Rufibach *et al. Trials* (2024) 25:353
https://doi.org/10.1186/s13063-024-08186-7

Trials

METHODOLOGY

Open Access

Survival analysis for AdVerse events with VarYing follow-up times (SAVVY): summary of findings and assessment of existing guidelines

Kaspar Rufibach[1*], Jan Beyersmann[2], Tim Friede[3], Claudia Schmoor[4] and Regina Stegherr[5]

Abstract

Background The SAVVY project aims to improve the analyses of adverse events (AEs) in clinical trials through the use of survival techniques appropriately dealing with varying follow-up times and competing events (CEs). This paper summarizes key features and conclusions from the various SAVVY papers.

Methods Summarizing several papers reporting theoretical investigations using simulations and an empirical study including randomized clinical trials from several sponsor organizations, biases from ignoring varying follow-up times or CEs are investigated. The bias of commonly used estimators of the absolute (incidence proportion and one minus Kaplan-Meier) and relative (risk and hazard ratio) AE risk is quantified. Furthermore, we provide a cursory assessment of how pertinent guidelines for the analysis of safety data deal with the features of varying follow-up time and CEs.

Results SAVVY finds that for both, avoiding bias and categorization of evidence with respect to treatment effect on AE risk into categories, the choice of the estimator is key and more important than features of the underlying data such as percentage of censoring, CEs, amount of follow-up, or value of the gold-standard.

Conclusions The choice of the estimator of the cumulative AE probability and the definition of CEs are crucial. Whenever varying follow-up times and/or CEs are present in the assessment of AEs, SAVVY recommends using the Aalen-Johansen estimator (AJE) with an appropriate definition of CEs to quantify AE risk. There is an urgent need to improve pertinent clinical trial reporting guidelines for reporting AEs so that incidence proportions or one minus Kaplan-Meier estimators are finally replaced by the AJE with appropriate definition of CEs.

Keywords Aalen-Johansen estimator, Adverse events, Competing events, Drug safety, Incidence proportion, Incidence density, Kaplan-Meier estimator

*Correspondence:
Kaspar Rufibach
kaspar.rufibach@roche.com
Full list of author information is available at the end of the article

© The Author(s) 2024. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

Background

In randomized clinical trials (RCT), an essential part of the benefit-risk assessment of treatments is the quantification of the risk of experiencing adverse events (AEs), and comparing these risk between treatment arms. Methods commonly employed to quantify absolute adverse event (AE) risk either do not account for varying follow-up times, censoring, or for competing events (CEs), although appreciation of these are important in risk quantification, see, e.g., O'Neill [1] and Procter and Schumacher [2].

Analyses of AE data in clinical trials can be improved through the use of survival techniques that account for varying follow-up times, censoring, and CEs. Varying follow-up times refer to the fact that if patients are assessed for AEs in regular intervals it may happen, even in absence of censoring, that depending on when a patient entered the trial, the follow-up time at a reporting event varies between patients. Similar to an efficacy endpoint, *censoring*, or rather *administrative censoring*, refers to the fact that for some of the patients we may only have incomplete observations in the sense that we know that they did not experience an AE up to the cut-off of the reporting timepoint, but that their observation time to have an AE continues beyond that. A thorough discussion of CEs in the context of AE risk quantification is provided below.

Precise definitions of estimators are also given below.

The AJE [3, 4] can be considered the non-parametric gold-standard method when quantifying absolute AE risk. The reason is that the AJE is the standard (non-parametric) estimator that accounts for CEs, censoring, and varying follow-up times simultaneously and, being non-parametric, does not rely on restrictive parametric assumptions, such as constant hazards. Any other estimator of AE probability, such as incidence proportion, probability transform incidence density, or one minus Kaplan-Meier, delivers biased estimates in general.

To quantify that bias for all these methods in an ideal scenario, Stegherr et al. [5] ran a comprehensive simulation study. Two key findings were (1) that ignoring CEs is more of a problem than falsely assuming constant hazards by the use of the incidence density and (2) that the choice of the AE probability estimator is crucial for the estimation of relative effects, i.e., comparison between groups.

To illustrate and further solidify these simulation-based results with real data the SAVVY consortium, a collaboration between nine pharmaceutical companies and three academic institutions meta-analyzed data from 17 randomized controlled trials (RCT).

In this article, we summarize the results of the empirical study, reported in two separate publications: Stegherr et al. [6] was concerned with estimation of AE risk in one treatment group and Rufibach et al. [7] with the comparison of AE risks between two groups in an RCT. A cursory assessment of how relevant guidelines recommend to estimate AE risk is given with a call for updates. We conclude with a discussion.

Definition of key terms

The target of estimation, or estimand, for the compared estimators is the probability $P(\text{AE in } [0, t])$. We will also call this quantity *risk*. In situations not additionally complicated by varying follow-up times or censoring, i.e., when all patients are observed for the same amount of time, this probability can easily be estimated using the incidence proportion, see below. However, as soon as we have varying follow-up and/or censoring, the incidence proportion will typically be a biased estimate of $P(\text{AE in } [0, t])$.

