data structure is simple, similar to that of a standard survival analysis, and shown in Table 2.

For each trial one dataset is needed in which all AE-specific data for the selected AEs are set below each other. The different AEs are distinguished by the AE ID, matching the AE ID given in the trial characteristics table filled for trial preregistration. The treatment group ID will be used by the SAS or R code but is not included in the results where

TABLE 1 Characteristics of the selected trials and AEs captured during pre-registration

Trial / AE characteristic	Explanation, possible entries
Unique trial ID	Clinical trials registry number according to a publicly accessible trial registry, for example, clinicaltrials.gov, German Clinical Trials Register
Indication	Indication/therapeutic area investigated in the trial
Type of comparison	Active or placebo controlled
End of the trial	Year of last patient / last visit of the trial
Maximum follow-up time for primary efficacy endpoint	Maximum follow-up time of the patients for recording the primary efficacy endpoint (in days)
AE ID	AE identifier, incremental AE numbering (from 1 to total number of selected AEs) per trial
Maximum follow-up time for AE	Maximum follow-up time of the patients for recording the AE (in days)
Seriousness of AE	Specify if serious AE, any AE or NA
Severity of AE	Specify CTCAE toxicity grade (e.g., ≥ 3), any toxicity grade or NA
MedDRA system organ class (SOC) of AE	Specify SOC(s), any SOC or NA
MedDRA preferred term (PT) of AE	Specify PT(s), any PT or NA
Special interest AE	Specify special interest AE, any AE or NA
Hard competing events	Specify the type(s) of the hard competing event(s) as defined in Section 2.2, for example, death
Soft competing events	Specify the type(s) of the soft competing event(s) as defined in Section 2.2, for example, end of treatment, withdrawal of consent, progression

TABLE 2 Required data structure for application of program code

Column	Description
AE ID	Number from 1 up to the number of selected AEs within the trial
Patient ID	Trial-specific unique patient ID
Treatment group ID	Identifying which treatment is experimental and which is control
Time to event	In days
Type of event	1 = AE of interest 2 = hard competing event 3 = soft competing event 0 = censored

experimental and control groups are coded as “A” and “B”, respectively. Observations with missing data, negative event times, or type of event not in {0, 1, 2, 3} are automatically excluded from the analyses.

The SAVVY trial ID is inserted in the SAS or R code. The program returns the aggregated data summarizing the results of the statistical analysis methods described in Section 4 and the results of some further descriptive analyses on the AE and the competing event, namely, mean, median, minimum and maximum event time by type of event and overall, and the total numbers of AEs, competing events, and censored observations, each overall and per treatment group. The dataset containing all results is named automatically identical to the SAVVY trial ID and is sent to the analysis center for further processing.

3.3 | Meta-analysis

Once the results of all registered trials are received, the analysis center performs the meta-analyses described in Section 5 of the estimated parameters described in Section 4. The results of the meta-analyses will be presented and discussed within the SAVVY project group without identifying individual trials and sponsors.

4 | STATISTICAL METHODS ON TRIAL LEVEL

4.1 | Quantifying length of follow-up

We will use two approaches to both quantify length of follow-up and to study its impact on analyzing the occurrence of AEs. The first approach is guided by the implicit choice made when calculating incidence proportions. The second approach is guided by the concern that overly small risk sets late in time may lead to unstable probability estimators (Pocock, Clayton, & Altman, 2002).

The first approach is to choose a time τ_A as the largest observed time (censored or AE or competing event) which was (if observed AE) or could have been (if censored or competing event) an observed AE time in group A. Time point τ_B is defined analogously for group B. Then one step to account for different lengths of follow-up between groups is to restrict statistical inference to the smaller of the two time points. Hence, let $\tau = \min(\tau_A, \tau_B)$. The motivation behind this approach is twofold: First, the commonly used incidence proportions are calculated in the complete dataset, that is, for the data available on $[0, \tau_A]$ and $[0, \tau_B]$, respectively. Second, group comparisons for time-to-event data are typically restricted to the smaller of these two time intervals only. For instance, the common log-rank test only compares groups as long as the risk sets are non-empty in both groups.

The second approach follows a suggestion of Pocock et al. (2002). It covers a range of choices for quantifying length of follow-up and will depend on the proportion of patients still at risk. So, additionally, choose $\tilde{\tau}_A(p) = \tilde{\tau}_A$ as the time such that $100 \cdot p\%$ of all patients in group A are still at risk just prior time $\tilde{\tau}_A$ and may have an observed event at time $\tilde{\tau}_A$, $p \in [0, 1]$. To be precise, let $\tilde{\tau}_A$ be the $100 \cdot p\%$ quantile of the usual empirical distribution function \hat{F} of the observed times (irrespective of censoring status),

$$\tilde{\tau}_A = \inf\{t : \hat{F}(t) \geq p\}.$$

Choose $\tilde{\tau}_B$ analogously in group B, and let $\tilde{\tau} = \min(\tilde{\tau}_A, \tilde{\tau}_B)$. We will consider $p \in \{0.3, 0.6, 0.9\}$. The relationship between $\tilde{\tau}_A$ and τ_A is that both time points coincide for the choice of $p = 1$.

We also note that the different choices of τ account for different evaluation times underlying the estimation methods within groups, but they do not account for differential dropout between groups. It is, for example, possible that $\tau_A = \tau_B$, but that dropout rates differ between groups. Such differential dropout would therefore be potentially treatment group related, suggesting to handle dropouts as competing risks (see Section 2.2).

4.2 | One-sample estimators

Methods are exemplarily discussed for group A and for τ . The estimators of “absolute AE risk” that we will consider fall into three groups (Allignol et al., 2016). First, the incidence proportion accounts for competing risks but not for censoring (Equation 1 below). Second, one minus the Kaplan–Meier estimator accounts for censoring but not for competing risks (Equation 4), and this is also true for a standard conversion of the incidence density to a probability (Equation 3). Third, the Aalen–Johansen estimator generalizes the Kaplan–Meier estimator to competing risks and will later serve as a benchmark or method of choice for non-parametric estimation of the cumulative AE probability. The connection to the incidence proportion is that both coincide in the absence of censoring. The connection to one minus the Kaplan–Meier estimator is that both coincide in the absence of competing risks. A parametric counterpart (Equation 7) of the Aalen–Johansen estimator (Equation 8) based on incidence densities is also considered. All of these estimators of “absolute AE risk” are probability estimators, which is in contrast to incidence densities, which estimate hazards under a constant hazards assumption. However, the transformations that we use to transform the latter onto the probability scale mirror those underlying the non-parametric estimators Aalen–Johansen and Kaplan–Meier; see Appendix A.2.4 in Aalen, Borgan, and Gjessing (2008) for a textbook account. Also note that the Aalen–Johansen estimator of the cumulative event probability of a competing event is often called “cumulative incidence function,” referring to a probability and, in plain language, equating estimator and estimated parameter.

