

GRADE SERIES

GRADE Guidelines: 29. Rating the certainty in time-to-event outcomes—Study limitations due to censoring of participants with missing data in intervention studies

Marius Goldkuhle^{a,*}, Ralf Bender^b, Elie A. Akl^{c,d}, Elvira C. van Dalen^e, Sarah Nevitt^f, Reem A. Mustafa^{d,g}, Gordon H. Guyatt^{d,h,i}, Marialene Trivella^j, Benjamin Djulbegovic^k, Holger Schünemann^{d,h,i}, Michela Cinquini^l, Nina Kreuzberger^a, Nicole Skoetz^a, GRADE Working Group

^aDepartment of Internal Medicine, University of Cologne, Faculty of Medicine and University Hospital Cologne, Center for Integrated Oncology Aachen Bonn Cologne Duesseldorf, Kerpener Str. 62, 50937, Cologne, Germany

^bDepartment of Medical Biometry, Institute for Quality and Efficiency in Health Care, Im Mediapark 8, D-50670 Cologne, Germany

^cDepartment of Internal Medicine, American University of Beirut, P.O.Box 11-0236, Lebanon, Canada

^dDepartment of Health Research Methods, Evidence, and Impact, McMaster University, 1280 Main St. W., Hamilton, Ontario, L8S 4K1, Canada

^ePrincess Máxima Center for Pediatric Oncology, Heidelberglaan 25, 3584CS, Utrecht, The Netherlands

^fDepartment of Biostatistics, University of Liverpool, Block F, Waterhouse Building, 1-5 Brownlow Street, Liverpool, L69 3GL, UK

^gDepartment of Medicine, University of Kansas Health System, 3901 Rainbow Blvd, MS3002, Kansas City, KS, 66160, USA

^hDepartment of Medicine, McMaster University, 1280 Main St. W., Hamilton, Ontario, L8S 4K1, Canada

ⁱMcMaster GRADE Centre & Michael G DeGroot Cochrane Canada Centre, McMaster University, Hamilton, Ontario, Canada

^jCentre for Statistics in Medicine, University of Oxford, Botnar Research Centre, Windmill Rd, Oxford, OX3 7LD, UK

^kCity of Hope, 1500 Duarte Rd, Duarte, CA, 91010, USA

^lUnit of Systematic Reviews and Guidelines Production, Mario Negri Institute for Pharmacological Research IRCCS, Via Giuseppe La Masa 19, 20156, Milan, Italy

Accepted 2 September 2020; Published online 30 September 2020

Abstract

Objectives: To provide Grading of Recommendations, Assessment, Development, and Evaluation (GRADE) guidance for the consideration of study limitations (risk of bias) due to missing participant outcome data for time-to-event outcomes in intervention studies.

Study Design and Setting: We developed this guidance through an iterative process that included membership consultation, feedback, presentation, and iterative discussion at meetings of the GRADE working group.

Results: The GRADE working group has published guidance on how to account for missing participant outcome data in binary and continuous outcomes. When analyzing time-to-event outcomes (e.g., overall survival and time-to-treatment failure) data of participants for whom the outcome of interest (e.g., death and relapse) has not been observed are dealt with through censoring. To do so, standard methods require that censored individuals are representative for those remaining in the study. Two types of censoring can be distinguished, end of study censoring and censoring because of missing data, commonly named loss to follow-up censoring. However, both types are not distinguishable with the usual information on censoring available to review authors. Dealing with individuals for whom data are missing during follow-up in the same way as individuals for whom full follow-up is available at the end of the study increases the risk of bias. Considerable differences in the treatment arms in the distribution of censoring over time (early versus late censoring), the overall degree of missing follow-up data, and the reasons why individuals were lost to follow-up may reduce the certainty in the study results. With often only very limited data available, review and guideline authors are required to make transparent and well-considered judgments when judging risk of bias of individual studies and then come to an overall grading decision for the entire body of evidence.

Conflict of interest statement: None.