Scientific questions of the SAVVY project

The overarching scientific questions of SAVVY can be phrased as follows:

1) For estimation of the probability of an AE, how biased are commonly used estimators, especially the incidence proportion and one minus Kaplan-Meier, in presence of censoring, varying follow-up between patients, CEs, and in the case of incidence densities a restrictive parametric model?
2) What is the bias of common estimators that quantify the relative risk of experiencing an AE between two treatment arms in a RCT?
3) Can trial characteristics be identified that help explain the bias in estimators?
4) How does the use of potentially biased estimators impact qualification of AE probabilities and relative effects in regulatory settings?

Within the SAVVY project, these questions were approached in two ways: first, in Stegherr et al. [5], via simulation of clinical trial data. This approach has the advantage that the true underlying data generating mechanism is specified by the researcher and therefore known. This allows to exactly quantify the bias of a given estimator, i.e., to answer 1) and 2) above (it would also allow to answer Question 4, but that was not addressed in Stegherr et al. [5]). Second, in Stegherr et al. [6] and Rufibach et al. [7], biases of commonly used estimators of absolute and relative AE risks were estimated by comparing them

to the best available estimator. Having real clinical trial data available also allows to answer Question 3 above, through meta-analytic methods.

Competing events and their connection to the ICH E9 estimands addendum

In what follows, we will use *competing event* (apart from direct quotes) and consider it synonymous to *competing risk*.

An important but largely unrecognized aspect when quantifying AE risk is the likely presence of CEs. Gooley et al. [8] define a CE as

> *"We shall define a competing risk as an event whose occurrence either precludes the occurrence of another event under examination or fundamentally alters the probability of occurrence of this other event".*

whereas the ICH E9(R1) estimands addendum [9] defines an intercurrent event as

> *"Events occurring after treatment initiation that affect either the interpretation or the existence of the measurements associated with the clinical question of interest".*

Here, *another event under examination* and *measurements associated with the clinical question of interest* refer to a defined AE of interest. So, a CE in this context is any clinical event that precludes the occurrence of an AE, the most prominent example being death. The above two definitions appear to be, if not the same, then at least very related. However, the ICH E9(R1) addendum does not discuss CEs, so it is not entirely clear how to embed CEs into the addendum framework, i.e., whether and if yes which of the proposed strategies of the estimand addendum applies to the CE situation. More research and discussion is needed to align CEs (if necessary at all), and the analysis of complex time-to-event data with the addendum.

Stegherr et al. [6] and Rufibach et al. [7] took a pragmatic approach and defined events as "competing" that preclude the occurrence or recording of the AE under consideration in a time-to-first-event analysis. Specifically, one important CE is death before AE. In addition, any event that would both be viewed from a patient perspective as an event of his/her course of disease or treatment and would stop the recording of the interesting AE was viewed as a CE. Since all these CEs apart from *death* may be prone to some subjectivity in the empirical analysis reported in Stegherr et al. [6] and Rufibach et al. [7], a variant of the estimators with a CE of *death only* was also considered. Since results were in line with the broader definition of CEs as given above, we omit the results for the *death only* variant here.

Estimation methods

A precise mathematical definition of all estimators of the probability of an AE that were compared in SAVVY is provided in Stegherr et al. [10], a prospectively published statistical analysis plan for the SAVVY project. A short introduction in all estimators is also provided in Stegherr et al. [6]. In this overview article, we only provide very brief descriptions of the considered estimators.

Incidence proportion

By far the most commonly used estimator, e.g., in standard safety reporting that enters benefit-risk assessment for the approval of new medicines, to estimate the risk of an adverse event (risks are estimated for one AE at a time) up to a maximal observation timepoint τ is the *incidence proportion* [4]. It simply divides the number of patients with an observed AE on $[0, \tau]$ in group A by the number of patients in group A. The incidence proportion is an estimator of the probability that an AE happens in the interval $[0, \tau]$, *and* that this AE is observed, i.e., not censored. This illustrates that the incidence proportion is not properly dealing with censored observations. However, it correctly accounts for CEs; see Allignol et al. [4] for an exemplary illustration of that feature.

Incidence density

To account for the differing follow-up times between patients, researchers and guidelines (see Table 4) often advocate to use the *incidence density* or *incidence rate*, where the number of AEs in the nominator is divided by the total patient-time at risk instead of simply the number of patients. As such, the incidence density does not directly estimate the probability of an AE, but rather the AE hazard. As described in Stegherr et al. [10], this hazard estimator can easily be transformed to indeed estimate the probability of an AE. However, it only does so assuming that the AE hazard is constant, i.e., the probability-transformed incidence density is a fully parametric estimator. In addition, as such it does not correctly account for CEs, but it can be modified to do so, leading to the *probability transform incidence density accounting for CEs*.

