MAIN PAPER

On estimands and the analysis of adverse events in the presence of varying follow-up times within the benefit assessment of therapies

Steffen Unkel[1] | Marjan Amiri[2] | Norbert Benda[3] | Jan Beyersmann[4] | Dietrich Knoerzer[5] | Katrin Kupas[6] | Frank Langer[7] | Friedhelm Leverkus[8] | Anja Loos[9] | Claudia Ose[2] | Tanja Proctor[10] | Claudia Schmoor[11] | Carsten Schwenke[12] | Guido Skipka[13] | Kristina Unnebrink[14] | Florian Voss[15] | Tim Friede[1]

[1]Department of Medical Statistics, University Medical Center Goettingen, Goettingen, Germany

[2]Center for Clinical Trials, University Hospital Essen, Essen, Germany

[3]Biostatistics and Special Pharmacokinetics Unit, Federal Institute for Drugs and Medical Devices, Bonn, Germany

[4]Institute of Statistics, Ulm University, Germany

[5]Roche Pharma AG, Grenzach, Germany

[6]Bristol-Myers Squibb GmbH & Co. KGaA, München, Germany

[7]Lilly Deutschland GmbH, Bad Homburg, Germany

[8]Pfizer Deutschland GmbH, Berlin, Germany

[9]Merck KGaA, Darmstadt, Germany

[10]Institute of Medical Biometry and Informatics, University of Heidelberg, Heidelberg, Germany

[11]Clinical Trials Unit, Faculty of Medicine and Medical Center, University of Freiburg, Freiburg im Breisgau, Germany

[12]Schwenke Consulting: Strategies and Solutions in Statistics (SCO:SSIS), Berlin, Germany

[13]Institute for Quality and Efficiency in Health Care, Cologne, Germany

[14]AbbVie Deutschland GmbH & Co. KG, Ludwigshafen, Germany

[15]Boehringer Ingelheim Pharma GmbH & Co. KG, Ingelheim, Germany

Correspondence
Steffen Unkel, Department of Medical Statistics, University Medical Center Goettingen, Humboldtallee 32, 37073 Goettingen, Germany.
Email: steffen.unkel@med.uni-goettingen.de

The analysis of adverse events (AEs) is a key component in the assessment of a drug's safety profile. Inappropriate analysis methods may result in misleading conclusions about a therapy's safety and consequently its benefit-risk ratio. The statistical analysis of AEs is complicated by the fact that the follow-up times can vary between the patients included in a clinical trial. This paper takes as its focus the analysis of AE data in the presence of varying follow-up times within the benefit assessment of therapeutic interventions. Instead of approaching this issue directly and solely from an analysis point of view, we first discuss what should be estimated in the context of safety data, leading to the concept of estimands. Although the current discussion on estimands is mainly related to efficacy

evaluation, the concept is applicable to safety endpoints as well. Within the framework of estimands, we present statistical methods for analysing AEs with the focus being on the time to the occurrence of the first AE of a specific type. We give recommendations which estimators should be used for the estimands described. Furthermore, we state practical implications of the analysis of AEs in clinical trials and give an overview of examples across different indications. We also provide a review of current practices of health technology assessment (HTA) agencies with respect to the evaluation of safety data. Finally, we describe problems with meta-analyses of AE data and sketch possible solutions.

KEYWORDS

adverse events, benefit assessment, clinical trials, estimands, safety data

1 | SETTING THE SCENE

The analysis of adverse events (AEs) is a key component in the assessment of a drug's safety profile. Inappropriate analysis methods may result in misleading conclusions about a therapy's safety and consequently its benefit-risk ratio. A variety of methods are available for the analysis of AEs, but their complexity and the imposed assumptions clearly differ. The simplest methods for contingency tables,[1] eg, naïve proportions and derived effect measures such as risk differences, relative risks, and odds ratios, presuppose identical follow-up times and usually ignore recurrent events and competing risks. If follow-up times are different across treatment groups, then comparisons based on simple incidence proportions produce biased results. Consideration of varying follow-up times by means of incidence densities is possible. However, the incidence density relies on a rather restrictive constant hazard assumption. Standard procedures for event times (see, eg, Collett[2]), in turn, are based on noninformative censoring and are not readily suitable for recurrent events and competing risks. For this purpose, more complex methods in the area of event time analysis do exist (see, eg, Collett[2, chapters 12-13]); for whose application, however, the data must in turn meet the corresponding prerequisites.

The purpose of the present paper is to address the research gap in the analysis of AE data in the spirit of the current discussion on clinical trial estimands. Instead of approaching the problem of investigating AEs directly and solely from an analysis point of view, we discuss which quantities should be estimated in the context of safety data, leading to the concept of safety estimands. Although the current debate on estimands and their role in clinical trials appears to be mainly related to efficacy evaluation,[3,4] the concept is applicable to safety endpoints as well. Only very recently, this perspective has found its way into publications such as Akacha et al.[5] Within the framework of estimands, we present estimation functions for estimating the quantities of interest, that is, statistical methods that map the AE data to a single value. In this context, we also describe problems related to meta-analyses of AE data and sketch possible solutions. Finally, we provide a review of current practices of health technology assessment (HTA) agencies with respect to the evaluation of safety data.

An AE is any unfavourable and unintended sign including an abnormal laboratory finding, symptom, or disease temporarily associated with the exposure to an investigational product, whether or not considered related to the product.[6] The term "treatment emergent" is often added to an AE as a modifier in order to remove manifestations of preexisting conditions from consideration.[7] AEs are documented by the investigator and coded with the Medical Dictionary for Regulatory Activities (MedDRA), which provides clinically validated medical terminology (https://www.meddra.org/). The MedDRA includes symptoms, diseases, diagnoses, investigation names and qualitative results, medical and surgical procedures, and social and family history. AEs are coded with the "lowest level terms." These are combined for the analyses to so-called preferred terms. The latest MedDRA version 21.0 contains more than 22 000 preferred terms. As the present paper focuses on methods for analysing AEs, we do not report further on definitions related to AEs or standards for the collection and documentation of AE data.

The remainder of the paper is organized as follows. To begin with, our work is motivated in Section 2 by giving a brief overview of examples with respect to the analysis of AE data in clinical trials in different indications. We also provide a review of current practices of HTA agencies with respect to the evaluation of AE data. The estimand framework is established in Section 3. We discuss what should be estimated in the context of safety data both from a regulatory context and from an HTA perspective. In Section 4, we provide statistical methods for analysing AEs with the focus being on the

TABLE 1 Some examples from early benefit assessments with considerably different follow-up times[a]

Dossier evaluation	Intervention	Control	Ratio of follow-up times
Oncology	Median follow-up + safety follow-up		
A14-48 prostate	16.6 mo + 28 d	4.6 mo + 28 d	31%
A15-17 lung	336 + 28 d	105 + 28 d	37%
A15-33 melanoma	168 + 90 d	63 + 90 d	59%
A16-04 mantle cell lymphoma	14.4 mo + 30 d	3.0 mo + 30 d	26%
Hepatitis C	Planned follow-up + safety follow-up		
A14-44	8 to 12 wk + 30 d	24, 28, or 48 wk + 30 days	23% to 57%
A16-48	12 wk + 30 d	24 wk + 30 d	57%

[a] The ratio of the follow-up times in the fourth column is computed by dividing the follow-up time of the treatment group with shorter follow-up by the follow-up time of the group with longer follow-up. The dossier assessments can be obtained from https://www.iqwig.de.

time to the occurrence of first AEs. We also describe some problems and their solutions for meta-analyses of AE data. In Section 5, we state some recommendations which estimators fit best to the described estimands. A discussion in Section 6 concludes the paper.

2 | CURRENT PRACTICE OF REGULATORY AND HTA AGENCIES

2.1 | Examples

Based on the experience with early benefit assessments by the Institute for Quality and Efficiency in Health Care (IQWiG) in Germany, we looked over some examples whether variable follow-up times for AEs between individual patients, treatment groups, and studies are common across different indications.

In general, trials with a primary time-to-event endpoint usually have variable follow-up times for each individual patient. However, depending on the indication the average follow-up time between treatment groups can be very different; see Table 1.

In oncology, study treatment is often given until disease progression only with limited follow-up time for AEs after discontinuation of trial treatment due to subsequent therapies, resulting in differential follow-up times for the different treatment arms and censored observations for the occurrence of AEs. Due to the dependency of follow-up time of AEs on progression-free survival, the treatment group with longer progression-free survival has a higher likelihood of observing an AE. In such situations, a simple comparison of incidence proportions between arms is biased in favour of the inferior treatment. In time-to-event trials with average follow-up times not considerably different between treatment groups, there can still be variable follow-up times for individual patients. Examples are large trials in cardiovascular disease, metabolism (diabetes), or respiratory disease (COPD) with cardiovascular outcomes or mortality as primary endpoint, where mortality is relatively low.

Other trial designs than time-to-event trials with variable follow-up times were identified in infectious diseases and central nervous system disorders. For example, there are several trials with planned trial length in Hepatitis C that allowed shortening the treatment time either for all patients in the experimental arm or for those in the experimental arm who achieved an early response. Therefore, the follow-up time in the arm with the experimental drug was shorter than in the control group; see Table 1. Here, the follow-up times not only differ between the treatment groups but also between studies of a drug in the same indication (eg, Hepatitis C). Comparison of AEs between treatment groups can be undertaken via incidence proportions only for the duration of the shorter treatment. It is not possible to demonstrate a potential advantage in terms of lower AE probabilities over a longer period of time. On the other hand, safety evaluations using naïvely all AEs on treatment are biased in favour of the treatment group with shorter treatment duration. For instance, in multiple sclerosis, trials with fixed treatment duration have frequently been used, but if control patients switch to treatment in an extension study, this would lead to longer follow-up in the experimental group and AE probabilities that are biased in favour of the control group. Figure 1 illustrates different scenarios of typical AE follow-up periods in clinical trials.