To begin, the incidence proportion divides the number of patients with an observed AE on $[0, \tau]$ in group A by the number n_A of patients in group A. More precisely, introduce individual first-AE-counting processes

$$N_i(t) \in \{0, 1\}, i \in \{1, \dots, n_A\}, t \in [0, \tau], N_i(0) = 0,$$

where $N_i(t) = 1$ denotes that an AE has been observed for patient i in the time interval $[0, t]$ and that no competing event has been observed before the AE. Analogously, let

$$\bar{N}_i(t) \in \{0, 1\}, i \in \{1, \dots, n_A\}, t \in [0, \tau], \bar{N}_i(0) = 0,$$

denote i 's counting process of observed competing events. Because we consider time-to-first-event and type-of-first-event, we have that

$$N_i(t) + \bar{N}_i(t) \leq 1,$$

and both $N_i(t)$ and $\bar{N}_i(t)$ change their value from 0 to 1 at most once, when a time-to-first-event has been observed. The aggregated processes are

$$N_A(t) = \sum_{i=1}^{n_A} N_i(t), \bar{N}_A(t) = \sum_{i=1}^{n_A} \bar{N}_i(t).$$

In the absence of censoring, the sum of the two aggregated processes will eventually be equal to n_A , but in general we have $N_A(\infty) + \bar{N}_A(\infty) \leq n_A$. The incidence proportion now is

$$IP_A(\tau) = \frac{N_A(\tau)}{n_A}. \quad (1)$$

The incidence proportion is an estimator of the probability that an AE happens in the interval $[0, \tau]$, and that this AE is observed. The incidence density has the same numerator, but divides by person-time-at-risk. Again, to be precise, introduce individual at-risk-processes

$$Y_i(t) \in \{0, 1\}, i \in \{1, \dots, n_A\}, t \in [0, \tau], Y_i(0) = 1,$$

where for $t > 0$, $Y_i(t) = 1$ denotes that the patient is still under observation on $[0, t)$ and that neither an AE nor a competing event have happened on $[0, t)$. Note that the at-risk-processes are left-continuous, such that $Y_i(t)$ denotes the at-risk status just prior time t . If $Y_i(t) = 1$, an event may happen and be observed at time t . Otherwise, $Y_i(t) = 0$. The incidence density now is

$$ID_A(\tau) = \frac{N_A(\tau)}{\int_0^\tau \sum_{i=1}^{n_A} Y_i(u) du}. \quad (2)$$

The incidence density is not a probability but estimates the AE hazard with values in $[0, \infty)$ under a constant hazard assumption. A typical transformation of this estimator onto the probability scale is

$$1 - \exp(-ID_A(\tau) \cdot \tau). \quad (3)$$

In the following, we will call this quantity probability transform of the incidence density ignoring competing events, or, simply, probability transform of the incidence density. Assuming a constant AE hazard, estimator (3) estimates the same quantity as one minus the Kaplan–Meier estimator, which only codes observed AEs as an event and censors anything else. To be precise, introduce increments

$$\Delta N_i(t) = N_i(t) - \lim_{u \nearrow t} N_i(u),$$

which equal one, if an AE (before any competing event) is observed for patient i at time t , and $\Delta N_i(t) = 0$ otherwise. Defining $\Delta \bar{N}_i(t)$ analogously, the increments of the aggregated processes are

$$\Delta N_A(t) = \sum_{i=1}^{n_A} \Delta N_i(t), \Delta \bar{N}_A(t) = \sum_{i=1}^{n_A} \Delta \bar{N}_i(t).$$

The size of the risk set is

$$Y_A(t) = \sum_{i=1}^{n_A} Y_i(t),$$

and one minus the Kaplan–Meier estimator which only codes observed AEs as an event and censors anything else can be expressed as

$$1 - \hat{S}_A(\tau) = 1 - \prod_{u \in (0, \tau]} \left(1 - \frac{\Delta N_A(u)}{Y_A(u)} \right) = 1 - \prod_{u \in (0, \tau]} (1 - \Delta \hat{\Lambda}_A(u)). \quad (4)$$

Here, $\Delta \hat{\Lambda}_A(u)$ is the increment of the non-parametric Nelson–Aalen estimator of the cumulative AE hazard

$$\hat{\Lambda}_A(\tau) = \sum_{u \in (0, \tau]} \frac{\Delta N_A(u)}{Y_A(u)}, \quad (5)$$

where the product in Equation 4 and the sum in Equation 5 is over all observed, unique event times u . Also note that we are slightly abusing notation in Equation 4, because $\hat{S}_A(\tau)$ is not estimating a proper survival function because of the presence of competing risks. The Nelson–Aalen estimator, however, is a proper estimator of the cumulative AE hazard. Assuming a constant AE hazard, $\hat{\Lambda}_A(\tau)$ and $ID_A(\tau) \cdot \tau$ estimate the same quantity.

Accounting for competing risks now requires to acknowledge that there also is a competing hazard. To begin, we introduce a competing incidence density

$$\overline{ID}_A(\tau) = \frac{\tilde{N}_A(\tau)}{\int_0^\tau \sum_{i=1}^{n_A} Y_i(u) du}. \quad (6)$$

Also using $ID_A(\tau)$ as defined above, we obtain an estimator of the cumulative AE probability based on incidence densities and accounting for competing risks,

$$\hat{p}_{ID;A}(\tau) = \frac{ID_A(\tau)}{ID_A(\tau) + \overline{ID}_A(\tau)} \left(1 - \exp(-\tau \cdot [ID_A(\tau) + \overline{ID}_A(\tau)]) \right). \quad (7)$$

This estimator will be called probability transform of the incidence density accounting for competing events in the remainder of this paper. The connection of this estimator to the incidence proportion is that the leading factor on the right-hand side of the previous display equals $IP_A(\tau)$ in the absence of censoring and if $\tau = \tau_A$ (Beyersmann & Schrade, 2017). In words, both of these quantities estimate the anytime-AE-probability in this situation. In the presence of censoring, quantity (7) estimates the cumulative AE probability assuming all hazards to be constant. The non-parametric counterpart is the Aalen–Johansen estimator of the so-called cumulative incidence function, that is, the probability of an AE occurring until τ ,

$$CIF_A(\tau) = \sum_{u \in (0, \tau]} \prod_{v \in (0, u)} \left(1 - \Delta \hat{\Lambda}_A(v) - \Delta \hat{\Lambda}_A^*(v) \right) \Delta \hat{\Lambda}_A(u), \quad (8)$$

where $\Delta \hat{\Lambda}_A^*(v)$ now is the increment of the competing Nelson–Aalen estimator in analogy to \overline{ID}_A . Note that we have again slightly abused notation, writing $CIF_A(\tau)$, although this quantity is an estimator.