One minus Kaplan-Meier

Researchers are often aware of the inability of the incidence proportion to properly deal with varying follow-up times and censoring. As a remedy, they (and many guidelines, see Table 4) then advocate to consider *time to AE* and estimate the probability of an AE by reading off the one minus Kaplan-Meier estimator at a timepoint of interest, e.g., the end of observation time τ or at an earlier timepoint. While indeed, this estimator properly

accounts for varying follow-up times and censoring the question remains how to deal with CEs. Numerous papers have been written [4, 8] and providing technical arguments explaining why the Kaplan-Meier estimator is a biased estimator of the probability of an AE. Intuitively, one minus the Kaplan-Meier estimator estimates the distribution function of the time of interest, that is, it tends towards one as we move to the right on the time axis. However, if we add up the probability of an AE and the probability of a CE in a truly CE scenario that also must be equal to one, implying that the probability of an AE is *strictly smaller* than one. As a consequence, the Kaplan-Meier estimator to estimate the probability of an AE will be biased upwards.

Aalen-Johansen estimator

Finally, there is an estimator that at the same time accounts for (random or independent) censoring, respects varying follow-up times, accounts for CEs in the right way, and is fully nonparametric and therefore free of bias introduced by any of these processes: the AJE [3]. It is therefore considered the gold standard estimator. In the empirical analysis of the 17 RCTs in the SAVVY project, it served as a benchmark against which all estimators were measured against. The term *bias* was therefore used for deviations of the estimators from this benchmark estimator, or gold standard. For an evaluation of the true bias, i.e., the deviation of estimators to the true underlying value, we refer to Stegherr et al. [5].

Table 1 in Stegherr et al. [6] concisely summarizes the properties of each considered estimator with respect to whether it accounts for censoring and CEs and whether it makes a parametric assumption, and we therefore reproduce it here in Table 1 for the estimators discussed here.

Quantification of bias—the SAVVY project

One of the goals of the SAVVY project is to quantify the bias of standard estimators of the probability of an AE. Based on simulations, i.e., comparing estimated to true underlying values from which the data was simulated, this has been done in Stegherr et al. [5]. Key findings in this study were that ignoring CEs is more of a problem than falsely assuming constant hazards. The one minus Kaplan-Meier estimator may overestimate the true AE probability by a factor of four at the largest observation time. Moreover, the choice of the AE probability estimator is crucial for group comparisons.

Table 1 Overview whether the estimators deal with the possible sources of bias, reproduced from Stegherr et al. [6]

	Accounts for censoring	Makes no constant hazard assumption	Accounts for CEs
Incidence proportion	No	Yes	Yes
1-Kaplan-Meier	Yes	Yes	No
gold-standard AJE	Yes	Yes	Yes

To confirm these results in real clinical trial data, three academic institutions and nine pharmaceutical companies collaborated within the SAVVY consortium. In order not to have to share the trial data, macros to compute the above estimators were developed centrally. Companies then ran these macros on trial data of their choice and only returned high-level results (estimated values of AE probabilities for each estimator and some basic trial characteristics) to the central data analysis unit hosted at one of the academic institutions. The central data analysis unit then meta-analyzed these results to quantify biases for estimators and impact on regulatory decision-making. A statistical analysis plan for these analyses was published [10]. Results for estimation of the probability of an AE in one arm are reported in Stegherr et al. [6] and for the comparison between two groups in Rufibach et al. [7].

In this overview paper, we focus attention on summarizing results for the most commonly used estimators of AE risks, namely the incidence proportion and one minus Kaplan-Meier in comparison to the gold standard AJE. Furthermore, we report results for the maximal evaluation time as defined in the above two papers, i.e., the latest time at which a dataset has an observation, either event or censored. Results for shorter evaluation times were in line with those for the maximal evaluation time.

Results

Ten organizations provided 17 trials including 186 types of AEs (median 8; interquartile range [3, 9]). Twelve (71.6% out of 17) trials were from oncology, nine (52.9%) were actively controlled, and eight (47.1%) were placebo controlled. The trials included between 200 and 7171 patients (median = 443; interquartile range [411, 1134]). Median follow-up of the treatment group was 927 days (interquartile range [449, 1380]).

For one of the 17 trials, details of the trial and the analysis of one of the AEs by the different methods investigated in this paper are presented in Stegherr et al. [5].

Note that for RRs only those AEs were considered where neither the estimated AE probability in the experimental arm nor the estimated AE probability in the control arm is 0 ($n = 156$ for incidence proportion and $n = 155$ for one minus Kaplan-Meier). This also applies to the categories in Table 2.