AEs are assessed on a regular basis at the visit at the beginning of the trial (V0) and during treatment (V1, … ,Vn). The end of treatment (EoT) triggers a safety follow-up visit (Saf-FU) marking the last regular safety assessment. AEs occurring on treatment or during safety follow-up are analysed as treatment-emergent AEs (TEAEs, marked by bold symbols). First occurrences of AEs (marked by triangles) are in general considered only when occurring during the TEAE period. Serious AEs may be reported spontaneously after the safety follow-up visit.
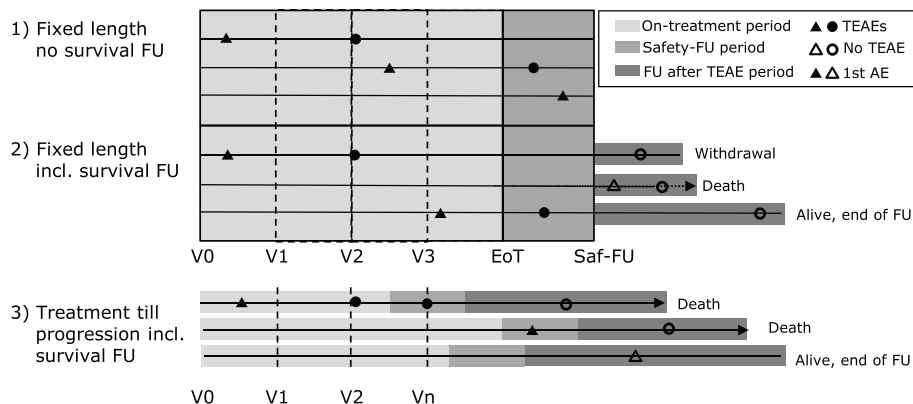
FIGURE 1 Description of different scenarios for typical adverse event (AE) follow-up (FU) in clinical trials (EoT, end of treatment; Saf-FU, safety follow-up; TEAEs, treatment emergent AEs [marked by bold symbols]; V0, visit at the beginning of the trial; V1, … ,Vn, visits during treatment). First occurrences of AEs are marked by triangles

2.2 | Regulatory context

Drug approval usually requires a confirmatory proof of efficacy in at least two well conducted randomized controlled trials followed by a benefit-risk assessment that shows that the treatment's benefit outweighs the expected risk associated with the new treatment. In contrast to the demonstration of efficacy as compared with a control, the benefit risk assessment is far less standardized with respect to properly balancing benefit and expected side effects. Nevertheless, details on clinical trial specifications with respect to the investigated population, the study duration and the number of patients to be studied are given in several regulatory guidelines, as in therapeutic EMA guidelines (the *CHMP clinical efficacy and safety guidelines*), ICH guidelines,[8-11] and U.S. Food and Drug Administration (FDA) guidelines.[12]

Safety assessments in drug approval are usually done with respect to the number and proportion of patients with specific AEs that occurred in the individual clinical trials, primarily focusing on the estimated probability of experiencing a given event within the predefined study duration potentially stratified in relevant subpopulations. It is based on a limited database available at the marketing authorization application where uncertainties about important risks may either prevent from authorization or imply the requirement of postauthorization safety investigations. Due to a number of different side effects, and because the absence of evidence of an increased risk is not necessarily evidence of absence of an increased risk,[13] the analysis of AEs is often crude and merely descriptive aiming to detect safety signals, although proposals have been made in the literature on the use of more sophisticated methods.[14-16] Study duration may be too short or different studies may have different durations leading to different rates. In addition the database from phase III studies may be insufficient to detect infrequent but serious events. In case of uncertainty, the EMA's Pharmacovigilance Risk Assessment Committee (PRAC) may require that a postauthorization safety study (PASS) be carried out after a medicine has been approved. Nevertheless, difficulties in the assessment at the time of marketing authorization are highly relevant especially in indications with a small patient population and a high unmet medical need imposing high pressure to the regulatory system.

Hence, an assessment of safety often targets the probability of an AE within a subject and a given period of time. The safety of a new treatment is assessed using all data available for this treatment. The FDA, eg, asks for a document called integrated summary of safety (ISS) as described in U.S. FDA.[17] In contrast to efficacy assessments, comparative safety assessments are difficult and the appraisal of side effects is usually done in absolute terms. Even if relative (comparative) safety assessments are given and relate to individual studies, accounting for different observational times due to study discontinuation, other events and the presence of changing individual proneness over time is difficult. Competing risks for a number of targeted adverse effects add another layer of complexity. For example, in two trials involving patients with type 2 diabetes and an elevated risk of cardiovascular disease, patients treated with canagliflozin had a lower risk of cardiovascular events than those who received placebo but a greater risk of amputation of toes, feet, or legs with canagliflozin than with placebo (6.3 vs 3.4 participants with amputation per 1000 patient-years, corresponding to a hazard ratio [HR] of 1.97; 95% confidence interval [CI], 1.41-2.75).[18] In such a situation, the increased risk of amputation might be difficult to determine because of the much larger risk of mortality in comparison.

2.3 | Health technology assessments

2.3.1 | The international perspective

The evaluation of safety is considered as an important element of HTA.[19] However, due to different approaches to HTA in different health care systems and little methodological guidance given generally,[20] the integration of safety data and with it, the analyses and interpretation of safety data differs considerably.

Some health care systems, such as the one in England, focus on the economic value of a new technology by implementing a cost effectiveness threshold. Those are usually based on an incremental cost effectiveness ratio (ICER) or quality-adjusted life years (QALY) by comparing against the standard of care.[21]

In this context data on AEs that are considered most relevant from the patients' quality of life (QoL) and/or costing perspective are usually integrated by means of utility functions.[22] A comprehensive review of the current practice of economic models concludes that there appears to be an implicit assumption within modelling guidance that adverse effects are very important. There appears to be a lack of clarity how they should be dealt with and considered for modelling purposes.[23]

Other health care systems, such as the one in Germany, base their decision on the incremental medical benefit against the standard of care.[24] Usually, data from clinical trials are used in order to evaluate the added benefit for all patient relevant endpoints separately to demonstrate an overall added benefit of the drug for the population in scope. Typically, a comprehensive description of AEs is provided in those assessments. However, no specific guidelines with respect to safety analyses are in place in countries following this approach. In the following, we discuss one specific example.

2.3.2 | The current practice at IQWiG in Germany

At the beginning of 2011, the early benefit assessment of new drugs was introduced in Germany with the Act on the Reform of the Market for Medicinal Products (AMNOG). The Federal Joint Committee (G-BA) generally commissions the Institute for Quality and Efficiency in Health Care (IQWiG) with this type of assessment, which examines whether a new therapy shows an added benefit (a positive patient-relevant treatment effect) over the current standard therapy. The IQWiG is required to assess the extent of added benefit on the basis of a dossier submitted by the pharmaceutical company responsible. In this assessment, the qualitative and quantitative certainty of results within the evidence presented, as well as the size of observed effects and their consistency, are appraised. The general methods of IQWiG are described in the Allgemeine Methoden.[24, Version 5.0] In accordance with §35b (1) Sentence 4 Sozialgesetzbuch (SGB) V, the following outcomes related to patient benefit are to be given appropriate consideration: increase in life expectancy, improvement in health status, and quality of life, as well as reduction in disease duration and adverse effects. In the benefit assessment, all patient-relevant endpoints play a role, and for safety, particular consideration is given to serious and severe AEs as well as treatment discontinuations. In addition, AEs that are of special interest within the context of the disease or drug class considered may play a role.

For the assessment of the extent of added benefit, the effect sizes are of main interest. An effect size in this context is defined as the (relative) difference between the new treatment and the appropriate comparator therapy. It is an important step to grade the qualitative certainty of the estimated effect size, eg, based on trial design, data quality, and estimation method. Patients not included into the analyses and patients with incomplete data increase the risk of bias. Therefore, the following information is gathered and considered to assess the risk of bias: study design (randomized, open-label, or double-blind), proportion of patients without consideration in the analyses per study arm, proportion of patients with incomplete data per study arm (censored data, lost-to-follow-up, …), reasons for censoring (informative, noninformative, and competing risks), and distribution of censoring times. Generally, this information is used to assess the direction (in favour of arm x) and the strength (low or high) of the risk of bias.

In case of varying follow-up times, methods based on survival time analyses are preferred compared with analyses based on four-fold contingency tables.[15] To classify the risk of bias, the number of the censored patients and the reasons for censoring have to be considered. Different types of censoring are possible: "uninformatively censored," "patients with competing risks," and "informatively censored," which influence the risk of bias. We will elaborate further on this issue in Section 4.3.

3 | ESTIMANDS

3.1 | Framework

It is paramount to agree upon the relevant target of estimation defined by the question what would happen to a specific patient or what is the patient's risk with respect to a specific event or multiple events when treated with a given drug as compared with another drug or to not being treated at all. In the context of *efficacy* assessments, this concept has recently been introduced within the framework of *estimands* in the new draft addendum R1 to the ICH E9 guideline on statistical principles in clinical trials entitled *Estimands and Sensitivity Analyses in Clinical Trials*; this addendum is referring to the precise parameter or function of parameters to be estimated in situations where *intercurrent events* as treatment discontinuation, death, rescue medication, or switch to the other study treatment may influence subsequent measurements.[7,25] We would like to stress the fact that the above-mentioned addendum is not yet finalized; hence, our expositions in the sequel can only reflect the current state of discussion. Parts of the draft addendum are seen critically by HTA agencies,[26] and the addendum does not prescribe the use of a particular estimand in a certain situation. Moreover, the discussion of the application of the estimands approach to safety data and questions related to benefit-risk assessments is ongoing and only started recently.