4.3 | Two-sample comparisons

In principle, many methods of two-sample comparisons are conceivable. We here aim to consider one method that applies to all one-sample estimators and provides a quantification of risk differences and relative risks, where *risk* here refers to a probability estimator as defined earlier. So for each of the estimation methods discussed in Section 4.2 and for each

evaluation time point defined in Section 4.1, we estimate the risk difference with corresponding 95% confidence interval and the relative risk whose 95% confidence interval can be derived with the help of a log-transformation and the delta-method and is given by

$$\hat{RR} \cdot \exp(\pm z_{0.975} \cdot \hat{\sigma}), \quad (9)$$

where $\text{var}(\log \hat{q}_A - \log \hat{q}_B) = \hat{\sigma}_A^2 + \hat{\sigma}_B^2 = \hat{\sigma}^2$. Thereby, \hat{q}_A and \hat{q}_B denote the estimators of Section 4.2 and $\hat{\sigma}_A^2$ and $\hat{\sigma}_B^2$ their corresponding variance estimates. The primary comparison of methods will be based on probability estimators as explained. However, because of the omnipresence of hazard ratios for group comparisons, we will also consider comparisons on the hazard scale as detailed below:

1. An estimated hazard ratio (output from standard Cox software) only coding AEs as “observed event,” together with an estimator of its variance and a confidence interval for the hazard ratio.
2. Ditto for the competing event, now only coding competing events as “observed event.” One rationale here is to check whether relevant signals on the hazard scale would have been missed by ignoring competing risks. We reiterate that, as for all competing event analyses, this will be done with the hard and the soft definition given in Section 2.2.
3. Ratios of incidence densities for AEs, with variance estimation analogous to above. Since under the assumption of constant hazard the incidence density is an estimator of the hazard, the rationale here is to check whether the simple constant hazard framework, although potentially misspecified, leads to a reasonable approximation of the hazard ratio estimated in semiparametric fashion.
4. Ratios of incidence densities for competing events.
5. Ratios of Nelson–Aalen estimators for AEs, with variance estimation analogous to above. The rationale here is to compare the usual hazard ratio estimator not only with a very simple parametric counterpart, but also with a fully non-parametric competitor. Under a proportional hazards assumption, the ratio of Nelson–Aalen estimators also estimates the hazard ratio, but not under non-proportional hazards (Andersen, 1983).
6. Ditto for the competing event.

4.4 | Assessment of differences of estimators

The estimators in Section 4.2 and the derived information on estimated risk differences and risk ratios are of standard input form for a meta-analysis. However, the aim is a methodological comparison of different methods for quantifying AE risk when applied to the very same data. To this end, the information on variances so far does not suffice, but what we need is an estimator of the variance between, for example, the incidence proportion and the Aalen–Johansen estimator when calculated on the same dataset. As is obvious from the formulae in Subsection 4.2, such estimators will in general be dependent. Closed form variance estimators might be obtained using the *functional* delta method (Gill & Johansen, 1990), but one would need to derive and implement such estimators for every single methodological comparison. We have therefore decided to follow the advice of Andersen, Borgan, Gill, and Keiding (1993) and to resort to bootstrap variances, drawing with replacement from the individual patients under an *i.i.d.* set-up. In addition, Hjort (1992) notes that bootstrapping yields valid estimators of the variance even if a (semi-) parametric model such as constant hazards or proportional hazards is misspecified. We also note that in a meta-analysis of published data on the overestimation of the cumulative revision arthroplasty using a Kaplan–Meier-type estimator, Lacny et al. (2015) used common approximations of the estimated variance of the hazard ratio for this purpose. However, this approach estimates a different variance as the correlation structure is not accounted for.

A mathematical justification for bootstrapping is obtained by obtaining the difference (or the ratio) of estimators via a composition of simpler building steps all of which are Hadamard differentiable. For instance, Gill and Johansen (1990) outline this program for the Kaplan–Meier estimator. The analogous (de-) composition for the Aalen–Johansen estimator would, for example, allow to “build” both the Aalen–Johansen and the Kaplan–Meier estimator and to subsequently consider their difference or ratio. Finally, an empirical process result, for example, Theorem 23.9 in van der Vaart (1998), justifies use of the bootstrap. The main practical argument for using the bootstrap in a meta-analytical study as the present one is to avoid additional programming burden which might be error-prone.

TABLE 3 Planned comparisons. The quantities marked with ★ are calculated in both groups

Target quantity	Gold-standard	Compared against
AE probability ★	Aalen–Johansen (Equation 8)	Incidence proportion (Equation 1)
	Aalen–Johansen (Equation 8)	Probability transform incidence density (Equation 3)
	Aalen–Johansen (Equation 8)	1-Kaplan–Meier (Equation 4)
	Aalen–Johansen (Equation 8)	Probability transform incidence. Density accounting for competing events (Equation 7)
	Aalen–Johansen (Equation 8)	Aalen–Johansen only treating death as competing event (see Section 2.2)
Composite endpoint ★	1-Kaplan–Meier	Incidence proportion
Hazard ratio	Cox	Ratio incidence densities
	Cox	Ratio Nelson–Aalen estimators

4.5 | Implementation

The estimators displayed in the Sections 4.2 and 4.3 are readily available in the statistical software SAS (SAS Institute, Cary, NC, USA) and R (R Core Team, 2018). The implementation of the estimators of the incidence proportion and the two estimators based on the incidence density is straightforward by the use of the formulae. In SAS software, the `proc lifetest` calculates the one minus Kaplan–Meier estimator. In R it can either be obtained by the `survfit` function of the `survival` package (Therneau & Grambsch, 2000) or, as it is a special case of a competing risk setting, the `etm` function of the identically named package (Allignol, Schumacher, & Beyersmann, 2011) can also be used to calculate both one minus the Kaplan–Meier estimator and the Aalen–Johansen estimator. Depending on which SAS version is used, the Aalen–Johansen estimator can, on the one hand, be computed by the predefined %CIF Macro. On the other hand, in newer versions of SAS software the `proc lifetest` specifying the event of interest in the option `failcode` can be used.

The first part of the two-sample comparisons, the risk differences and relative risks with corresponding variances, can be directly calculated by implementing the formulae. The Cox model and therefore the estimated hazard ratio may be obtained by the use of the `proc phreg` in SAS and with the function `coxph` in R. In order to estimate (event-specific) hazard ratios, for example, for AE, a well-known coding method is to also code observed competing events as “censored.” A brief look at the simple incidence densities illustrates correctness of this method for analyzing hazards. The estimator of the cumulative AE probability in (7) demonstrates that all hazards enter probability calculations, and hence the “code as censored” approach is only available on the hazard level. In both of the software, it can be easily switched which event is of interest, such that the hazard analysis of the competing event can as easily be obtained. The ratios of the incidence densities for the adverse as well as for the competing event are easily calculated once the incidence density for the one-sample estimators has been saved. The `proc lifetest` with the `nelson` option gives the Nelson–Aalen estimator. Moreover, the `mvna` function of the `mvna` package (Allignol, Beyersmann, & Schumacher, 2008) returns this estimator in R.

SAS macro code and the corresponding function in R are available as supplementary material. The main macro code has been written in SAS 9.4 software and checked in R by one of the authors (RS). It has subsequently been checked in a small-scale pilot study (VJ, KR, CS).

5 | META-ANALYSIS

Once the trial-level data have been analyzed using the methods described in Section 4, results will be summarized across trials using the approaches listed below. Whereas individual trial data analyses will be run within the sponsor company, meta-analyses will be run on the calculated probability and hazard (ratio) estimates centrally at the analysis center, that is the institutions of the academic project group members.