Empirical bias of common estimators of absolute risk with respect to the gold-standard AJE

For the comparison of the AE probabilities, Stegherr et al. [6] focused on the experimental arm. Median

follow-up was 927 days (interquartile range [449, 1380]). The median of the gold-standard AJE was 0.092 (minimum 0 and maximum 0.961), i.e., the estimated probability of an AE was 9.2% on average over all 186 AEs over all trials. The one minus Kaplan-Meier estimator was on average about 1.2-fold larger than the AJE and the probability transform of the incidence density ignoring CEs was even two-fold larger. The average empirical bias (i.e., the difference between the considered estimator and the gold-standard AJE) using the incidence proportion was less than 5%. Assuming constant hazards using incidence densities was hardly an issue provided that CEs were accounted for. However, beyond these average biases it was striking how extreme the bias could become: For one minus Kaplan-Meier, in our empirical analysis using real clinical trial data, we observed an overestimation of the AE probability up to a factor of five, whereas for the incidence proportion we observed underestimations of up to a factor of three. This is in line with the findings of the simulation study in Stegherr et al. [5] and illustrates that using too simplistic estimators for the probability of an AE can be truly misleading. To evaluate what study characteristics impact the bias, a meta-regression was performed. For that, the response variable was defined as the AE probability estimates obtained with the different estimators divided by the AE probability obtained with the gold-standard AJE, considered on the log-scale. Covariates were the proportion of censoring, the evaluation time point τ (i.e., the maximal time to event in years – AE, CE or censoring – observed), and the size of the AE probability estimated by the gold-standard AJE. The meta-regression showed that the bias depended primarily on the amount of censoring and on the amount of CEs.

Finally, according to the European Commission's guideline on summary of product characteristics (SmPC) [11] and based on the recommendations of the *Council for International Organizations of Medical Science* (CIOMS) Working Groups III and V [12], the AE risk is classified into frequency categories which are defined by "very rare," "rare," "uncommon," "common," and "very common" when the risk is < 0.01%, < 0.1%, < 1%, < 10%, ≥ 10%, respectively. Stegherr et al. [6] (Table 2) assigned these categories to AE probability estimates from all estimators (and all 186 AEs) and compared them to the categories resulting from the AJE. As an example, systematic overestimation of the one minus Kaplan-Meier resulted in upgrading of 2/6 AEs from "uncommon" to "common" and 14/86 from "common" to "very common."

Empirical bias of probability-based estimators of relative risk with respect to the gold-standard AJE

Naively, one could ask the question whether when we have biased estimates of the AE probability in two groups, that go in the same direction in both groups, and want to compare them, the bias "cancels out." To assess this hypothesis, Rufibach et al. [7] extended the work from Stegherr et al. [6] to quantify the bias when comparing AE risks between groups in a randomized trial. The focus of Rufibach et al. [7] is not to define what a fit-for purpose estimand to quantify safety risk could be, but rather to evaluate statistical properties of commonly used estimators in the presence of varying follow-up and CEs. A thorough discussion of effect measures to quantify the relative risk is given there as well (Section 2.7-Effect Measures). Without going into causal or estimand details, the effects to be compared between groups are to be understood as total effects, comparing patients' AE risk in this world and in the presence of CEs. The estimators that were finally assessed are

- The risk ratio (RR) $\widehat{RR} = \hat{q}_E/\hat{q}_C$ for estimators $\hat{q}_E$ and $\hat{q}_C$ of AE probabilities calculated at a specific evaluation time within each treatment arm (E for experimental, C for control)
- The hazard ratio (HR) for both, the AE of interest and the CE

In the one-sample case, estimates of AE probabilities were benchmarked on the gold-standard AJE. This, because the latter is a fully nonparametric estimator that accounts for censoring, does not rely on a constant hazard assumption, and accounts for CEs, as discussed above. So, as a straightforward extension for the comparison of AE probabilities between two arms using the RR, we benchmark the latter on the RR estimated using the AJE in each arm, with variance derived using the delta rule. The gold-standard for estimates of the HR will be the HR from Cox regression. This is because the latter is typically used to quantify a treatment effect not only for efficacy, but also for time-to-first-AE type endpoints. Variances of comparisons of different estimators of the RR and HR will be received via bootstrapping, because of the dependency of estimators computed on the same dataset; see Stegherr et al. [5].

For the one-sample case of estimation of absolute AE risk direction of biases can be derived analytically: incidence proportion systematically under- and one minus Kaplan-Meier systematically overestimates the true AE risk. However, there is no such derivation possible for direction of the bias for the RR or HR. Rufibach et al. [7] found the RR based on the incidence proportion overestimates the RR computed using the gold-standard AJE by up to a factor of three, and one minus Kaplan-Meier underestimates up to a factor of four, see Fig. 1. Interestingly, dividing the two biased estimates of the AE probability based on the incidence proportion, which both

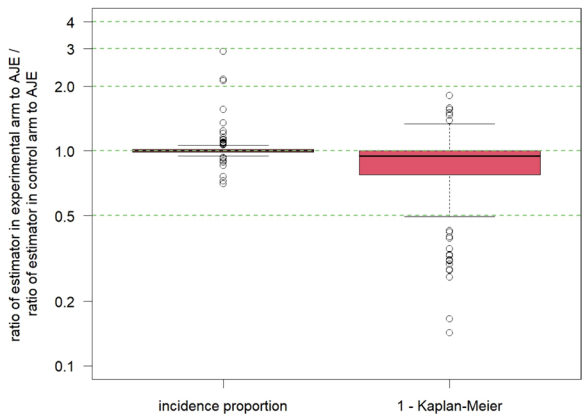

Fig. 1 Ratio of RRs estimated with estimator of interest and the gold-standard AJE

tend to *underestimate* the true AE probability, leads to an estimate of the RR that on average performs comparably to the AJE. Apart from shedding light on estimation quality of the incidence proportion and one minus Kaplan-Meier to estimate the RR, we conclude that different patterns of under- or overestimation of absolute AE probabilities can lead to similar performance for RR. This implies that in general, one cannot conclude how an estimator of the relative AE risk performs based on looking how these same estimators perform on estimation of arm-wise AE probabilities.