Four different elements are required to describe the estimand of interest: the *targeted population*, the *endpoint* (variable), the *intervention effect* that describes how intercurrent events that potentially influences the endpoint are accounted for, and the *summary measure* that summarize the comparison of the two treatments under investigation.

Whereas the nomenclature of types of estimands has been developed and changed during the last few years, currently, the following classes of estimands are discussed in the regulatory context:

- **Treatment policy**: *treatment policy estimand* do not account for any intercurrent event. The treatment effect is measured irrespective of any intercurrent event, as treatment discontinuation or additional medication given.
- **Composite**: *composite estimands* combine the variable of interest with the intercurrent event, eg, by defining a treatment failure by the lack of response or treatment discontinuation.
- **Hypothetical**: *hypothetical estimands* target an effect that would occur in the overall population in a hypothetical scenario, in which no patient experienced the intercurrent event. For example, the effect of all patients adhering to treatment constitutes a hypothetical effect when some patients in fact do not adhere to treatment.
- **Prinicpal stratum**: *principal strata estimands* are defined by the subset of patients in whom the intercurrent event occurs either under one of the treatments or under both. Because a group comparison trial cannot directly identify these patients with respect to the not-administered treatment, causal inference methods using specific assumptions would be required for the analysis.
- **While on treatment**: *while on treatment estimands* relate to the effect prior to the occurrence of an intercurrent event, eg, before intake of rescue medication or the effect while being alive.

Although the current discussion is related to the *efficacy* evaluation, the concept is applicable to *safety* endpoints, usually the occurrence of a specific side effect, as well. Considering the variable of interest as the time to a specific side effect, summary measures might be given after a specific period of time. Relevant *intercurrent events* are treatment discontinuation or switch, death, or other side effects that may prevent from the event of interest.

Whereas the basic idea of an estimand is not restricted to the efficacy assessment, different issues related to the large number of event types and the desired "equivalence proof" combined with the cautionary principle point to somewhat different difficulties. Envisaging the chances of a beneficial treatment in the sense of both, efficacy and safety, for a given patient may be seen as a concept of incorporating efficacy and safety in an estimand but may be still difficult to be interpreted in the comparison of two treatments with different safety and efficacy patterns. In that sense, it appears sensible to go along the lines that are conceived for the efficacy assessment and clarify the precise parameters in the event analysis in the presence of other concurring events either in relation to the given treatment, the patient's condition, or competing side effects.

3.2 | Estimands in the regulatory context and in HTAs—What to aim for?

The estimand framework is not specific to the clinical development of new drugs in the regulatory context but is also relevant to address needs of HTA bodies. The aim of the HTA process is the assessment of evidence as a basis for further decisions about reimbursement, pricing and market access.

Estimands are supposed to focus and describe the research question in detail. As the aims of drug approval agencies and HTA bodies differ to some extent, different estimands may be of primary interest, but some overlap in secondary considerations can be expected. The following considerations focussing on HTA may also apply to the benefit risk assessment in drug approval. In the German early benefit assessment according to AMNOG, ie, § 35a SGB V, there is a need for the HTA authorities to identify the estimands, which are not necessarily the same as in the regulatory context to obtain marketing authorization.

For marketing authorization the current practice, in general, is to report estimates for what could be considered while on treatment estimands to provide evidence for the safety profile of the treatment of interest. Specific information, however, on AEs occurring in a long-term follow-up regardless of study drug adherence may be requested in special cases, which may, however, be hampered by the limited observational period. Potentially diluting effects of treatment policy estimands in case of treatment discontinuation or switch may, depending on the treatment comparison and the disease, be anticonservative for comparative safety assessments, hence favouring while on treatment estimands for regulatory purposes. Certainly, within the context of recent regulatory discussion on estimands in efficacy, a new regulatory framework also on estimands in safety is needed.

HTA bodies are most interested in the *treatment policy estimand*, independently of occurrences like rescue therapy (eg, rheumatoid arthritis) or subsequent therapies (eg, oncology). However, in indications like oncology, the AEs are frequently only collected up to a certain point after last dose of study medication. These data do not support a treatment policy estimand. In such situations, the evidence for risk assessment is less strong. It may be difficult or impossible to cover the treatment policy estimand with the data usually obtained. To obtain sufficient data for a treatment policy estimand and provide a solid basis for an early benefit assessment, the current practice of how data are collected in clinical studies needs to be changed.[15]

A major challenge in oncology is the treatment change after progression of the disease, where in many cases patients enter a subsequent clinical study, eg, in malignant melanoma where about 4 years ago, the only treatment option was dacarbacin, and most patients entered clinical trials after progression. These studies were under evaluation by the HTA bodies recently due to the time gap between study conduct and marketing authorization. However, in most studies, a patient is not allowed to enter a new clinical study, if they are still participants of the prior study. As a consequence, they need to withdraw consent for the first study to enter the next. Therefore, AEs cannot be collected after progression for the first study in general. Exceptions from this practice are observed,[27] following a protocol recording new onset of serious AEs up to 90 days after last dose of study treatment and those serious AEs considered related any time after discontinuation of treatment. Another example consists of the German Society of Pediatric Oncology and Hematology (GPOH),[28] which records further follow-up data on children after EoT and with this provides the possibility for long-term surveillance and follow-up and late effect evaluation in paediatric oncology patients. Apart from these examples, the while on treatment estimand is used in clinical trials to avoid bias due to unbalanced withdrawals in the treatment groups. For example, if the control treatment is a standard first line treatment and a subsequent study in second line requires the standard treatment as first line, then only patients of the control group are allowed to enter the second line study leading to unbalanced subsequent therapy along with biases in efficacy but also safety.

Four different scenarios, which are displayed in Figure 2, can be distinguished to describe safety estimands in an HTA system.

The scenarios in Figure 2 differ according to the lengths of the planned and observed follow-up times in the study. The definition of estimands becomes increasingly complex with more pronounced differences in follow-up time due to intercurrent events. In the first two scenarios, the planned follow-up times of AEs in the study population are similar, eg, in trials with fixed trial lengths (see Section 2.1). Whereas in Scenario 1, there are no or only minor differences in observed follow-up times between individual patients within treatment groups or between treatment groups, in Scenario 2 medium to large differences in observed follow-up times do exist, due to, eg, high level of treatment or study discontinuations. Scenarios 3 and 4 consider studies with expected differences in follow-up times by treatment group, eg, in oncology (see Section 2.1), where AE reporting stops a certain number of days after last dose of study drug and the time on study drug differs between treatment groups. In all four scenarios, HTA bodies usually aim for the treatment policy estimand. However, studies are commonly planned to collect data that are appropriate for the while on treatment estimand, which is usually the focus of regulatory agencies in the marketing authorization process. When benefit dossiers are based on the same studies, it is often not possible to provide estimates of the treatment policy estimand desired by the HTA agencies due to lack of adequate data. Hence, it is apparent that the described requirements by HTA bodies with regard to safety analyses need to be taken into account already in the planning phase of a clinical study.
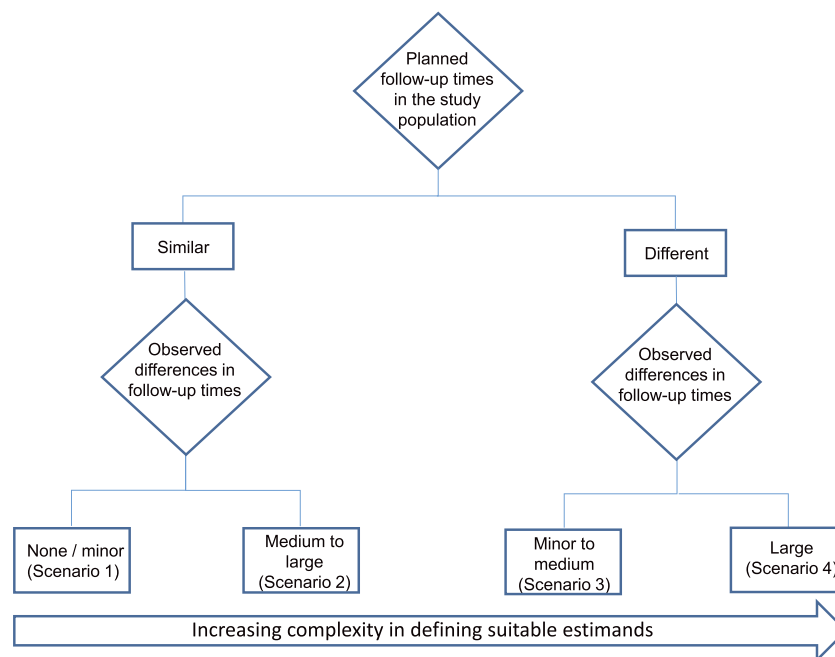
FIGURE 2 Flow chart displaying four different scenarios across indications for the consideration of safety estimands in an HTA system

4 | STATISTICAL METHODOLOGIES

The aim of this section is to discuss methods for *analysing* AEs. We focus on the occurrence of the first AE of a specific type because the statistical considerations for the analysis of the first AE are relevant also for the analysis of recurrent AEs. We found a great variety of methods in the literature, some of which were not well defined, making it difficult to identify both estimator and estimand. Main methods that focussed on AEs occurrence in one group were described in the literature as crude rate, incidence proportion, incidence rate, exposure-adjusted incidence rate, hazard function for AEs, Kaplan-Meier, and cumulative incidence function considering competing risks. We first describe methods of estimation within one treatment group in Section 4.1 and subsequently, the comparison of AE occurrence between samples in Section 4.2. The time point of AE occurrence or comparison thereof will be made explicit. A major issue will be that safety data need not be completely observed over the whole study period for all patients and will possibly be right-censored. This requires the use of time-to-event methodology[14,16] taking different follow-up times into account, which is common in efficacy analyses, but less so for safety. In this context, it is often discussed which kind of censoring is informative, and the role of censoring will briefly be revisited in Section 4.3. Methods for meta-analyses of AE data are discussed in Section 4.4.