* Corresponding author. Department of Internal Medicine Center for Integrated Oncology Aachen Bonn Cologne Duesseldorf University Hospital of Cologne Kerpener Str. 62 50637 Köln, Germany. Tel.: +49 221 478-62032; fax: +49 221 478-96654.

E-mail address: marius.goldkuhle@uk-koeln.de (M. Goldkuhle).

Conclusion: Concern for risk of bias resulting from censoring of participants for whom follow-up data are missing in the underlying studies of a body of evidence can be expressed in the study limitations (risk of bias) domain of the GRADE approach. © 2020 Elsevier Inc. All rights reserved.

Keywords: GRADE; Certainty of the evidence; Time-to-event outcomes; Survival analysis; Risk of bias; Loss to follow-up; Censoring missing data

1. Introduction

The Grading of Recommendations, Assessment, Development, and Evaluation (GRADE) working group has defined domains that can limit the certainty in a body of evidence [1–6]. Within its study limitations domain (i.e., risk of bias), the GRADE approach has issued guidance on how to account for missing participant outcome data for binary and continuous outcomes [6,7]. That guidance proposes conducting sensitivity meta-analyses making assumptions about the outcomes of participants with missing data, to test the robustness of the findings of the primary meta-analysis [7,8].

Although the basic principles for assessing risk of bias associated with missing participant outcome data in binary outcome analysis also apply to time-to-event analysis, there are issues uniquely applicable to time-to-event outcomes. In contrast to binary data analysis, time-to-event studies, which assess not only whether an event of interest occurs but also when it occurs, typically follow patients for varying periods of time. Because time-to-event analyses include data from individuals with variable lengths of follow-up, those for whom follow-up data becomes absent during the study interval are typically treated in the same way as those with regular follow-up until the end of the analysis (i.e., they provided complete data). Therefore, we here refer to missing follow-up data to characterize the situation when information for an individual becomes absent at a time point within the intended and prespecified observation period. This article discusses GRADE rating of study limitations associated with missing follow-up data when dealing with time-to-event analysis.

2. Background

2.1. Time-to-event analysis and censoring

Time-to-event analysis is also often referred to as survival analysis, in which the “survival time” describes the time until an event such as death occurs. The most prominent methods to analyze time-to-event outcomes include Kaplan–Meier curves along with the log-rank test and the Cox proportional hazards regression model [9,10]. Time-to-event outcomes are often described by survival rates, defined as the probability that an individual will not have experienced an event (e.g., “survived”) up to a certain time point or hazard rates, which can be interpreted as

instantaneous failure rates, meaning an individual’s likelihood of experiencing an event (e.g., “death”) at a certain time point given that the event has not occurred up to this time point.

The most prominently applied relative effect measure is the hazard ratio, which is the ratio of hazards between two groups. It is commonly obtained from the Cox proportional hazards regression model, which adjusts for relevant covariates and confounders. An unadjusted hazard ratio can also be derived indirectly using other analytical techniques, such as the Kaplan–Meier method or the log-rank test [10,11].

A core feature of time-to-event analysis is the consideration of “censoring” which occurs when patients complete their follow-up period without having experienced the event of interest. Censored observations are included in analyses to optimize the efficiency that time-to-event analysis provides over binary data analysis [12]. If the time to an event and censoring are not included in the calculation of the (log) hazard ratio, it equals the (log) relative risk.

To include censored observations in time-to-event analyses, general methods of survival analysis require an assumption of noninformative and independent censoring. Violations of this assumption introduce risk of bias. [Appendix A1](#) provides a short review of the definition of noninformative censoring and its relation to independent censoring. In accordance with established training resources for time-to-event analysts [13], we will use the concept of independent and dependent censoring to describe situations under which censoring may lead to distortion of the analysis results.

Independent censoring occurs when censored participants and those remaining under observation have the same probability of experiencing the event of interest, as if the censored individuals were “randomly drawn” during the course of follow-up [13,14]. An example for censoring mechanisms independent from the survival time (and also noninformative) is administrative closure of a study. Differences in the observation times of participants then are solely a result of the staggered study entry times and the fixed study closure time ([Fig. 1](#)) [13,15].