As discussed in Stegherr et al. [6], one reason for the good performance of the incidence proportion might be a high amount of CEs before possible censoring. However, not only the proportion of censoring but also the timing of the censoring is relevant.

Meta-analyses confirmed the above impressions. Meta-regressions showed that (1) the key difference between estimators lies in the value of the average RR and (2) the impact of covariates is overall limited, compared to the average RR the estimated coefficients are close to 1. This emphasizes that the choice of the estimator is key and that this holds true over a wide range of possible data configurations quantified through the considered covariates.

A key finding of Rufibach et al. [7] was that categorization of evidence based on RR crucially depends on the estimator one uses to estimate the RR. We therefore reproduce Table 1 in their paper here in a version trimmed down to incidence proportion and one minus Kaplan-Meier; see Table 2. Overall, we find quite a number of switches to neighboring categories, more so than for estimators of the absolute AE risk in one arm. Reasons for switches are wider CIs of the AJE as well as RR estimates/CI bounds that are close to the cutoffs between categories. As the incidence proportion on average estimates the RR well, we see a similar number of switches to a higher ($n = 8$, below the diagonal in Table 2) and lower ($n = 9$) evidence category. Not surprisingly, for one minus Kaplan-Meier that underestimates the RR with respect to the gold-standard AJE, we see relevantly more switches to a lower than higher evidence category, namely $n = 41/n = 16$, 32/8, and 28/6, respectively.

In summary, the choice of the estimator of the RR does have a relevant impact on the conclusions. Note that there is no universally accepted standard how one should combine a point estimate and its associated variability, in our case RR, into evidence categories. As an example, Rufibach et al. [7] used a categorization motivated by the methods put forward by the IQWiG [13] for severe AEs (Table 14) to be used for the German benefit-risk assessment.

Table 2 The impact of the choice of relative effect estimators, incidence proportion, and one minus Kaplan-Meier, for AE probabilities on qualitative conclusions. Diagonal entries are set in bold face. Deviations from the gold-standard AJE are the off-diagonal entries. Off-diagonal zeros are omitted from the display

		Gold-standard Aalen-Johansen			
		(0) no effect	**(a) minor**	**(b) considerable**	**(c) major**
Incidence proportion	(0) no effect	**84**	5		
	(a) minor	3	**10**	2	
	(b) considerable	1	2	**12**	2
	(c) major	1		1	**33**
One minus Kaplan-Meier	(0) no effect	**84**	9	4	8
	(a) minor	3	**6**	3	3
	(b) considerable	2	1	**7**	5
	(c) major		1	1	**18**

Table 3 Conclusions of the RR calculated with the gold-standard Aalen-Johansen estimator compared to the conclusions of the HR calculated with the Cox model. The table shows the analysis of those $n = 94$ AEs that were observed with an absolute frequency of ≥ 10 in each arm. Off-diagonal zeros are omitted from the display

		HR Cox for AE			
		(0) no effect	**(a) minor**	**(b) considerable**	**(c) major**
RR gold-standard Aalen-Johansen	(0) no effect	**42**	3	3	1
	(a) minor	9	**2**	1	
	(b) considerable	4	1	**3**	2
	(c) major	2		4	**17**

Table 4 Coverage of relevant time-to-event aspects for quantification of AE risk for an opportunistic sample of safety guidelines. "x" means that this aspect is mentioned in the respective guideline

Guideline	Acknowledges varying FU	Proposes incidence density	Acknowledges constant hazard assumption	Proposes life-table/one minus Kaplan-Meier	Acknowledges CEs
ICH E3 [19]	x			x	
ICH E9 [20]	x	x		x	
SmPC [11]	x	x			
CIOMS [18]	x	x	x	x	
FDA premarketing [21]	x	x	x	x	
CONSORT Harm 2022 update [22]	x				

Empirical bias of hazard-based estimators of relative risk with respect to the gold-standard AJE