4.1 | Methods of estimation within one treatment group

One major common method is the so-called crude rate, which despite its name is in fact a proportion and is defined as $\hat{P}(AE) = a/n$, where a is the number of patients observed to experience at least one AE of a specific type, and n is the total number of study patients, see other studies[14,16,29-31] and references therein. The crude rate is a correct estimator of the probability to experience at least one AE of the interesting type in case of complete data and identical follow-up times for all patients. With different follow-up times in different samples, the crude rate will estimate the AE probability at different points in time. This difficulty is resolved by considering the incidence proportion (Allignol et al[14] and references therein):

$$\hat{P}(AE \text{ in } [0, t]) = \frac{\sum_{u \le t} a_u}{n}, \tag{1}$$

where a_u denotes the number of patients observed to experience at least one AE of a specific type at time u. The expression (1) estimates the probability of experiencing at least one AE within some time interval $[0, t]$ but is again only valid for complete data over the considered time interval. In the presence of censoring, both the crude rate and the incidence proportion *underestimate* the AE probability.[14] The reason is that these methods in fact estimate the probability of *both* the AE occurrence *and* the nonoccurrence of censoring.

Ioannidis et al[30] and Amit et al,[32] therefore, suggest to use one minus the Kaplan-Meier estimator for estimating $P(AE\ in[0,t])$, censoring time to AE by both the end of follow-up and by competing events (CE) that preclude AE occurrence such as death without a prior AE. However, this method *overestimates* the AE probability.[14,33] The reason is that one minus the Kaplan-Meier estimator approximates a cumulative distribution function, implicitly assuming that eventually 100% of all patients experience the AE under consideration, possibly after death. Therefore, one minus the Kaplan-Meier estimator must not be used to estimate $P(AE\ in[0,t])$ in the presence of CEs that prevent the occurrence of the AE under consideration.

It is the Aalen-Johansen estimator[34] that generalizes the Kaplan-Meier estimator to multiple event types and nonparametrically estimates the so-called cumulative incidence function $P(AE\ in\ [0,t])$, accounting for both CEs[35,36] and the usual censoring due to end of follow-up. The Aalen-Johansen estimator for the probability of AE occurrence is (Allignol et al[14] and references therein)

$$\hat{P}(T \leq t, AE) = \sum_{u \leq t} \hat{P}(T > u-) \cdot \frac{a_u}{n_u}, \tag{2}$$

where T is the time until occurrence of an AE or of a CE, $\hat{P}(T > u-)$ denotes the estimate of the probability of not experiencing an AE or the CE just prior to time u, and n_u is the number of patients at risk of observing an AE or a CE just prior to u. The right-hand side of equation 2 is a sum over the empirical probabilities of experiencing an AE at the observed event times. Here, for estimation of $P(T > u-)$, the Kaplan-Meier method is used, because the definition of T encompasses all CEs. Time-to-event analyses are based on hazards, because in general, follow-up times are incomplete. In fact, the sum over the quotients on the right hand side in (2) is the Nelson-Aalen estimator of the cumulative AE hazard $\int_0^t \alpha_{AE}(u)du$. For analysing AEs, the Nelson-Aalen estimator is key in three ways. Firstly, it enters the computation of the Aalen-Johansen estimator. Secondly, it is closely linked to the Mean Cumulative Function, which is based on the Nelson-Aalen estimator and is also used in safety analyses.[29] Thirdly, the Nelson-Aalen estimator is the cumulative nonparametric counterpart of the commonly used incidence rate (or incidence density) of AEs[14,30,31]:

$$IR_{AE} = \frac{a}{\sum t_i}, \tag{3}$$

where t_i is the time at risk for patient i, and $\sum t_i$ denotes the population time (person-years) at risk. The incidence rate (3) is an estimator of the AE hazard $\alpha_{AE}(t)$ under a constant hazard assumption, $\alpha_{AE}(t) = \alpha_{AE}$ for all times t. The incidence rate is popular, because its denominator accounts for varying follow-up times. Sometimes, the exposure-adjusted incidence rate is reported by counting in the denominator only the population time during exposure to study treatment. However, it is *not* a probability estimator (and should better not be reported as a percentage). In fact, it is easily seen that depending on how time is measured (think of milliseconds or decades), the denominator can be made arbitrarily large or small, possibly resulting in values larger than one.

In perfect analogy to the Aalen-Johansen estimator, translating incidence rates into probability statements requires incorporating CEs, eg, death without prior AE. For instance, IR_{AE} and the incidence rate of the CE, $IR_{CE} = c/\sum t_i$ (writing c for the number of patients observed to experience the CE), can be used to obtain a parametric counterpart of the Aalen-Johansen estimator. If constant event-specific hazards are assumed, the cumulative incidence function of the event type AE is explicitly given as

$$\begin{aligned} P(T \leq t, AE) &= \int_0^t \alpha_{AE} \cdot \exp\left(-(\alpha_{AE} + \alpha_{CE})s\right)\ ds \\ &= \frac{\alpha_{AE}}{\alpha_{AE}+\alpha_{CE}} \left(1 - \exp(-(\alpha_{AE} + \alpha_{CE})t)\right), \end{aligned} \tag{4}$$

where α_{CE} denotes the CE hazard. By plugging in both event-specific incidence rates, the cumulative incidence function can be estimated *parametrically* under a constant hazards assumption.

We also note that both the Nelson-Aalen estimator and the incidence rate allow for AEs to be recurrent. In this situation, the translation into probability statements becomes more complex because of the more complicated recurrent events structure, but also with recurrent AEs, CEs have to be taken into account.

4.2 | Comparison of treatment groups

When comparing two treatment groups with respect to AE occurrence, often measures like risk difference, relative risk, or odds ratio of crude rates are suggested (eg, Amit et al[32]). However, if such relative measures are used in the presence of censoring and are based on biased one-sample estimators as discussed above, the result of such a comparison will

be biased too, but the direction of the bias is uncertain. For instance, a ratio of incidence proportions calculated from censored data will divide something too small by something too small.

As a parametric analysis, the ratio of incidence rates is an appropriate estimator of the hazard ratio under a constant hazard assumption. The obvious semiparametric extension is to use a Cox proportional hazards model,

$$\alpha_{AE}(t|Z) = \alpha_{AE;0}(t) \exp(\beta_{AE}^\top Z) , \tag{5}$$

where $\alpha_{AE;0}(t)$ is an unspecified baseline AE hazard, β_{AE} is the vector of regression coefficients, and Z a vector of baseline covariates including treatment group. In other words, if in (5) the only covariate is treatment group, $Z \in \{0, 1\}$, then the ratio of the incidence rate in group 1 to the incidence rate in group 0 estimates the hazard ratio $\exp(\beta_{AE})$ under the assumption of a constant baseline hazard for AEs, $\alpha_{AE;0}(t) \equiv$ constant. If this assumption is in doubt, any Cox regression software *technically* censoring the time to AE by observed CEs will yield the usual maximum partial likelihood estimator of $\exp(\beta_{AE})$. Technically, censoring by observed CEs is in perfect analogy to calculation of the incidence rates, but, again in analogy to the incidence rates, it does not allow for probability statements. In other words, the analysis remains somewhat incomplete without consideration of the hazard of the CE, eg, via a second Cox model,

$$\alpha_{CE}(t|Z) = \alpha_{CE;0}(t) \exp(\beta_{CE}^\top Z) ,$$

which *technically* censors the time to the CE by observed AEs, see Beyersmann et al[37] for a practical in-depth discussion. A reasonable method of choice will be a Cox regression model for the event-specific hazards. The important point is that it requires as many Cox regression models as there are event-specific hazards present. Although fitting two Cox models is straightforward from a computational perspective, the presence of two hazards is not without subtleties.