When individuals are censored because of missing follow-up data, this assumption is likely to be violated. Examples of such situations which may bias results include:

- Participants withdraw consent because of physical or mental side effects of an intervention;

What is new?

Key findings

- Analysis methods for time-to-event outcomes deal with participants for whom outcome data are unavailable through censoring. Two types of censoring, the end of study censoring and censoring because of missing data (commonly named loss to follow-up censoring), have to be differentiated.
- Censoring of individuals with missing follow-up data is likely to violate the assumption of independence of censoring and increases the risk of biased results.
- The magnitude of bias resulting from censoring of participants with missing data depends on several factors. An increasing degree of dependent censored observations and difference among the study arms increases the risk of bias.

What are the implications and what should be changed?

- Often, reasons why individuals in studies were censored and the time points of censoring are unavailable to systematic review and guideline authors who therefore have to make risk of bias judgments for primary studies based on the distribution of censoring over time or the degree of missing participant follow-up data.
- Systematic review and guideline authors need to make GRADE judgments across the body of evidence for study limitations resulting from censoring of participants with missing data considering all available information, including the possibility of carrying out sensitivity analysis by assessing whether studies at high risk of bias or studies in which there are concerns yield different results.

- Participants are withdrawn from the observation and censored after switching treatment as a result of progressive disease;
- Investigators fail to locate study participants.

2.2. Reporting time-to-event data and censoring in primary studies

Flaws in reporting time-to-event analyses may complicate their adequate appraisal by systematic review authors including assessing risk of bias resulting from censoring of individuals with missing follow-up data [16–19]. Suboptimal reporting includes, but is not limited to outcome definitions, the extent and duration of follow-up, precision measures such as the number of participants at risk at

certain time points, and details of statistical model building. Authors often fail to precisely define censoring mechanisms, omit the number of censored participants, and fail to state why individual study participants were censored [16–19].

Studies published in leading medical journals are not immune to reporting limitations: for instance, one methodological study found inconsistency between the number of participants reported in the text/tables as “lost before the end of the study” and those assessed from Kaplan–Meier curves [20]. Prior work has specified minimal reporting items for time-to-event analyses and survival curves [17,18,21,22]. Appendix A2 outlines reporting requirements that allow systematic review and guideline authors to assess possible risk of bias resulting from informative censoring.

3. Methods

This guidance was developed by the members of the GRADE working group. They included methodologists, clinical epidemiologists, and biostatisticians with experience in systematic reviews and/or guideline development. The group developed the guidance based on iterative discussions by e-mail, on conference calls, and at a GRADE working group meeting in Manchester, UK, in October 2018. The final draft of the guidance was presented during the GRADE working group meeting in Hamilton in June 2019 and was approved following the group’s standard approval process.

4. Scope

This guidance aims to support systematic review and guideline authors in the assessment of study limitations (risk of bias) due to missing follow-up data for time-to-event outcomes in intervention studies. We describe an approach that takes a systematic reviewer perspective relying on information that one could typically obtain from only the trial report and its accompanying records. To comply with well-known resources for systematic review authors to assess the risk of bias in individual studies and with reference to previous GRADE guidance for rating the certainty of the evidence with focus on study limitations (risk of bias), we refer to missing follow-up data as the unavailability of follow-up data for individuals during the study interval [6,23,24]. This includes all types of missing data and situations in which the outcome status of study participants becomes unavailable during the study period irrespective of the reason (e.g., patients not available or inappropriately excluded) [25,26].

The concerning risk of bias arises, for example, when investigators censor individuals for whom data are missing and include them in the computation of effect measures

in the same way as participants with independent censoring (e.g., those whose follow-up ended appropriately at the end of the data collection period). Systematic review and guideline authors seldom have information regarding the reasons for censoring for each participant in every eligible study. Consistent with well-known instructions for systematic review authors, we therefore provide guidance that is primarily aimed at detecting a potential bias in individual studies [23,24]. Judgments on study level then inform the risk of bias assessment for an overall body of evidence separately for each outcome.