5.1 | Method comparisons

Table 3 gives an overview of the planned method comparisons. We will distinguish between the two types of the competing events introduced in Section 2.2, that is, all comparisons will be performed for both types of competing events (death-only

and all-events). The benchmark or gold-standard estimator here is the Aalen–Johansen estimator, because it accounts for both censoring and competing events. Thereby, the main interest is in the “all-events” competing event. Especially, the comparison of Aalen–Johansen estimators based on the different competing event definitions is of interest. The two Aalen–Johansen estimators will be compared for the AE as well as for the competing event. Moreover, the comparisons in terms of hazard ratios will be conducted for the AE and for the competing event. All comparisons are conducted at the five follow-up times, that is, at the p -quantiles ($p = 0.3, 0.6, 0.9, 1$) accounting for different follow-up times in both groups and at the maximum follow-up time not accounting for differences in follow-up between the groups, as defined in Section 4.1.

5.2 | Assessment of bias

The bias of the estimators with respect to the gold-standard estimator will be assessed by comparison of the estimators with the gold-standard estimator. We call such a difference “bias,” although, of course, the comparison is between estimators. We note however that, say, a comparison of one minus Kaplan–Meier and Aalen–Johansen will converge in probability towards the true or asymptotic bias. Bias will be assessed visually using Bland–Altman plots of the AE probability, the risk difference, and the (log) relative risks (Altman & Bland, 1983). As we consider a comparison of an estimator to a benchmark, the gold-standard estimator is plotted on the x-axis instead of the mean (Krouwer, 2008). With these plots, both the one-sample (AE probability) as well as the two-sample (risk difference, relative risk) situation can be considered.

5.3 | Frequency categories

For the one-sample estimators, the possible change in frequency categories depending on estimation method will also be investigated. According to the European Commission’s guideline on summary of product characteristics (SmPC) (EMA, 2009) the frequency categories are, respectively, classified as “very rare,” “rare,” “uncommon,” “common,” and “very common” when found to be $< 0.01\%$, $< 0.1\%$, $< 1\%$, $< 10\%$, $\geq 10\%$. Frequency categories obtained with the estimators will be compared to frequency categories obtained with the gold-standard estimator, that is, the Aalen–Johansen estimator.

The comparison of the conclusions about the therapies’ safety derived from the two-sample comparisons of the various approaches shall be compared in terms of statistical significance, clinical relevance and benefit assessment criteria (IQWiG, 2017; Kieser & Hauschke, 2005) against the Aalen–Johansen approach as benchmark in frequency tables.

5.4 | Assessment of precision

As assessment of precision, the standard errors or the width of the confidence intervals of the estimators will be compared to the gold-standard ones. This is done in terms of plots of the ratios of the standard errors for the methods with at most small to moderate bias. The consideration of precision is deemed useful only in the absence of any substantial bias.

5.5 | Random effects meta-analysis and meta-regression

A more formal assessment of difference between estimators and possible factors influencing these is carried out in form of random effects meta-analyses and meta-regressions. These will model the ratios of the estimators considered in Section 5.1 (i.e., respective estimator divided by benchmark). The standard errors of these ratios will be needed for the meta-analysis. As noted in Section 4.4, the derivation of these standard errors is complicated by the dependence of the estimators. Therefore, they will be obtained with a bootstrap (see Section 4.4).

To be more precise, the estimator of the log-ratio ($\log(\text{estimator}/\text{benchmark})$) $\hat{\theta}_k$ is observed with bootstrapped variance $\hat{\sigma}_k^2$ for each adverse event $k = 1, \dots, K$. Then a normal–normal hierarchical model (NNHM) (Hedges & Olkin, 2014) of the form $\hat{\theta}_k | \theta_k \sim N(\theta_k, \sigma_k^2)$ with $\theta_k | \theta, \rho \sim N(\theta, \rho^2)$, $k = 1, \dots, K$, is fitted, with ρ^2 denoting the between AE heterogeneity. Thereby, between adverse events variability is introduced via $\theta_k = \theta + \epsilon_k$ with $\epsilon_k \sim N(0, \rho^2)$. As the main interest is in the mean parameter θ , the marginal model $\hat{\theta}_k | \theta \sim N(\theta, \sigma_k^2 + \rho^2)$, $k = 1, \dots, K$, will be used.

TABLE 4 Parameter used to simulate eight trials each contributing one AE

Study	n_A	n_B	$\lambda_A(t)$	$\tilde{\lambda}_A^{\text{death}}(t)$	$\tilde{\lambda}_A^{\text{soft}}(t)$	$\lambda_B(t)$	$\tilde{\lambda}_B^{\text{death}}(t)$	$\tilde{\lambda}_B^{\text{soft}}(t)$	Censoring
S1	200	200	0.00265	0.00151	0.00227	0.00245	0.00207	0.00322	$U[0, 1000]$
S2	200	200	0.00124	0.00193	0.00249	0.00112	0.00268	0.00373	$U[800, 1000]$
S3	400	400	$(1/2)t$	$(2/25)t$	$0.8/(t+2)$	$(2/9)t$	$(2/9)t$	$0.8/(t+2)$	$U[0.5, 5]$
S4	500	500	$4/(t+2)$	$(1/2)t$	0.5	$4/(t+2)$	$(1/2)t$	0.5	$U[0.15, 4]$
S5	1000	1000	$2/(t+2)$	0.0037	$(1/2)t$	$(1/2)t$	$(2/25)t$	$0.8/(t+2)$	$U[1, 5]$
S6	150	100	$(3/8)t^2$	$(1/8)t$	$0.8/(t+2)$	$(2/25)t$	$(3/125)t^2$	$0.8/(t+2)$	$U[0.5, 5]$
S7	825	800	$1/(t+4)$	$0.2/(t+3)$	$(1/8)t$	$1/(t+5)$	$0.2/(t+3)$	$(2/25)t$	$U[0.2, 4]$
S8	700	700	$2/(t+2)$	0.0013	$(1/9)t^2$	$2/(t+2)$	0.0043	$(2/9)t$	$U[2, 3]$

As we are also interested in exploring any heterogeneity identified, possible sources of heterogeneity are assessed using meta-regression models including, for example, the proportion of censoring, of adverse event or the competing event recording over time. The model of the meta-regression is of the form $\hat{\theta}_k = \theta + \beta_1 x_{1k} + \dots + \beta_p x_{pk} + \epsilon_k$, where x_{ik} denote the i th of p possible source of heterogeneity of the k th AE. Thereby, β_1, \dots, β_p are fixed effects and ϵ_k is random.

The random effects model will be fitted with a two-step approach. First, the between adverse event variability ρ^2 is estimated by the Paule–Mandel estimator as recommended by Veroniki et al. (2016). Since we expect a large number of trials and types of AEs Wald-type confidence intervals with normal quantiles will be used for inference rather than t -quantiles as for example in the Knapp–Hartung–Sidik–Jonkman approach. For smaller numbers, that is fewer than 20, the Knapp–Hartung–Sidik–Jonkman approach is of course preferable and would be used (Veroniki et al., 2019). As the second step, θ or β_1, \dots, β_p are estimated with the help of weighted least squares with weights equal to $w_i = 1/(\sigma_k^2 + \rho^2)$. The meta-analysis and meta-regression will be conducted in R by the use of the function `rma()` of the `metafor`-package (Viechtbauer, 2010).

The meta-analysis and meta-regression will first be performed on the AE level and not on the trial level, that is, AEs of the same trial will be assumed to be independent. In a next step, as the structure of these data are more complex than in standard meta-analyses, potentially additional hierarchy levels will be considered. In the NNHM described above, it is assumed that it is sufficient to model the between AE heterogeneity. However, it might be necessary to consider in addition, for instance, any heterogeneity between studies or indications. Therefore, random effects not only for AE but also for trial or indication are considered in subsequent analyses to explore whether additional hierarchy levels improve model fit.