We want to illustrate this using a toy example, assuming, for ease of presentation, constant hazards. We consider a treatment that modifies the AE hazard by a factor of 0.5 and the CE hazard by a factor of 0.25. As $t \to \infty$, one can see from (4) that P(AE | group 1) in treatment group 1 becomes

$$\frac{0.5 \cdot \alpha_{AE_0}}{0.5 \cdot \alpha_{AE_0} + 0.25 \cdot \alpha_{CE_0}} , \tag{6}$$

where α_{AE_0} and α_{CE_0} denote the AE hazard and CE hazard in group 0, respectively. The expression (6) is greater than P(AE|group 0), although the AE hazard has been reduced. The reason is simple. In our toy example, treatment reduces both hazards, thus, delaying both events. Because the effect is larger on the CE than on the AE, there will eventually be more AEs in treatment group 1 than in treatment group 0, such that the cumulative AE probabilities cross at some point in time. This is illustrated in Figure 3, showing the cumulative AE probabilities in group 0 and group 1 over time for the situation of constant hazards described above.
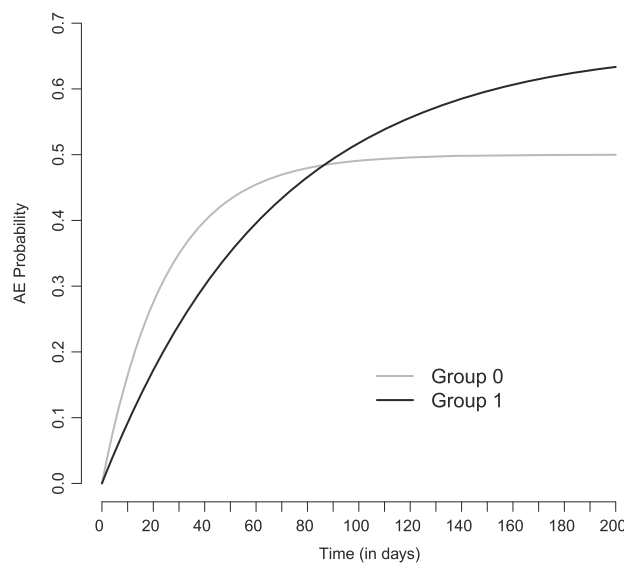

FIGURE 3 Cumulative adverse event (AE) probabilities for two groups and constant hazards. Although in group 1 the AE hazard is lower compared to group 0, the cumulative AE probability in group 1 is eventually greater than in group 0

In group 0, both the AE hazard rate and the CE hazard rate were set to 0.02 events per day, eventually leading to an AE probability of 1/2 in group 0 and of 2/3 in group 1. In group 1, the AE hazard is modified by a factor of 0.5 and the CE hazard by a factor of 0.25.

Because multiple hazards are present and an analysis of only one hazard does not suffice for probability statements, so-called direct approaches such as the Fine and Gray model for the so-called subdistribution hazard[38] or, easier to interpret, the proportional odds cumulative incidence function model[39] have been developed. Most popular is perhaps the Fine and Gray approach, which interprets one minus the cumulative incidence function as a survival function and fits a Cox model to the corresponding hazard, the so-called subdistribution hazard. The approach is useful in that a subdistribution hazard ratio greater (smaller) than one translates into an increase (decrease) of the cumulative incidence function but is otherwise difficult to interpret,[40] because the subdistribution hazard, say $\lambda(t|Z)$, can be expressed as

$$\lambda(t|Z) = \frac{P(T > t|Z)}{1 - P(T \le t, \text{AE}|Z)} \cdot \alpha_{\text{AE}}(t|Z),$$

which results in a complicated mixture of effects on the hazard scale and on probability scales. Alternatives include group comparisons based on confidence bands of the cumulative incidence functions (eg, Beyersmann et al[41]) or the proportional odds cumulative incidence function model[39] mentioned above. The latter is a generalization of the logistic regression model to binomial probabilities $P(T \le t, \text{AE}|Z)$ as a function of time t and in the presence of censoring.

4.3 | Censoring: Independent or informative?

Our presentation so far has demonstrated that the analysis of AE occurrence in trials with varying follow-up times has to account for differences in follow-up, in particular in the form of *censoring*. Survival methodology should not only be used for efficacy but also for safety analyses. However, the safety estimands of interest may still be a matter of debate (see Section 5). Above, we have demonstrated that the relationship between hazards and probabilities is more subtle when CEs that preclude AE occurrence are present. But even when this is accounted for, there may be a choice between, say incidence rates and "exposure-adjusted incidence rates" (which are also incidence rates, but with a different at-risk period, see Section 4.1).

Censoring is a concept that pertains to all these aspects, but it is more subtle than may seem at first glance. So far, we have argued that more standard statistical techniques not from the field of survival analysis are inappropriate because the analysis will be about AEs that are *observed* rather than about AEs that the patient experiences. The observation time of AEs is often restricted, eg, in oncological trials often to progression of disease + 30 days. The patient may experience AEs after this period, but this is not reported in the case report form (CRF). Next, we have found that one minus the Kaplan-Meier estimator censoring the time to AE by CEs overestimates the cumulative AE *probability* but that such a censoring approach yields a valid analysis of the AE *hazard*. Whether or not censoring yields a valid analysis has an impact on the estimand at hand, of course.

When survival methods for AE analysis are discussed, authors often warn against *informative censoring* (eg, Bender et al[15]), but this discussion on censoring is somewhat complicated by inconsistent terminology in the literature. *Random censoring* typically refers to the situation where time to event and time to censoring are independent random variables taking values in $[0, \infty)$ (eg, Aalen et al[42, p30]). This is also called *independent censoring* or *noninformative censoring* (eg, O'Quigley[43]) in the literature, and it is neither uncommon that these last two terms are used interchangeably (eg, Collett[2]) nor that they refer to different censoring mechanisms (eg, Kleinbaum and Klein[44]). This bedevils the discussion both on AE analyses and estimands, because one must first define what is meant by, eg, the term *independent censoring*, because different authors may use the term differently, referring to different censoring mechanisms.

In our context, one will rarely be willing to assume that the time to a certain AE and the time to a CE such as death (or progression) without prior AE are independent and present themselves as an example of random censoring. In fact, and more importantly, it is entirely unclear how to define time-to-AE for a patient who has died as the value of a random variable in the positive real numbers. Such a value would suggest that there is an AE *after death*, which is an awkward concept (to say the least), and we prefer an agnostic point of view.

Above, we have found that observed occurrence of a CE can be regarded as "independent censoring" in the sense that technically, treating it as a censored observation allows for a correct analysis of the AE hazard. In other words, the analysis of the AE hazard *does not depend* on whether censoring was due to administrative closure of the study or whether it was due to a CE. On the other hand, observed occurrence of a CE can be regarded as "informative censoring" in the sense

that technically treating it as a censored observation does not allow for a correct analysis of the AE probability as in a Kaplan-Meier procedure.

These ideas are made rigorous in the counting process approach to survival analysis,[42,45,46] see also Allignol et al[14] for a non-technical account. In a nutshell, censoring by a CE is independent censoring in that it preserves the desired form of the intensity of the counting process of the event under consideration, but it becomes independent yet informative censoring if the target parameter is the cumulative incidence function.

It is worthwhile to reflect on these concepts in oncology trials, where common endpoints are progression-free survival and overall survival. It is not uncommon that recording of AEs is stopped for patients who progress and undergo a second line treatment. Of course, these patients may still experience AEs, and it is generally assumed that the hazard of an AE after progression is different as compared with before progression. Progression then is a CE for AE without prior progression. And progression is a CE for death without prior progression *and* without prior AE. Hence, censoring by the progression event will yield a valid analysis of the hazards of the other two CEs, but any probability statement will need to account for all hazards involved. A different question, however, is what kind of censoring by progression is with respect to AE occurrence *after* progression. In a way, the answer is easy: if censoring by progression events yields a valid analysis of AE occurrence *before* progression, but if the AE hazard *after* progression changes, progression cannot be independent (and, hence, not noninformative) censoring with respect to AE occurrence *after* progression.

The argument can be made rigorously by showing that censoring by progression does not preserve the desired form of the intensity of the AE counting process, if the latter is not restricted to AEs before progression, but the bottom line is obvious: if recording of AEs is stopped for patients with diagnosed progression, inference for postprogression AEs is impossible.

4.4 | Meta-analyses of AE data

When data from more than one study are available, it is not uncommon to naïvely pool the data across the studies by, eg, "simply combin[ing] the numerator events and the denominators for the selected studies."[47] McEntegart[48] and later Rücker and Schumacher[49] as well as Chuang-Stein and Beltangady[50] warned of such naïve pooling as results might be biased due to Simpson's paradox. The International Conference on Harmonization (ICH) E9 states that "any statistical procedures used to combine data across trials should be described in detail" and that "attention should be paid [...] to the proper modelling of the various sources of variation." The use of meta-analysis techniques is encouraged,[51] because these techniques allow for *variation in baseline* (control group) outcomes across the various studies. Random-effects meta-analysis *in addition* allows for *variation in treatment effects* across studies (so-called *between-trial heterogeneity*). Therefore, this type of models is appropriate to formally combine several studies in one analysis. In the context of safety analyses, a number of specific problems arise (see, eg, Berlin et al[52]), some of which will be considered in the following.

Meta-analysis can be carried out using *aggregated data* of the individual studies or, if available, *individual patient data* (IPD). IPD meta-analyses have some advantages over aggregate data meta-analyses,[53] in particular with time-to-event data considered in this manuscript. If time-to-event data are considered and the meta-analysis is based on published data, it is sometimes necessary to reconstruct the data by using appropriate methods.[54,55]

As explained in Section 4.2, effect measures such as the risk difference, relative risk, or odds ratio of crude rates are not appropriate when analysing AE with varying follow-up times. Alternatives include the ratio of incidence rates, which would be appropriate for instance under a constant hazard assumption; the hazard ratio estimated in a Cox proportional hazards model; or the subdistribution hazard ratio of the popular Fine and Gray model. For the purpose of meta-analyses, these effect measures would be log-transformed to estimate a combined effect on the log-scale, eg, log rate ratio or log hazard ratio. Technically, this is fairly straightforward. However, some challenges arise if, for example, the follow-up times vary considerably between the studies. Under assumptions such as proportional hazards the formal combination of studies with different follow-up times are justified. In practice, however, such assumptions might be challenged. Whereas with a single study, the hazard ratio estimated by weighted Cox regression might be interpreted as an average effect over follow-up when the proportional hazards assumption does not hold true,[56] this interpretation in the presence of considerably different follow-up times across studies in a meta-analysis does not apply in the same way. Furthermore, with these arguments, the variation in follow-up times across studies is likely to yield some level of between-trial heterogeneity in treatment effects.