In accordance with previous GRADE guideline for missing participant outcome data for binary and continuous outcomes, we provide guidance for systematic review and guideline authors who assess comparative clinical trials based on aggregated data [7]. We differentiate the issue of adequately accounting for loss to follow-up from that of adherence to the intention-to-treat principle, which relates to analyzing study participants with known data in the groups to which they were allocated [7,27].

We focus on risk of bias in the outputs of the “standard” methods of survival analysis and the Cox model hazard ratio as the single comparative relative effect size measure [16–19]. Within the context of this guidance, we assume that the primary study investigators and subsequently the authors of meta-analyses have chosen the correct method for analyzing competing events for the intended research question.

5. How censoring participants with missing follow-up data may affect the results of the study

5.1. Censoring of participants leading to overestimation and underestimation of the survival probability

Similar to binary outcome analysis, the distortion of the outcome probability of the group under study depends on the outcome probability of those for whom data are missing. When individuals who are more likely to experience the (negative) event of interest (e.g., death) are also more likely to be missing (positive correlation between the occurrence of the event and missingness of data), for example, because they are more likely to be lost to follow-up, the true survival probability of a study group will inevitably be overestimated [12,23]. This means that the corresponding true risk of the (negative) event occurrence will be underestimated. Such an association may occur, for example, if participants with treatment-related adverse events are no longer followed up and are censored at the time of loss to follow-up.

On the other hand, in case of a negative correlation between the occurrence of the event and the probability of being censored, the true survival probability for a study group may be underestimated (and the corresponding true event risk overestimated) [23]. For example, underestimation of

the event-free survival probability will occur if in a study comparing the impact of psychiatric interventions on time-to-treatment failure participants in one arm benefit so substantially that they fail to return and are therefore lost from the study.

5.2. Effect of censoring of participants with missing follow-up data on the hazard ratio

Factors that might result in a biased hazard ratio are the frequency of the outcome event of interest, the treatment effect in terms of the distribution of the outcome event between the study arms, and the frequency and distribution of censoring because of missing data (e.g., effect of intervention on the frequency of loss to follow-up). As the impact of dependent censoring on the hazard ratio cannot be determined based on the observed data (because the true outcome of censored individuals is not observable), quantifications of the associated bias are difficult [15].

Nevertheless, the potential bias resulting from censoring of missing follow-up data can be substantial, especially when the outcome probability for those with missing data is considerably increased. In studies evaluating antiretroviral treatment programs for HIV in settings with limited resources, loss to follow-up rates are typically high. Performing a systematic review and meta-analysis of studies of such programs in which individuals lost to follow-up were actively traced by telephone calls or social networks; Brinkhof et al. [28] found that the mortality among patients lost to follow-up was considerably increased. In a subsequent study, Brinkhof et al. [29] then used the mortality estimates from their previous systematic review to impute representative mortality data for individuals lost to follow-up in an evaluation of five antiretroviral treatment programs in sub-Saharan Africa and found that survival analysis ignoring increased mortality among participants lost to follow-up greatly underestimated overall mortality and leads to a biased evaluation of the programs.

In most situations, however, the reasons for censoring and the associated prognosis will be unavailable to systematic review and guideline authors. Therefore, similar to assessments of a risk of bias in binary data analysis, one has to rely on the simplified principle that the higher the frequency of dependent censoring of participants in relation to the event rates and the greater the difference between the groups, the higher the potential for biased results [6]. Simulations of single arm studies show that the degree of bias is more strongly influenced by the overall proportion of participants that are censored with an increased/decreased risk of experiencing the outcome, rather than the difference in the hazard of study participants who are remaining at risk until the end of the observation period and those who are censored [30]. Between-group comparison simulations show that the degree of bias in settings with proportional hazards in Cox models is mainly enhanced by