6 | EXAMPLE

The aim of this section is to emphasize the motivation of the SAVVY project to show what kind of results can be expected, to demonstrate the relevance of our investigation, and to illustrate the comparison approaches of Section 5. For this, we simulated eight datasets representing trials each contributing one AE. Table 4 displays the simulation parameters, and Table 5 gives an overview over the datasets. The simulated datasets cover common and very common AEs and different constellations of soft and hard (death) competing events.

6.1 | Example of the estimation of the AE probability

At first, we only consider one of the eight trials to display the differences in the results of the AE probability estimators introduced in Section 4.2 for one AE. We chose trial S4.

Figure 1 displays the calculated AE probability using the estimators of Section 4.2. Furthermore, for the Aalen–Johansen estimator both definitions of a competing event are considered whereas for the probability transform of the incidence density accounting for competing events only the all events definition is used. The Aalen–Johansen estimator only considering death as a competing event is called Aalen–Johansen (death only) in the following.

At the maximum follow-up time (no common tau), all of the estimates obtained with the estimators of Section 4.2 are different. The greatest value is obtained for the one minus Kaplan–Meier estimator. In group B, it is one as the last event

TABLE 5 Overview over the simulated datasets. Gold-standard Aalen–Johansen estimator (AJE) and event numbers (no.) in both groups at the end of follow-up. CE stands for competing events

Study	AJE in A	AJE in B	no. cens in A	no. AE in A	no. death in A	no soft CE in A	no. cens in B	no. AE in B	no. death in B	no. soft CE in B
S1	0.4242	0.4195	28	74	67	31	29	72	76	23
S2	0.3950	0.4700	2	79	89	30	0	94	72	34
S3	0.4011	0.5880	60	155	109	76	61	218	103	18
S4	0.3342	0.4453	62	96	228	114	71	116	193	120
S5	0.1632	0.3508	56	140	318	486	100	347	307	246
S6	0.4172	0.6674	24	60	39	27	26	57	16	1
S7	0.2667	0.2468	386	97	205	137	427	96	147	130
S8	0.0739	0.0593	81	39	106	474	73	35	168	424

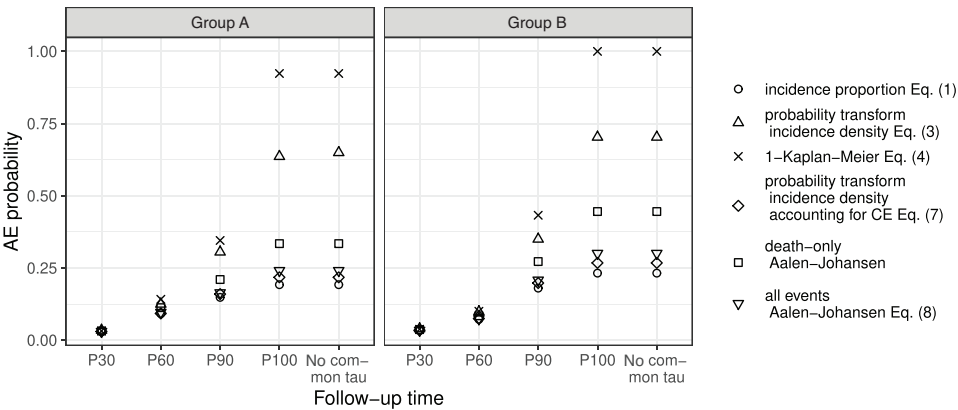


FIGURE 1 Calculated AE probability using the six different estimators for the AE of study S4. CE stands for competing events

in group B is an AE. The probability transform of the incidence density and the Aalen–Johansen (death only) estimator are also greater than the estimate of the gold-standard Aalen–Johansen estimator. The incidence proportion and the probability transform of the incidence density accounting for competing events are smaller than the gold standard. At earlier follow-up times, the differences between the estimators are smaller as usually at earlier follow-up times there are less observations censored and less competing events. As an example for the group comparison, we consider the relative risks with corresponding 95% confidence interval of the estimators displayed in Figure 1 at the maximum follow-up time accounting for the same length of follow-up in both groups called “P100.” Table 6 displays the results.

The relative risk of the incidence proportion, of the probability transform of the incidence density accounting for competing events and of the all events Aalen–Johansen estimator are similar. The probability transform of the incidence density and the one minus Kaplan–Meier estimator estimate a greater relative risk than the all-events Aalen–Johansen estimator. A relative risk smaller than one is calculated using the Aalen–Johansen estimator that treats only death as a competing event. For this Aalen–Johansen (death only) estimator, the corresponding 95% confidence interval does not

TABLE 6 Relative risks (RR; treatment divided by control) and 95% confidence interval (95% CI) of the AE probability at the maximum follow-up time accounting for the same follow-up time in both groups (P100). CE stands for competing events

Estimator	RR	95% CI
Incidence proportion	0.8276	[0.6508, 1.0524]
Probability transform incidence density	0.9050	[0.7800, 1.0501]
1-Kaplan–Meier	0.9234	[0.8054, 1.0587]
Probability transform incidence density accounting for CE	0.8134	[0.6428, 1.0294]
Death-only Aalen–Johansen	0.7505	[0.6064, 0.9288]
All events Aalen–Johansen	0.8019	[0.6383, 1.0074]

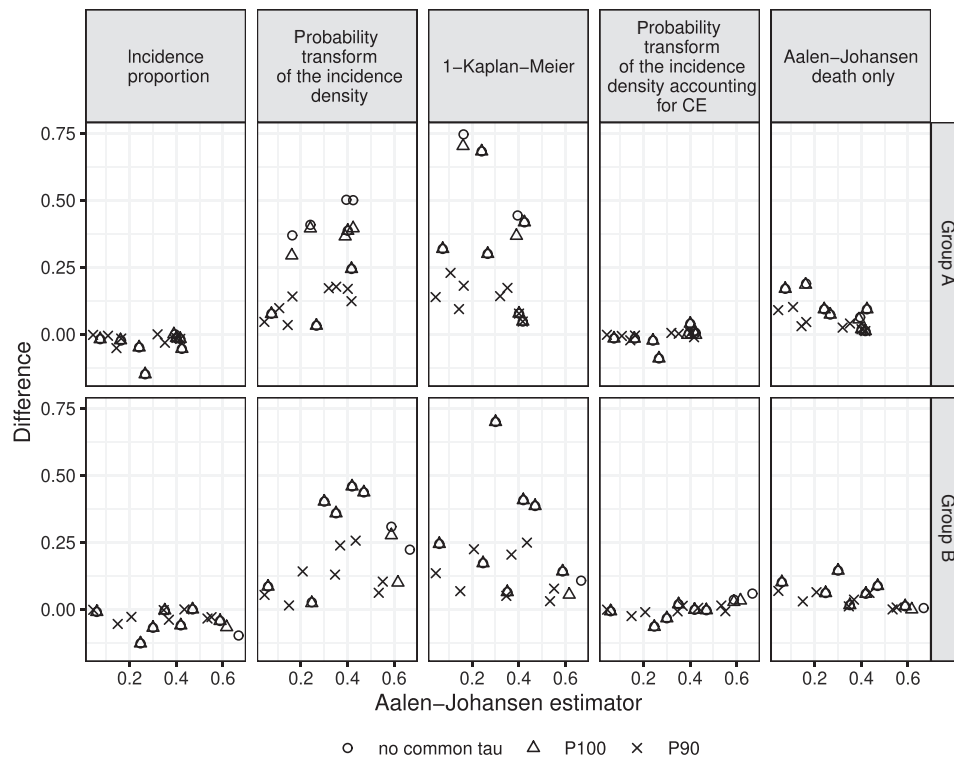


FIGURE 2 Bland-Altman plots of the AE probability in the group A and B at the three late follow-up times called “no common tau,” “P100,” and “P90.” CE stands for competing events

include the one whereas all other confidence intervals do. So using the Aalen-Johansen (death only) estimator instead of the gold-standard Aalen-Johansen estimator leads to a difference conclusion about a therapy’s safety.