As for hazard-based inference, it is important to note that, even if the event of interest is AE, effects on all other CEs is generally recommended for any (hazard-based) analysis of CEs [14]. Rufibach et al. [7] found that the hazard of AE is generally larger for the experimental compared to the control arm, meaning that the instantaneous risk of AE is typically higher in the experimental arm of an RCT, not unexpectedly. For the hazards of CEs, for both variants, i.e., considering death only as CE or including more reasons as described above, what we find is that the hazard in the experimental arm is generally lower than in the control arm, i.e., there is an effect of the experimental treatment on the CE. If we simply censored at CEs, we would thus introduce arm-dependent censoring, a feature that may lead to biased effect estimates [15, 16]. The ratio of the incidence densities of the AE in the two arms underestimates with respect to the Cox regression HR while for the other two endpoints on the median they turn out to be approximately unbiased compared to the Cox HR, with a tendency to overestimation when accounting for all CEs. To appreciate the differences between the two estimators of the RR based on hazards, i.e., the incidence density ratio and the gold-standard Cox regression HR, recall the properties of the two methods: Both properly account for censoring and they properly estimate event-specific hazards, or rather the relative effect based on these. The only difference between the two methods is what they assume about the shape of the underlying hazard: the incidence density assumes them to be *constant* up to the considered follow-up time, which also implies that they are *proportional*.

The impact of the use of the different estimators on the conclusions derived from the comparison of treatment arms was again investigated by the use of categories. These are typically derived from comparing the confidence interval (CI) of the RR to thresholds. In contrast to the usual IQWiG procedure, however, they did not only categorize the benefit of a therapy, but also the harm, where they did not distinguish between a positive and a negative treatment effect. Four categories were possible: (0) "no effect" if 1 is included in the CI, (a) "minor" ("gering") if the upper bound of the CI is in the interval [0.9, 1] for a RR< 1 or the lower bound in the interval [1, 1.11] for a RR> 1, (b) "considerable" ("beträchtlich") if the upper bound of the CI is in the interval [0.75, 0.9] for a RR< 1 or the lower bound in the interval [1.11, 1.33] for a RR> 1, and (c) "major" ("erheblich") if the upper bound is smaller than 0.75 for a RR< 1 or the lower bound greater than 1.33 for a RR> 1. The same categorization was used for the HR instead of RR.

Rufibach et al. [7] have considered two effect measures to quantify the RR of an AE in two arms: the RR based

on AE probabilities and the HR computed from Cox regression. Rufibach et al.'s [7] analyses revealed that all the considered estimators are overall inferior to the two gold standards that were considered, either the RR based on the arm-wise AJE or the HR based on Cox regression. A question that remains is whether the qualitative conclusions drawn based on the two gold standards are relevantly different when relying on the criteria put forward by the IQWiG (Table 14 in their general methods document [13]); see Table 9 in Rufibach et al. [7], which we reproduce here as Table 3. Quite some different classifications based on the two estimators of the RR were observed. However, this is not a surprise, as the *estimand* the two methods look at is not the same (see Varadhan et al. [17] for an exposition of this issue): Cox HR quantifies a relative effect based on an endpoint of AE *hazard*, whereas RR based on gold-standard Aalen-Johansen is based on a comparison of *probabilities* at a evaluation time; see Rufibach et al. [7] for an extended discussion of this issue.

Overall conclusions from the empirical analysis

To conclude this section, based on theoretical and empirical evidence, Stegherr et al. [6] clearly recommend the AJE as the non-parametric, unbiased estimator in the presence of both CEs and censoring. If a parametric analysis based on incidence densities is considered, they strongly recommend to incorporate incidence densities of CEs as they detail in their paper. This is also preferable over the one minus Kaplan-Meier approach. A conclusion of the empirical study of the SAVVY project not discussed here was also that ignoring CEs appeared to be worse than assuming constant hazards. This illustrates the importance of careful consideration of CEs when aiming at properly estimating and comparing AE risks.

Analysis of safety in key guidelines

To understand the landscape of existing guidelines on safety reporting which, ideally, at some point will be updated based on the conclusions from SAVVY, we reviewed an opportunistic sample of these and collected results in Table 4. Many of these guidelines mention at least that varying follow-up times may be relevant to quantify AE risk. For example, the CIOMS Working Group VI handbook [18], which forms the basis of several guidance, already admits that "ICH Guideline E3 mentions survival analysis methods for analysing safety data, but it appears that this has often not been followed in practice". Looking at Table 4 two aspects are remarkable: first, the heterogeneity in how related guidelines treat the issues of varying follow-up, use of the incidence density, constant hazard assumption for the latter, proposing life-table or one minus Kaplan-Meier techniques; second, the complete absence of any consideration of CEs, although at least death seems to be quite an obvious CE in estimation of AE risk, with many others potentially relevant. Taken together with the frequent mentioning of life-table/one minus Kaplan-Meier methods to account for varying follow-up time, this is specifically concerning given the findings in Stegherr et al. [6] and Rufibach et al. [7] about the bias of one minus Kaplan-Meier in the presence of CEs.

Overall, it appears somewhat surprising that all guidelines exhibit awareness of the varying follow-up time issue and even discuss potential remedies, but that in routine reporting of safety and quantification of AE risks the incidence proportion appears still to be the overwhelmingly preferred approach.

Implementation

All methods introduced in the SAVVY project have been implemented in the R package savvyr [23]. This package is available from CRAN.