Summarizing AE data from a clinical development programme, typically only a small number of studies is available. Meta-analysis of (very) few studies has recently attracted more attention as it is also frequently accounted in settings other than the one considered here. An overview and discussion of the various methods in the context of benefit assessments

is provided by Bender et al.[57] Specifically, empirical studies demonstrated that between-trial heterogeneity is likely to be present,[58] suggesting the use of random-effects rather than fixed-effect meta-analysis. With *few* studies, however, the *between-trial heterogeneity* is difficult to assess with standard methods for random-effects meta-analysis based on normal approximation yielding CIs for the combined effect which are too short and which have coverage probabilities well below the nominal confidence level. This is due to an underestimation of the between-trial heterogeneity and a failure to account for the uncertainty in the estimation of the heterogeneity. *Bayesian* random-effects meta-analysis with weakly informative prior on the between-trial heterogeneity (and an uninformative prior on the treatment effect) has been suggested for meta-analysis with few studies.[59,60] This avoids zero estimates of the between-trial heterogeneity and accounts for uncertainty in the estimation yielding satisfactory coverage probabilities and interval lengths.[61,62] Application of the DIRECT algorithm,[63] which is faster than MCMC sampling and does not require inspection of convergence diagnostics, means that computations are fairly straightforward. An implementation is available in form of the R package `bayesmeta`, which can be downloaded from CRAN (`https://cran.r-project.org/package=bayesmeta`).

Neal et al[18] report an integrated analysis of two large-scale randomized placebo trials assessing the efficacy and safety of canagliflozin in patients with type 2 diabetes and elevated risk of cardiovascular disease. Patients were followed up for varying lengths of time as the programme was event driven with a constraint on minimum follow-up. In a stratified Cox regression, a beneficial effect of canagliflozin versus placebo on the primary outcome time to death from cardiovascular causes, nonfatal myocardial infarction, or nonfatal stroke, whatever occurred first, was demonstrated (HR = 0.86; 95% CI, 0.75-0.97). For the purpose of illustration, we consider here the AE "low trauma fracture." Figure 4 is a forest plot of the HR with 95% CI from the two studies CANVAS and CANVAS-R.

The figure also includes a fixed-effect meta-analysis, which is in fact very similar to the stratified Cox regression reported by Neal et al.[18] As noted by Neal et al,[18] there was considerable between-trial heterogeneity. Therefore, the use of a random-effects meta-analysis is indicated. The forest plot includes results from three random-effects meta-analyses, modified Knapp-Hartung as a frequentist method suggested for meta-analyses with few studies,[64] and Bayesian random-effects meta-analyses with two choices of the prior for the heterogeneity parameter τ.[62] As can be seen from Figure 4, the modified Knapp-Hartung method yields a very wide, noninformative interval whereas the Bayesian intervals are much shorter. In comparison with the fixed-effect model, the Bayesian intervals are considerably longer as they account for the pronounced between-trial heterogeneity.

Meta-analyses of AE data are further complicated when the events considered are *rare*. Normal approximations of the distributions of effects, eg, log hazard ratios or log odds ratios, break down with low event rates. In particular, if some studies result in zero events in both arms. Then measures such as (log-)odds ratios cannot be calculated. Among the remedies proposed for such problems are continuity corrections[65] and models of the counts such as binomial distributions. The latter can be fitted using likelihood or Bayesian methods.[59,66,67] Very recently, the use of weakly informative priors for the treatment effect as well as for the between-trial heterogeneity has been shown to result in satisfactory properties in meta-analyses with few studies and rare events.[68]
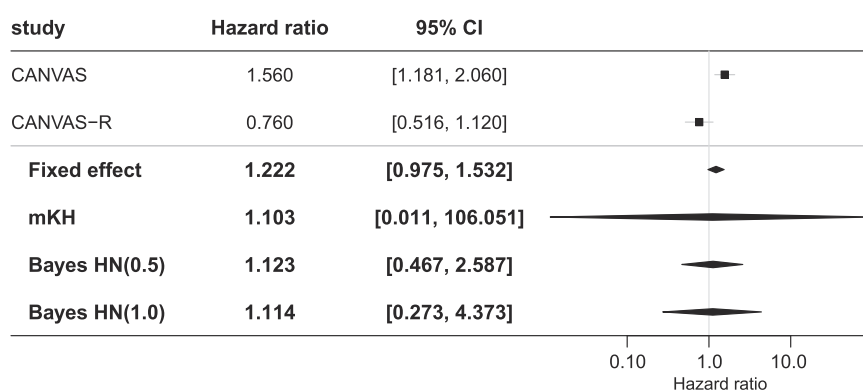
study	Hazard ratio	95% CI	
CANVAS	1.560	[1.181, 2.060]	
CANVAS−R	0.760	[0.516, 1.120]	
Fixed effect	**1.222**	**[0.975, 1.532]**	
mKH	**1.103**	**[0.011, 106.051]**	
Bayes HN(0.5)	**1.123**	**[0.467, 2.587]**	
Bayes HN(1.0)	**1.114**	**[0.273, 4.373]**	

FIGURE 4 Illustrating example for meta-analyses. Forest plot of hazard ratios for low trauma fractures as observed in CANVAS and CANVAS-R with 95% confidence intervals (CIs) and four combined hazard ratios from a fixed-effect meta-analysis, modified Knapp-Hartung (mKH) meta-analysis, and Bayesian random-effects meta-analysis with two half-normal (HN) priors for the heterogeneity parameter τ

5 | ESTIMATORS FOR DESCRIBED ESTIMANDS

In the following, the different statistical methods discussed in the previous section are put into the context of the estimand framework described in Section 3.1 for the specific situation of the analysis of AEs. The decision on the estimand is made on the study level. Although in the following we consider estimands of individual studies, the same principle applies to meta-analyses of AE data. A disadvantage of aggregate data meta-analyses in this context is of course that different estimands might have been used in the studies whereas in IPD meta-analyses, the same or at least similar estimands can be applied to the studies. As with Section 3, our statements are based on the estimands as described in the draft addendum R1 to the ICH E9 guideline[25]; hence, the estimands we are referring to in this section may be subject to change.

We focus specifically on trials in *oncology* with differences in follow-up times driven by progression, discontinuation of treatment, death, or end of follow-up. As progression in general leads to treatment discontinuation, we consider only the intercurrent events *treatment discontinuation* and *death*. For all estimands considered in Section 3.1, except the principal stratum estimand, the population is defined by the inclusion and exclusion criteria of the study and contains the treated patients (first element of estimand description). With respect to the second element of estimand description, for all estimands, the endpoint is the time to the first AE of a specific type. The estimands differ as a consequence of the following aspects: the follow-up time over which AEs are included in the analysis and the way how the intercurrent events treatment discontinuation and death are accounted for (third element of estimand description) and by the population-level summary for the endpoint (fourth element of estimand description). Here, by "intercurrent events," we really mean "*postrandomization events*," because death is both postrandomization and terminal but not truly intercurrent.

Considering the *treatment policy estimand*, the interest is in the comparison of treatment groups with respect to AE occurrence until death or end of follow-up, irrespective of the intercurrent event discontinuation of treatment. So it includes all AEs until death or end of study and, therefore, requires the collection of AE data after treatment discontinuation. The AE hazards of the treatment groups can be compared by calculating the hazard ratio in a Cox regression model where for patients without AE, the time to AE is censored by death or by end of follow-up. In the interpretation of the result, it has to be considered if treatment has an effect on the CE death without prior AE, ie, if the hazard ratio between treatment groups with respect to death without prior AE is different from one. The hazards of the CE death without prior AE have also to be taken into account in the estimation of the AE probabilities within the treatment groups by using the Aalen-Johansen estimator. Treatment groups can also be compared with respect to the AE probabilities by estimating the difference between AE probabilities at a specified time point or over the whole study period, by calculating the subdistribution hazard ratio from a Fine and Gray model,[38] or by calculating the odds ratio from a proportional odds cumulative incidence function model.[39] The decision between treatment comparison by means of hazard functions or by means of probability functions defines the fourth element of estimand description mentioned above. The *while on treatment estimand* includes the AEs until discontinuation of treatment and requires the collection of AE data up to this event. Treatment groups can be compared with the same methods as used for the treatment policy estimand, now treating discontinuation of treatment before AE and death without prior AE as CEs.

The *composite estimand* would combine the interesting event AE with the intercurrent events treatment discontinuation and death without prior AE. The endpoint is then time to AE, treatment discontinuation, or death whatever occurs first. So AE data after treatment discontinuation are not required. However, this does not hold in general for any intercurrent event. In this situation, no CEs are present, and standard survival analysis techniques can be applied, ie the composite event hazards of the treatment groups can be compared by calculating the hazard ratio in a Cox regression model, and the probabilities of the composite event can be estimated by one minus the Kaplan-Meier estimator, where for patients without the composite event, the time to event is censored by end of follow-up. However, whether such a composite endpoint is an adequate safety outcome is a different question. We reiterate that any effect on the composite may be disentangled, and effects on the single components of the composite may be analysed using the techniques discussed in Section 4. The *hypothetical estimand* targets the effect of treatment on AE occurrence in the hypothetical scenario in which the intercurrent event treatment discontinuation would not occur. This estimand requires the collection of AE data only up to this event, under the assumption that both the AE hazard and the death hazard remain unchanged in this hypothetical situation. The estimator used for the while on treatment estimand would be valid also for the hypothetical estimand but now handling treatment discontinuation as a pure censoring event and not as a competing risk. As the assumption is obviously both untestable and rather strong, sensitivity analyses are a minimum requirement, when this estimand is targeted. An even more hypothetical estimand would target the effect of treatment on AE occurrence, assuming that neither treatment discontinuation nor death before AE occurs. This estimand is even more hypothetical in the sense that one might

be able to imagine situations where treatment continuation is enforced, but enforcing the absence of death is much more speculative.