6.2 | Example of the meta-analysis

Figure 2 displays an example for Bland-Altman plots of the AE probability for the eight AEs. The three different follow-up time points are displayed as different symbols. The incidence proportion is often smaller than the gold-standard Aalen-Johansen estimator as the incidence proportion does not account for censoring. The probability transform of the incidence density accounting for competing events is about equal the gold-standard Aalen-Johansen estimator, but in some scenarios it is greater whereas in others it is smaller than the gold standard. The other three estimators are always greater than the gold standard.

Table 7 displays the frequency categories obtained with the incidence proportion or the one minus Kaplan-Meier estimator compared to the all events Aalen-Johansen estimator as the gold standard.

The incidence proportion and the gold-standard Aalen-Johansen estimator always obtain the same frequency category for the eight AEs. Using the one minus Kaplan-Meier estimator one AE (S8) is categorized to “very common” instead to “common.”

Figure 3 shows an exemplary display of the assessment of precision. For all of the eight AEs, the standard error of the Aalen-Johansen estimator is smaller than the standard error of the one minus Kaplan-Meier estimator. A major reason for this is that one minus Kaplan-Meier equals zero at the largest observed time if and only if that time corresponds to an AE. Otherwise, one minus Kaplan-Meier is larger than zero. This is in contrast to the Aalen-Johansen estimator whose plateau approaches the cumulative anytime AE probability.

The meta-analysis considers the log-ratio, but for easier interpretation of the results, these are back-transformed to the original scale to interpreted as the average ratio of the two estimators. The average ratio of the one minus Kaplan-Meier and the Aalen-Johansen estimator at the maximum follow-up time is 2.454. At earlier follow-up times, the ratio is smaller (see Table 8).

TABLE 7 Frequency categories of the AE probability in group A at the maximum follow-up time (no common tau)

		All events Aalen–Johansen				
		Very rare	Rare	Uncommon	Common	Very common
Incidence proportion	Very rare	0	0	0	0	0
	Rare	0	0	0	0	0
	Uncommon	0	0	0	0	0
	Common	0	0	0	1	0
	Very common	0	0	0	0	7
1-Kaplan–Meier	Very rare	0	0	0	0	0
	Rare	0	0	0	0	0
	Uncommon	0	0	0	0	0
	Common	0	0	0	0	0
	Very common	0	0	0	1	7

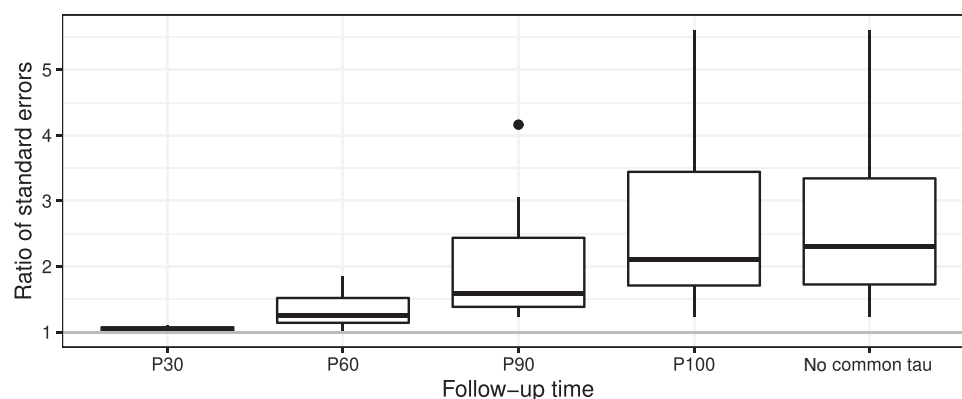


FIGURE 3 Boxplots of the ratios of the standard errors of the one minus Kaplan–Meier estimator divided by the gold-standard Aalen–Johansen estimator at the five follow-up times in group A

As an example for a meta-regression we investigate the effect of the percentage of censoring on this ratio. Table 9 displays the resulting estimates.

In a trial with 16.5% censoring the average ratio of one minus Kaplan–Meier and the gold-standard Aalen–Johansen estimator is 2.491. An increase in the percentage of censoring by 0.1 decreases the average ratio by the factor $\exp(-1.111 \cdot 0.1) = 0.895$ (using $\exp(-1.111) = 0.329$).

TABLE 8 Exemplary results of the meta-analysis for the comparison one minus Kaplan–Meier against gold-standard Aalen–Johansen estimator

Follow-up time point	$\exp(\hat{\theta})$	95% Confidence interval	ρ
No common tau	2.454	[1.595, 3.775]	0.617
P100	2.414	[1.573, 2.535]	0.613
P90	1.835	[1.328, 2.535]	0.463
P60	1.261	[1.103, 1.442]	0.182
P30	1.027	[1.002, 1.052]	0.022

TABLE 9 Exemplary results of the meta-regression for the comparison one minus Kaplan–Meier against gold-standard Aalen–Johansen estimator with input centered percentage of censoring at the maximum follow-up time (no common tau)

	Estimate	95% Confidence interval	p-value	ρ
$\exp(\hat{\theta})$	2.491	[1.583, 3.919]	<.0001	0.646
$\exp(\hat{\beta}_{\text{cens}})$	0.329	[0.010, 11.025]	.535	

7 | DISCUSSION

We have presented the rationale and statistical concept of the empirical, meta-analytical SAVVY study, which is presently on-going. The study aims to investigate the impact of commonly used methods to quantify AE incidence which fall short of accounting for both varying follow-up times and competing risks.