Discussion

Main conclusion of the SAVVY project

The main conclusion from the SAVVY project is that commonly used methods such as incidence proportions, incidence densities (with and without ignoring CEs), and one minus Kaplan-Meier are all biased and therefore inappropriate to quantify AE risk in the presence of varying follow-up times, CEs, and censoring. Estimators are biased for estimation for absolute as well as relative AE risks. It is important to note that this bias is a statistical property of any of these estimators and independent of the purpose we use any of these estimators for, i.e., whether we quantify the risk for a prespecified or emerging AE, or estimate AE risk in a given therapeutic area, or want to detect a different AE signal between two treatment arms. Taking together, Stegherr et al. [5], Stegherr et al. [6], and Rufibach et al. [7] provide theoretical (i.e., based on simulated) as well as empirical (based on data from 17 RCTs) evidence for the bias of all estimators apart from the gold-standard AJE, and also quantify this bias. This supports decade-long theoretical investigations into the bias of, e.g., the one minus Kaplan-Meier estimator in this field. We are of the strong opinion that all relevant stakeholders, among them clinicians, statisticians, and regulators, should collaborate to finally implement fit-for-purpose methods to quantify AE risk, and update pertinent guidelines. In our opinion, the implementation of the ICH E9(R1) estimands addendum—so far primarily for efficacy—offers a window of opportunity to push for a change also in reporting safety information. In drug development, safety contributes as much to the benefit-risk assessment of a medicine as efficacy, so the same estimand, estimation, and reporting standards should apply to both.

SAVVY—template for sharing of summary data

A special feature of SAVVY was the way data from the 17 RCTs was shared and analyzed: In a big collaborative effort, data had been gathered within 10 sponsor organizations (nine pharmaceutical companies and one academic trial center). In order to avoid challenges with data sharing, SAVVY used an approach familiar from Health Informatics, see, e.g., Budin et al. [24]. A standardized data structure was defined [10] based on which SAS and R macros were developed by the academic project group members. These macros where then shared with all participating sponsor organizations and run by them locally on their individual trial data. Only aggregated data necessary for meta-analyses were forwarded to the academic group members to centrally run meta-analyses.

This appears to be an approach that could also be applied in other applications, as long as the analysis of interest can be done on summary data provided by single organizations.

Limitations

There may be trials where varying follow-up times and/or competing events do not necessarily (relevantly) bias the estimation of AE risks, for example, in trials with complete and identical follow-up for all patients. In such cases, the incidence proportion might be a *good enough* estimator.

AJE is clearly identified as the most unbiased estimator of AE risk in the presence of varying follow-up times and CEs, which is already clear from its theoretical properties. Theoretically, compared to a parametric counterpart (e.g., an estimator assuming both, the AE and CE hazard are constant), as a completely nonparametric estimator AJE will have larger variance. However, as discussed in Stegherr et al. [5], the increase in variance is small.

Next steps for the SAVVY project and the analysis of safety data

Work within the SAVVY project continues. Concrete plans are an analysis restricted to the oncology trials within the 17 RCTs, discussing in more detail the issue of CEs in a typical oncology clinical trial. Collaborations with clinicians in other therapeutic areas to define AEs of interest for which "proper" estimation of their risk would be informative and what clinical events to define as competing are envisaged.

SAVVY's long-term vision is indeed to further familiarize trialists with the AJE and have this method be recommended in future revisions of pertinent guidelines. This in connection with developing pragmatic approaches how to properly identify and define CEs in therapeutic areas.

Abbreviations

AE	Adverse event
CE	Competing event
SAVVY	Survival analysis for AdVerse events with VarYing follow-up times
HR	Hazard ratio
RR	Relative risk
RCT	Randomized clinical trial

Acknowledgements

We thank Thomas Künzel and Daniel Sabanés Bové for implementing earlier code into the R package savvyr [23].

Authors' contributions

KR, JB, TF, CS, and RS conceived the idea for article and drafted it. All authors critically reviewed the manuscript and approved its final version.

Funding

Not applicable.

Availability of data and materials

Individual trial data analyses were run within the sponsor organizations using SAS and R software provided by the academic project group members. Only aggregated data necessary for meta-analyses were shared and meta-analyses were run centrally at the academic institutions. The SAVVY project has a webpage: https://numbersman77.github.io/savvy. Methods are implemented in the R package savvyr [23]. This package is available from CRAN.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

KR is employee of F. Hoffmann-La Roche (Basel, Switzerland). JB has received personal fees for consultancy from Pfizer and Roche, all outside the submitted work. TF has received personal fees for consultancies (including data monitoring committees) from Bayer, Boehringer Ingelheim, Janssen, Novartis, and Roche, 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. RS declares no competing interests.

Author details

[1]F. Hoffmann-La Roche, Basel, Switzerland. [2]Institute of Statistics, Ulm University, Ulm, Germany. [3]Department of Medical Statistics, University Medical Center Göttingen, Humboldtallee 32, 37073 Göttingen, Germany. [4]Clinical Trials Unit, Faculty of Medicine and Medical Center, University of Freiburg, Freiburg im Breisgau, Germany. [5]Charité - Universitätsmedizin Berlin, Corporate Member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Institute of Biometry and Clinical Epidemiology, Berlin, Germany.