The *principal stratum estimand* builds on the causal framework of potential primary outcomes. A potential outcome is defined for each possible treatment assignment, but only one potential outcome is observed, namely, that for the actually assigned treatment. This framework is now extended to *potential intercurrent events* and targets a causal treatment comparison within subpopulations defined by potential intercurrent events. In our example, the potential intercurrent event is treatment discontinuation, while alive, as a function of time and for each possible treatment assignment. A *basic* principal stratum contains all patients with identical values for all potential intercurrent events. For the example of treatment discontinuation (but ignoring *time* until discontinuation for ease of presentation) and two possible treatment assignments, one basic principal stratum is the set of all patients whose potential intercurrent event statuses are (yes, yes); this is the set of all patients who discontinue treatment under both treatment assignments. By construction, measuring the difference between potential outcomes on such a basic principal stratum yields a causal effect "adjusting" for the intercurrent events. Next, a principal stratum is a union of basic principal strata. A common example is the set of patients where the potential intercurrent events are unaffected by treatment—(yes, yes) and (no, no) in the example above—and the complement are all patients where the potential intercurrent events differ between treatments ([no, yes], [yes, no]). However, it is impossible to identify these patients in advance, and a post hoc analysis, eg, based only on those patients who did not stop treatment would be biased, one reason being time-dependent confounding. For the principal stratum estimand, causal inference methods would be required, extending the statistical methods presented in Section 4. Causal inference methods have been developed for the analysis of time-to-event data,[69,70] including methods for CEs,[71] but practical applications may be subtle. For instance, a key concept in causal reasoning is that of a treatment regimen, determined at time 0. In our context, one may, at least theoretically, envisage enforcing a treatment regimen of no discontinuation. However, progression events (which trigger treatment discontinuation) are not controllable in terms of a "progression regimen," which is one major motivation behind the principal stratum framework.[72]

6 | DISCUSSION

The introduction of estimands provides a framework to structure the discussion about the analyses of AE in terms of specific demands/needs and appropriateness of different analyses approaches. We formulated a framework based on safety estimands within which we proposed statistical methods including methods for evidence synthesis that map the AE data to a single value. For the described estimands, we have given recommendations which estimators should be used. In particular, we would like to advocate the use of time-to-event methodology for the analysis of AE data, although such a proposal is known for a quite a long time.[35]

As discussed in Section 3.2, estimands of primary interest may differ between drug approval agencies and the HTA bodies in certain instances. The regulatory agencies evaluate the safety profile of a new treatment. The assessment is based on all safety data available for the new treatment which is summarized in the summary of safety. There, the safety profile of the new treatment is provided in detail showing safety data of all kinds (ie, AEs, laboratory data, physical examination, and vital signs) of all available studies with the new treatment to assess the benefit and risk of this treatment. In addition, the safety is assessed across the whole lifetime of the new treatment by assessing the required periodic safety updates. The standard approach is to use descriptive statistics such as absolute and relative frequencies. Common practice evaluates the safety of the treatment itself using a while on treatment estimand, which could also be formulated as a hypothetical estimand under relevant assumptions as stated in Section 5. The HTA bodies on the other hand are interested in the relative effect on AEs of the new treatment in comparison with a chosen comparator in the indication of interest. The estimand of interest is the treatment policy. In indications such as diabetes mellitus, studies may cover both types of estimands if treatment times and follow-up are similar in both treatment groups. In indications such as oncological diseases, the follow-up of AEs is stopped at the planned end of the study treatment plus a certain number of days, which often results in considerable differences in follow-up times. In such scenarios the treatment policy estimands cannot be covered as AEs usually are not collected for subsequent therapies. A possible solution would be to plan studies that enable estimates for both estimands. All efforts should be undertaken while planning and conducting clinical trials to obtain similar follow-up times. In practice, however, this might not always be possible. Practical challenges include scenarios with patients moving on to different studies due to treatment failure. In such cases, similar follow-up times cannot be guaranteed. Even if the regulatory agencies have other tasks than HTA bodies, the general objective is to establish beneficial treatments for patients. The common parts should be identified in order to harmonize both perspectives.

Further work, some of it under way, can be done or is required in several areas, some of which have already been mentioned. Firstly, in the present paper, we restricted ourselves to methods for analysing the time to occurrence of first AEs. Therefore, the methodologies proposed in this paper need to be revisited with a view to analysing recurrent AEs, intermittent AEs, and AEs with varying severity. Secondly, our proposed methods for analysing AEs lack from adequately accounting for the occurrence of multiple, different AEs. Approaches that account for a number of types of possibly related AEs do exist.[73,74] Thirdly, an empirical investigation is underway to investigate in a large number of randomized controlled trials whether the different analyses of AEs lead to different decisions when comparing safety between groups. Fourthly, it is an open question what impact regulators and HTA agencies might have on the development of methods for the analysis of AE data through, eg, the ICH E9 addendum on estimands and sensitivity analysis in clinical trials. Finally, we have demonstrated that there is a gap between what would ideally be seen in benefit assessments from an HTA agency perspective and current practices of how data are collected in trials with an objective to achieve marketing authorization. It will not be easy to reconcile these two different views. However, the present article stimulated the discussion between different stakeholders.

ACKNOWLEDGEMENTS

ORCID

Steffen Unkel http://orcid.org/0000-0002-2083-0090
Norbert Benda http://orcid.org/0000-0001-5605-2414
Tim Friede http://orcid.org/0000-0001-5347-7441

REFERENCES

1. Agresti A. *Categorical Data Analysis*, 3rd ed. Hoboken, New Jersey: Wiley; 2013.

2. Collett D. *Modelling Survival Data in Medical Research*, 3rd ed. Boca Raton, Florida: Chapman & Hall/CRC Press; 2015.

3. Akacha M, Bretz F, Ohlssen D, Schmidli H. Estimands and their role in clinical trials. *Stat Biopharmaceutical Res*. 2015;22:1-4.

4. Leuchs A-K, Zinserling J, Brandt A, Wirtz D, Benda N. Choosing appropriate estimands in clinical trials. *Therapeutic Innov Regul Sci*. 2015;49:584-592.

5. Akacha M, Bretz F, Ruberg S. Estimands in clinical trials—broadening the prspective. *Stat Med*. 2017;36:5-19.

6. International Conference on Harmonisation. ICH E2A: clinical safety data management: definition and standards for expedited reporting. 1994. CPMP/ICH/377/95. http://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Efficacy/E2A/Step4/E2A_Guideline.pdf. Accessed on January 28, 2018.

7. International Conference on Harmonisation. ICH E9: statistical principles for clinical trials. 1998. CPMP/ICH/363/96. http://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Efficacy/E9/Step4/E9_Guideline.pdf. Accessed on January 28, 2018.

8. International Conference on Harmonisation. ICH E1: Population exposure: the extent of population exposure to assess clinical safety for drugs intended for long-term treatment of non-life-threatening conditions. 1994. CPMP/ICH/375/95. http://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Efficacy/E1/Step4/E1_Guideline.pdf. Accessed on January 30, 2018.

9. International Conference on Harmonisation. ICH E3: structure and content of clinical study reports. 1995. CPMP/ICH/137/95. http://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Efficacy/E3/E3_Guideline.pdf. Accessed on January 28, 2018.

10. International Conference on Harmonisation. ICH E5(R1): ethnic factors in the acceptability of foreign clinical data. 1998. CPMP/ICH/289/95. http://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Efficacy/E5_R1/Step4/E5_R1__Guideline.pdf. Accessed on January 30, 2018.

11. International Conference on Harmonisation. ICH M4E(R2): revision of M4E guideline on enhancing the format and structure of benefit-risk information in ICH. 2016. CPMP/ICH/2887/1999. http://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/CTD/M4E_R2_Efficacy/M4E_R2__Step_4.pdf. Accessed on August 20, 2018.

12. U.S. Food and Drug Administration (FDA). Guidance for industry. Premarketing risk assessment. 2005. https://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/ucm072002.pdf. Accessed on August 31, 2018.

13. Altman DG, Bland JM. Statistics notes: absence of evidence is not evidence of absence. *BMJ*. 1995;311:485.

14. Allignol A, Beyersmann J, Schmoor C. Statistical issues in the analysis of adverse events in time-to-event data. *Pharmaceutical Stat*. 2016;15:297-305.

15. Bender R, Beckmann L, Lange S. Biometrical issues in the analysis of adverse events within the benefit assessment of drugs. *Pharmaceutical Stat*. 2016;15:292-296.

16. Chuang-Stein C, Le V, Chen W. Recent advancements in the analysis and presentation of safety data. *Drug Inf J*. 2001;35:377-397.

17. U.S. Food and Drug Administration (FDA). Guidance for industry. Integrated summaries of effectiveness and safety: location within the common technical document. 2009. https://www.fda.gov/downloads/drugs/guidances/ucm136174.pdf. Accessed on February 3, 2018.

18. Neal B, Perkovic V, Mahaffey KW, et al. Canagliflozin and cardiovascular and renal events in type 2 diabetes. *N Engl J Med*. 2017;377:644-657.

19. Busse R, Orvain J, Velasco M, et al. Best practice in undertaking and reporting health technology assessments. *Int J Technol Assess Health Care*. 2012;18:361-422.

20. Pieper D, Antoine S-L, Morfeld J-C, Mathes T, Eikermann M. Methodological approaches in conducting overviews: current state in HTA agencies. *Res Synth Methods*. 2014;5:187-199.

21. Guide to the methods of technology appraisal 2013. National Institute for Health and Care Excellence, London, United Kingdom, 2013. https://www.nice.org.uk/process/pmg9/chapter/foreword. Accessed on January 28.