The study described in this paper considers time-to-first-AE only, but not recurrent AEs. The reasons are fourfold. First, we kept in mind the ultimate goal of safety evaluation in drug development that is accurately informing the product label adverse drug reaction section by providing the most relevant frequency category for SmPC or frequency for US PI (US prescribing information). Second, the incidence proportion is only meaningful as an estimator of absolute AE risk for first AEs, but not for recurrent ones. The incidence density could be computed for recurrent events in a meaningful way, but the assumption of a constant AE hazard would then be an even more restrictive parametric model (Windeler & Lange, 1995). Third, censoring, varying follow-up times and competing events will be no less important when AEs can be recurrent. For general recurrent events analyses, this has only very recently been emphasized by Andersen, Angst, and Ravn (2019). Fourth, in a time-to-first-event analysis, the absolute AE risk or cumulative AE probability, non-parametrically estimated by Aalen–Johansen, is a natural target quantity or estimand. In a recurrent events setting, the options for statistical modeling become more complex, because intermediate AEs will, in general, impact the incidence of subsequent AEs. One consequence is that in the time-to-first-event setting, the absolute AE risk can be expressed via fully conditional intensities, while the question of whether to use fully conditional or rather marginal approaches becomes a more pressing question when AEs are recurrent; see again Andersen et al. (2019). It is our intention that the SAVVY project will, in the future, also further investigate the analysis of recurrent AEs, and, to begin, such investigations shall be informed by the results of the study described in the present paper.

As a limitation, we also note that the empirical study of the SAVVY project is not able to generate generalizable results in terms of an average bias. In this first empirical study, we plan to collect data from various trials and types of AEs to see what extreme biases are possible. We are aware that our sample of included trials is not representative in any aspect and that the preregistration procedure of selected trials by the sponsor organizations will not eliminate the possibility of selecting trials depending on the expected results. It is our aim to provide a broader picture of possible biases that can occur in clinical trial than by an exemplary presentation of two or three preselected extreme example trials. As future research, such analyses for representative studies in only one indication or studies being representative according to specific other criteria are planned. For these, the same statistical analysis plan will be used.

Furthermore, there might be dependencies between the AEs of the same trial as one individual can contribute to more than one AE. These dependencies might impact the inference, confidence intervals, and testing of hypotheses in the meta-analysis. With the data at hand, it is impossible to check for such dependencies. But the meta-analysis is still an adequate way to explore possible biases by considering the ratio of two estimators as we do not aim to derive generalizable conclusions. In addition, we expect the possible dependencies to be limited, because the typical situation will be that one patient contributes one AE.

From a HTA point of view, a treatment policy estimand is of interest requiring the analysis of AEs over the entire follow-up period. In most clinical trials, it is current practice that AEs are only observed until some prespecified point in time after the last dose of treatment which makes it impossible to consider AEs thereafter. We also think that the recording of AEs after treatment discontinuation is important for a complete safety assessment, but it is neither our aim to investigate the impact of different AE recording period on the estimation of AE risk nor do the typically available data allow for such an investigation. The proposed investigation does not allow the conclusion that it is not required to document AE data after treatment discontinuation. Irrespective of the duration of AE recording, the occurrence of AEs is subject to competing events and censoring. Our aim is to investigate the methodological impact of inappropriate handling of competing events and censoring. Thereby, we consider two different ways to deal with treatment discontinuation; one censoring these premature ends of follow-up and the other treating them as competing events. The data typically

available from a trial allow to investigate this question which remains relevant even if AE data after discontinuation are available.

The current investigations within the SAVVY project focus on analyses of individual studies. In practice, these analyses would be integrated across trials. Particular problems arise when only a small number of trials is combined in a random effects meta-analysis (Bender et al., 2018) or the events are rare (Günhan, Röver, & Friede, 2019). Furthermore, strategies for signal detection in safety analyses are also not considered here, but are subject to an on-going research (see, e.g., Gould, 2018).


ACKNOWLEDGMENTS

Open access funding enabled and organized by Projekt DEAL.

CONFLICT OF INTEREST

VJ, KR, and FL are employees of Novartis Pharma AG (Basel, Switzerland), F. Hoffmann-La Roche (Basel, Switzerland), and Pfizer Deutschland (Berlin, Germany), respectively. TF has received personal fees for consultancies (including data-monitoring committees) from Bayer, Boehringer Ingelheim, Janssen, Novartis, and Roche, all outside the submitted work. JB has received personal fees for consultancy from Pfizer, all outside the submitted work. CS has received personal fees for consultancies (including data-monitoring committees) from Novartis and Roche, all outside the submitted work. The companies mentioned will contribute data to the meta-analysis. RS has declared no conflict of interest.

OPEN RESEARCH BADGES

 This article has earned an Open Data badge for making publicly available the digitally-shareable data necessary to reproduce the reported results. The data is available in the Supporting Information section.

This article has earned an open data badge “**Reproducible Research**” for making publicly available the code necessary to reproduce the reported results. The results reported in this article could fully be reproduced.

ORCID

Regina Stegherr  <https://orcid.org/0000-0002-0528-348X>

Jan Beyersmann  <https://orcid.org/0000-0002-3793-4611>

Kaspar Rufibach  <https://orcid.org/0000-0002-2634-1167>

Claudia Schmoor  <https://orcid.org/0000-0001-5610-9425>

Tim Friede  <https://orcid.org/0000-0001-5347-7441>

REFERENCES

- Aalen, O., Borgan, O., & Gjessing, H. (2008). *Survival and event history analysis*. Berlin, Germany: Springer Science & Business Media.
- Aalen, O., & Johansen, S. (1978). An empirical transition matrix for non-homogeneous Markov chains based on censored observations. *Scandinavian Journal of Statistics*, 5, 141–150.
- Allignol, A., Beyersmann, J., & Schumacher, M. (2011). mvna: An R package for the Nelson–Aalen estimator in multistate models. *R News*, 8, 48–50.
- Allignol, A., Beyersmann, J., & Schmoor, C. (2016). Statistical issues in the analysis of adverse events in time-to-event data. *Pharmaceutical Statistics*, 15, 297–305.
- Allignol, A., Schumacher, M., & Beyersmann, J. (2011). Empirical transition matrix of multi-state models: The etm package. *Journal of Statistical Software*, 38, 1–15.
- Altman, D. G., & Bland, J. M. (1983). Measurement in medicine: the analysis of method comparison studies. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 32, 307–317. <http://www.jstor.org/stable/2987937>.
- Andersen, P. K. (1983). Comparing survival distributions via hazard ratio estimates. *Scandinavian Journal of Statistics*, 10, 77–85.
- Andersen, P. K., Angst, J., & Ravn, H. (2019). Modeling marginal features in studies of recurrent events in the presence of a terminal event. *Lifetime Data Analysis*, 25, 681–695.
- Andersen, P. K., Borgan, O., Gill, R. D., & Keiding, N. (1993). *Statistical models based on counting processes*. Berlin, Germany: Springer Science & Business Media.
- Bender, R., & Beckmann, L. (2019). Limitations of the incidence density ratio as approximation of the hazard ratio. *Trials*, 20, 485.
- Bender, R., Beckmann, L., & Lange, S. (2016). Biometrical issues in the analysis of adverse events within the benefit assessment of drugs. *Pharmaceutical Statistics*, 15, 292–296.
- Bender, R., Friede, T., Koch, A., Kuss, O., Schlattmann, P., Schwarzer, G., & Skipka, G. (2018). Methods for evidence synthesis in the case of very few studies. *Research Synthesis Methods*, 9, 382–392.