Received: 27 February 2024 Accepted: 21 May 2024
Published online: 31 May 2024

References

1. O'Neill RT. Statistical analyses of adverse event data from clinical trials: Special emphasis on serious events. Drug Inf J. 1987;21:9–20.
2. Proctor T, Schumacher M. Analysing adverse events by time-to-event models: the CLEOPATRA study. Pharm Stat. 2016;15(4):306–314. https://doi.org/10.1002/pst.1758.
3. Aalen OO, Johansen S. An empirical transition matrix for non-homogeneous Markov chains based on censored observations. Scand J Stat. 1978;5(3):141–50.
4. Allignol A, Beyersmann J, Schmoor C. Statistical issues in the analysis of adverse events in time-to-event data. Pharm Stat. 2016;15:297–305.

5. 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. Pharm Stat. 2021;20(6):1125–46.
6. Stegherr R, Schmoor C, Beyersmann J, Rufibach K, Jehl V, Brückner A, et al. Survival analysis for AdVerse events with VarYing follow-up times (SAVVY) – estimation of adverse event risks. Trials. 2021;22(420). https://doi.org/10.1186/s13063-021-05354-x.
7. Rufibach K, Stegherr R, Schmoor C, Jehl V, Allignol A, Boeckenhoff A, et al. Comparison of adverse event risks in randomized controlled trials with varying follow-up times and competing events: results from an empirical study. Stat Biopharm Res. 0(0):1–14. https://doi.org/10.1080/19466315.2022.2144944.
8. Gooley TA, Leisenring W, Crowley J, Storer BE. Estimation of failure probabilities in the presence of competing risks: new representations of old estimators. Stat Med. 1999;18:695–706.
9. International Council For Harmonisation of Technical. Requirements For Pharmaceuticals For Human Use (ICH). ICH E9(R1) addendum on estimands and sensitivity analysis in clinical trials to the guideline on statistical principles for clinical trials. Step4_Guide line_2019_1203.pdf. 2019.
10. Stegherr R, Beyersmann J, Jehl V, Rufibach K, Leverkus F, Schmoor C, 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;63:650–70. https://doi.org/10.1002/bimj.201900347.
11. 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 27 May 2024.
12. CIOMS Working Groups III and V. Guidelines for preparing core clinical-safety information on drugs. Geneva: Council for International Organizations of Medical Sciences.
13. IQWiG. General Methods, Version 5.0. Institute of Quality and Efficiency in Health Care. https://www.iqwig.de/en/methods/methods-paper.3020.html. Accessed 27 May 2024.
14. Latouche A, Allignol A, Beyersmann J, Labopin M, Fine JP. A competing risks analysis should report results on all cause-specific hazards and cumulative incidence functions. J Clin Epidemiol. 2013;66(6):648–53.
15. Schemper M, Smith TL. A note on quantifying follow-up in studies of failure time. Control Clin Trials. 1996;17:343–6.
16. Clark TG, Altman DG, De Stavola BL. Quantification of the completeness of follow-up. Lancet. 2002;359:1309–10.
17. Varadhan R, Weiss CO, Segal JB, Wu AW, Scharfstein D, Boyd C. Evaluating health outcomes in the presence of competing risks: a review of statistical methods and clinical applications. Med Care. 48:S96–105. https://doi.org/10.1097/MLR.0b013e3181d99107.
18. CIOMS Working Groups VI. Management of Safety Informationfrom Clinical Trials; 2005. Geneva: Council for International Organizations of Medical Sciences.
19. ICH Harmonised Tripartite Guideline. Structure and content of clinical study reports E3. 1995. https://database.ich.org/sites/default/files/E3_Guideline.pdf. Accessed 27 May 2024.
20. ICH Harmonised Tripartite Guideline. Statistical principals for clinical trials E9. 1998. https://database.ich.org/sites/default/files/E9_Guideline.pdf. Accessed 27 May 2024.
21. U S Food and Drug Administration. Premarketing Risk Assessment. https://www.fda.gov/media/71650/download. Accessed 27 May 2024.
22. Junqueira DR, Zorzela L, Golder S, Loke Y, Gagnier JJ, Julious SA, Harms CONSORT, et al. statement, explanation, and elaboration: updated guideline for the reporting of harms in randomized trials. J Clin Epidemiol. 2022;158:149–65.
23. Kuenzel T, Rufibach K, Stegherr R, Sabanés Bové D. savvyr: Survival Analysis for AdVerse Events with VarYing Follow-Up Times. R package version 0.1.0. https://CRAN.R-project.org/package=savvyr. Accessed 27 May 2024.
24. Budin-Ljosne I, Burton P, Isaeva J, Gaye A, Turner A, Murtagh MJ, et al. DataSHIELD: an ethically robust solution to multiple-site individual-level data analysis. Public Health Genomics. 2015;18(2):87–96.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.