22. Ara R, Wailoo A. Using health state utility values in models exploring the cost-effectiveness of health technologies. *Value Health*. 2012;15:971-974.

23. Craig D, McDaid C, Fonseca T, Stock C, Duffy S, Woolacott N. Are adverse effects incorporated in economic models? An initial review of current practice. *Health Technol Assess*. 2009;13:1-71, 97–181.

24. IQWiG. Allgemeine Methoden, Version 5.0. Institute for Quality and Efficiency in Health Care, Cologne, Germany; 2017.

25. International Conference on Harmonisation. ICH E9(R1): addendum to the guideline on statistical principles for clinical trials on estimands and sensitivity analysis in clinical trials, step 2. 2017. Draft guideline. http://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Efficacy/E9/E9-R1EWG_Step2_Guideline_2017_0616.pdf. Accessed on January 28, 2018.

26. IQWiG. Comments from IQWiG on 'ICH E9 (R1) addendum on estimands and sensitivity analysis in clinical trials to the guideline on statistical principles for clinical trials' (EMA/CHMP/ICH/436221/2017). 2018. Institute for Quality and Efficiency in Health Care, Cologne, Germany. https://www.iqwig.de/download/2018-02-19_CommentsEMA_WC500233916_IQWiG.pdf. Accessed on April 18, 2018.

27. Robert C, Schachter J, Long GV, et al. Pembrolizumab versus ipilimumab in advanced melanoma. *N Engl J Med*. 2015;372:2521-2532.

28. Calaminus G, Kaatsch P. Position paper of the society of pediatric oncology and hematology (GPOH) on (long-term) surveillance, (long-term) follow-up and late effect evaluation in pediatric oncology patients. *Klinische Pädiatrie*. 2007;219:173-178.

29. Siddiqui O. Statistical methods to analyze adverse events data of randomized clinical trials. *J Biopharm Stat*. 2009;19:889-899.

30. Ioannidis JA, Evans SW, Gøtzsche PC, et al. Better reporting of harms in randomized trials: an extension of the CONSORT statement. *Ann Intern Med*. 2004;141:781-788.

31. Lineberry N, Berlin JA, Mansi B, et al. Recommendations to improve adverse event reporting in clinical trial publications: a joint pharmaceutical industry/journal editor perspective. *BMJ*. 2016;355:i5078.

32. Amit O, Heiberger RM, Lane PW. Graphical approaches to the analysis of safety data from clinical trials. *Pharm Stat*. 2008;7:20-35.

33. 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.

34. Aalen OO, Johansen S. An empirical transition matrix for non-homogeneous Markov chains based on censored observations. *Scand J Stat*. 1978;5:141-150.

35. O'Neill RT. Statistical analyses of adverse event data from clinical trials special emphasis on serious events. *Drug Inf J*. 1987;21:9-20.

36. Crowe BJ, Brueckner A, Beasley C, Kulkarni P. Current practices, challenges, and statistical issues with product safety labeling. *Stat Biopharmaceutical Res*. 2013;5:180-193.

37. Beyersmann J, Allignol A, Schumacher M. *Competing Risks and Multistate Models with R*. New York: Springer; 2012.

38. Fine J, Gray R. A proportional hazards model for the subdistribution of a competing risk. *J Am Stat Assoc*. 1999;94:496-509.

39. Eriksson F, Li J, Scheike T, Zhang M-J. The proportional odds cumulative incidence model for competing risks. *Biometrics*. 2015;71:687-695.

40. Andersen PK, Keiding N. Interpretability and importance of functionals in competing risks and multistate models. *Stat Med*. 2012;31:1074-1088.

41. Beyersmann J, Termini SD, Pauly M. Weak convergence of the wild bootstrap for the Aalen–Johansen estimator of the cumulative incidence function of a competing risk. *Scand J Stat*. 2013;40:387-402.

42. Aalen OO, Borgan Ø, Gjessing HK. *Survival and Event History Analysis: A Process Point of View*. New York: Springer; 2008.

43. O'Quigley J. *Proportional Hazards Regression*. New York: Springer; 2008.

44. Kleinbaum DG, Klein M. *Survival Analysis: A Self-Learning Text*, 3rd ed. New York: Springer; 2012.

45. Andersen PK, Borgan Ø, Gill RD, Keiding N. *Statistical Models Based on Counting Processes*. New York: Springer; 1993.

46. Keiding N, Andersen PK, eds.. *Survival and Event History Analysis*. Chichester: Wiley; 2006.

47. U.S. Food and Drug Administration (FDA). Reviewer guidance. Conducting a clinical safety review of a new product application and preparing a report on the review. 2005. https://www.fda.gov/ohrms/dockets/ac/05/briefing/2005-4143B1_06_Tab-13.pdf. Accessed on 28 January 2018.

48. McEntegart DJ. Pooling in integrated safety databases. *Drug Inf J*. 2000;34:495-499.

49. Rücker G, Schumacher M. Simpson's paradox visualized: The example of the rosiglitazone meta-analysis. *BMC Med Res Methodol*. 2008;8:34.

50. Chuang-Stein C, Beltangady M. Reporting cumulative proportion of subjects with an adverse event based on data from multiple studies. *Pharmaceutical Stat*. 2011;10:3-7.

51. Council for International Organisations of Medical Sciences (CIOMS) Working group X. Evidence synthesis for meta-analysis for drug safety. report of CIOMS Working Group X, Geneva Switzerland, Council for International Organizations of Medical Sciences (CIOMS); 2016.

52. Berlin JA, Crowe BJ, Whalen E, Xia HA, Koro CE, Kuebler J. Meta-analysis of clinical trial safety data in a drug development program: answers to frequently asked questions. *Clin Trials*. 2013;10:20-31.

53. Tierney JF, Vale C, Riley R, et al. Individual participant data (IPD) meta-analyses of randomised controlled trials: guidance on their use. *PloS Med*. 2015;12:e1001855.

54. Guyot P, Ades AE, Ouwens MJNM, Welton N. Enhanced secondary analysis of survival data: reconstructing the data from published Kaplan-Meier survival curves. *BMC Med Res Methodol*. 2012;12:9.

55. Liu Z, Rich B, Hanley JA. Recovering the raw data behind a non-parametric survival curve. *Systematic Rev*. 2015;3:151.

56. Schemper M, Wakounig S, Heinze G. The estimation of average hazard ratios by weighted Cox regression. *Stat Med*. 2009;28:2473-2489.

57. Bender R, Friede T, Koch A, et al. Methods for evidence synthesis in the case of very few studies. *Res Synth Methods*. 2018;9(3):382-392. https://doi.org/10.1002/jrsm.1297

58. Turner RM, Davey J, Clarke MJ, Thompson SG, Higgins JPT. Predicting the extent of heterogeneity in meta-analysis, using empirical data from the Cochrane Database of Systematic Reviews. *Int J Epidemiol*. 2012;41:818-827.

59. Spiegelhalter DJ, Abrams KR, Myles JP. *Bayesian Approaches to Clinical Trials and Health-Care Evaluation*. Chichester: Wiley; 2004.

60. Higgins JPT, Thompson SG, Spiegelhalter DJ. A re-evaluation of random-effects meta-analysis. *J R Stat Soc Ser A*. 2009;172:137-159.

61. Friede T, Röver C, Wandel S, Neuenschwander B. Meta-analysis of few small studies in orphan diseases. *Res Synth Methods*. 2017;8:79-91.

62. Friede T, Röver C, Wandel S, Neuenschwander B. Meta-analysis of two studies in the presence of heterogeneity with applications in rare diseases. *Biom J*. 2017;59:658-671.

63. Röver C, Friede T. Discrete approximation of a mixture distribution via restricted divergence. *J Comput Graph Stat*. 2017;26:217-222.

64. Röver C, Knapp G, Friede T. Hartung-Knapp-Sidik-Jonkman approach and its modification for random-effects meta-analysis with few studies. *BMC Med Res Methodol*. 2015;15:99.

65. Bradburn MJ, Deeks JJ, Berlin JA, Localio AR. Much ado about nothing: a comparison of the performance of meta-analytical methods with rare events. *Stat Med*. 2007;26:53-77.

66. Böhning D, Kuhnert R, Rattanasiri S. *Meta-analysis of Binary Data Using Profile Likelihood*. Boca Raton, Florida: Chapman and Hall/CRC Press; 2008.

67. Kuß O. Statistical methods for meta-analyses including information from studies without any events—add nothing to nothing and succeed nevertheless. *Stat Med*. 2015;34:1097-1116.

68. Günhan BK, Röver C, Friede T. Meta-analysis of few studies involving rare events. http://arxiv.org/abs/1809.04407 (submitted for publication); 2018.

69. Baker SG. Analysis of survival data from a randomized trial with all-or-none compliance: estimating the cost-effectiveness of a cancer screening program. *J Am Stat Assoc*. 1998;93:929-934.

70. Hernán MA, Robins JM. *Causal Inference*. Boca Ration, Florida: Chapman & Hall/CRC Press; 2018. forthcoming.

71. Andersen PK, Syriopoulou E, Parner ET. Causal inference in survival analysis using pseudo-observations. *Stat Med*. 2017;36:2669-2681.

72. Frangakis CE, Rubin DB. Principal stratification in causal inference. *Biometrics*. 2002;58:21-29.

73. Güttner A, Kübler J, Pigeot I. Multivariate time-to-event analysis of multiple adverse events of drugs in integrated analyses. *Stat Med*. 2007;26:1518-1531.

74. Berry SM, Berry DA. Accounting for multiplicities in assessing drug safety: a three-level hierarchical mixture model. *Biometrics*. 2004;60:418-426.