- Beyersmann, J., & Schmoor, C. (2019). The analysis of adverse events in randomized clinical trials. In S. Halabi & S. Michiels (Eds.), *Textbook of clinical trials in oncology: a statistical perspective*, (537–558). Boca Raton, FL: CRC Press.
- Beyersmann, J., & Schrade, C. (2017). Florence Nightingale, William Farr and competing risks. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 180, 285–293.
- Committee for Human Medicinal Products (2016). Guideline for good clinical practice E6(R2). Retrieved from <https://www.ema.europa.eu/en/ich-e6-r2-good-clinical-practice#current-version---revision-2-section>.
- Crowe, B. J., Xia, H. A., Berlin, J. A., Watson, D. J., Shi, H., Lin, S. L., .. et al., (2009). Recommendations for safety planning, data collection, evaluation and reporting during drug, biologic and vaccine development: A report of the safety planning, evaluation, and reporting team. *Clinical Trials*, 6, 430–440.
- EMA (2009). A guideline on summary of product characteristics (SmPC). Retrieved from https://ec.europa.eu/health/sites/health/files/files/eudralex/vol-2/c/smpc_guideline_rev2_en.pdf.
- Ferreira-González, I., Permanyer-Miralda, G., Busse, J. W., Bryant, D. M., Montori, V. M., Alonso-Coello, P., ... Gordon, H. Guyatt (2017). Methodologic discussions for using and interpreting composite endpoints are limited, but still identify major concerns. *Journal of Clinical Epidemiology*, 60, 651–657.
- Gill, R. D., & Johansen, S. (1990). A survey of product-integration with a view toward application in survival analysis. *The Annals of Statistics*, 18, 1501–1555.
- Gosho, M., Sato, Y., Nagashima, K., & Takahashi, S. (2018). Trends in study design and the statistical methods employed in a leading general medicine journal. *Journal of Clinical Pharmacy and Therapeutics*, 43, 36–44.
- Gould, A.L. (2018). Unified screening for potential elevated adverse event risk and other associations. *Statistics in Medicine*, 37, 2667–2689.
- Günhan, B.K., Röver, C., & Friede, T. (2020). Meta-analysis of few studies involving rare events. *Research Synthesis Methods*, 11, 74–90.
- Hedges, L. V., & Olkin, I. (2014). *Statistical methods for meta-analysis*. New York: Academic press.
- Hjort, N. (1992). On inference in parametric survival data models. *International Statistical Review*, 60, 355–387.
- Horton, N., & Switzer, S. (2005). Statistical methods in the journal. *New England Journal of Medicine*, 353, 1977–1979.
- IQWiG (2017). Allgemeine Methoden, Version 5.0. Institute of Quality and Efficiency in Health Care. Retrieved from <https://www.iqwig.de/de/methoden/methodenpapier.3020.html>.
- Kieser, M., & Hauschke, D. (2005). Assessment of clinical relevance by considering point estimates and associated confidence intervals. *Pharmaceutical Statistics: The Journal of Applied Statistics in the Pharmaceutical Industry*, 4, 101–107.
- Kraemer, H. C. (2009). Events per person-time (incidence rate): a misleading statistic? *Statistics in Medicine*, 28, 1028–1039.
- Krouwer, J. S. (2008). Why Bland–Altman plots should use X , not $(Y + X)/2$ when X is a reference method. *Statistics in Medicine*, 27, 778–780.
- Lacny, S., Wilson, T., Clement, F., Roberts, D. J., Faris, P., Ghali, W., & Marshall, D. (2015). Kaplan–Meier survival analysis overestimates the risk of revision arthroplasty: a meta-analysis. *Clinical Orthopaedics and Related Research*, 473, 3431–3442.
- Lacny, S., Wilson, T., Clement, F., Roberts, D. J., Faris, P., Ghali, W., & Marshall, D. (2018). Kaplan–Meier survival analysis overestimates cumulative incidence of health-related events in competing risk settings: a meta-analysis. *Journal of Clinical Epidemiology*, 93, 25–35.
- Lawless, J. F., & Cook, R. J. (2019). A new perspective on loss to follow-up in failure time and life history studies. *Statistics in Medicine*, 38, 4583–4610.
- O'Neill, R. T. (1987). Statistical analyses of adverse event data from clinical trials: Special emphasis on serious events. *Drug Information Journal*, 21, 9–20.
- Pocock, S. J., Clayton, T. C., & Altman, D. G. (2002). Survival plots of time-to-event outcomes in clinical trials: Good practice and pitfalls. *The Lancet*, 359, 1686–1689.
- R Core Team (2018). *R: a language and environment for statistical computing*. R foundation for statistical computing, Vienna, Austria. Retrieved from <https://www.R-project.org/>.
- Rufibach, K. (2019). Treatment effect quantification for time-to-event endpoints – estimands, analysis strategies, and beyond. *Pharmaceutical Statistics*, 18, 145–165.
- Sato, Y., Gosho, M., Nagashima, K., Takahashi, S., Ware, J. H., & Laird, N. M. (2017). Statistical methods in the journal—An update. *New England Journal of Medicine*, 376, 1086–1087.
- Schumacher, M., Ohneberg, K., & Beyersmann, J. (2016). Competing risk bias was common in a prominent medical journal. *Journal of Clinical Epidemiology*, 80, 135–136.
- Siddiqui, O. (2009). Statistical methods to analyze adverse events data of randomized clinical trials. *Journal of Biopharmaceutical Statistics*, 19, 889–899.
- Therneau, T. M., & Grambsch, P. M. (2000). *Modeling survival data: extending the Cox model*. Berlin, Germany: Springer Science & Business Media.
- Unkel, S., Amiri, M., Benda, N., Beyersmann, J., Knoerzer, D., Kupas, K., .. Friede, T. (2019). On estimands and the analysis of adverse events in the presence of varying follow-up times within the benefit assessment of therapies. *Pharmaceutical Statistics*, 18, 166–183.
- van der Vaart, A. W. (1998). *Asymptotic statistics*. Cambridge, UK: Cambridge University Press.
- van Walraven, C., & McAlister, F. A. (2016). Competing risk bias was common in Kaplan–Meier risk estimates published in prominent medical journals. *Journal of Clinical Epidemiology*, 69, 170–173.
- Veroniki, A.A., Jackson, D., Bender, R., Kuss, O., Langan, D., Higgins, J.P.T., .. Salanti, G. (2019). Methods to calculate uncertainty in the estimated overall effect size from a random effects meta analysis. *Research Synthesis Methods*, 10, 23–43.

- Veroniki, A. A., Jackson, D., Viechtbauer, W., Bender, R., Bowden, J., Knapp, G., & Salanti, G. (2016). Methods to estimate the between-study variance and its uncertainty in meta-analysis. *Research Synthesis Methods*, 7, 55–79.
- Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, 36, 1–48.
- Windeler, J., & Lange, S. (1995). Events per person year—A dubious concept. *British Medical Journal*, 310, 454–456.
- Yang, F., Wittes, J., & Pitt, B. (2019). Beware of on-treatment safety analyses. *Clinical Trials*, 16, 63–70. <https://doi.org/10.1177/1740774518812774>, PMID: 30445833.

SUPPORTING INFORMATION

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

How to cite this article: Stegherr R, Beyersmann J, Jehl V, Rufibach K, Leverkus F, Schmoor C, Friede T. Survival analysis for AdVerse events with VarYing follow-up times (SAVVY): Rationale and statistical concept of a meta-analytic study. *Biometrical Journal*. 2021;63:650–670. <https://doi.org/10.1002/bimj.201900347>