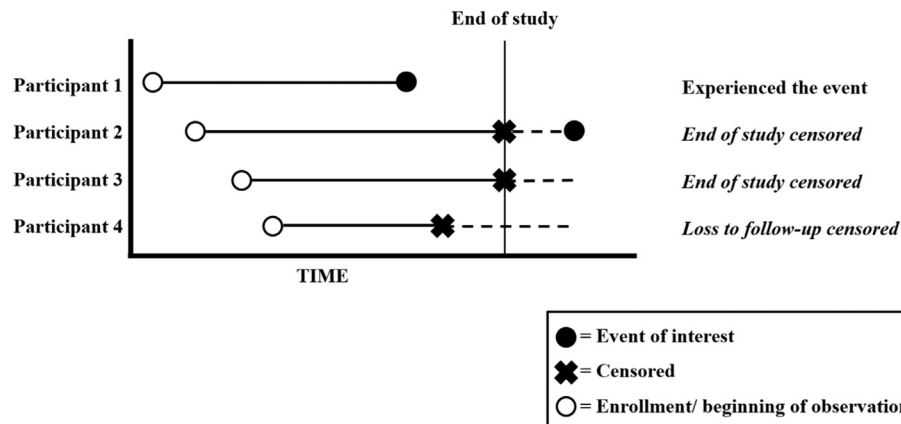


Fig. 1. Types of censoring: For participant 1, the occurrence of the outcome event is observable. Participants 2 and 3 are censored because of the administrative closure of the study. The variation in their duration of follow-up and the differing censoring time points result from the staggered recruiting phase of the study. Participant 4 is lost from the observation before the administrative ending of the study and censored for a different reason.

the overall degree and the early time points of censoring for any reason [31].

5.3. Illustration of the uncertainty introduced through early dependent censoring to comparisons

To illustrate the impact of early dependent censoring on survival analyses, we reconstructed individual participant data from the analysis of overall survival in a study by Denis et al. [32] (see also section 6.1). In this study example, the number of censored participants was different between the groups, particularly in the beginning of follow-up. Given the transparent reporting of outcome and censoring events in the available survival curve (Fig. 2), we were able to reconstruct event and censoring time points for the individuals in each group (see Appendix A3). Box 1 provides a detailed description of the study example, and Appendix A3 provides a summary of our proceeding to reconstruct survival data. We verified the consistency of our reconstructed data set with the approach presented by Guyot et al. [33] that allows recreating individual participant level data from published survival curves by assuming constant censoring within a given time interval and recalculated hazard ratios and Kaplan–Meier survival curves.

To demonstrate the impact of early censoring on the results, we considered a hypothetical scenario in which all participants who were censored before 7 months of follow-up experience the event 1 month after censoring, that is these data are no longer censored but are counted as events. This assumption represents the extreme case of a very large positive correlation between early censoring and the experience of the event of interest.

Appendix A4 Figures 1 and 2 show the Kaplan–Meier survival curves for the reconstructed data set and the hypothetical scenario. The original hazard ratio resulting from the authors' analysis is 0.32 (95% confidence interval (CI) 0.15 to 0.67). The hazard ratio resulting for the data

Box 1 Example 1: Denis et al. [32].

In a randomized trial comparing a web-mediated follow-up strategy with routine surveillance for participants suffering from lung cancer, the primary end point was overall survival defined from random assignment to death or to the last assessment of patient's status when the patient was censored. A hazard ratio between groups was calculated using a Cox proportional hazards model. A total of 133 participants were randomized, and after exclusions of participants found after randomization to be ineligible, 60 and 61 participants were included in the modified intention-to-treat analyses in the intervention and the control arms, respectively. The number of reported deaths per arm was 11 vs. 26 and the number of relapses 34 vs. 36. The study was closed early at an interim analysis by recommendation of the independent data monitoring board.

The degree of censoring was not reported throughout the study publication. However, an assessment of the presented survival curve (Fig. 2) shows substantially more censoring of participants in the experimental arm, particularly during early follow-up. Despite the visible survival benefits and the statistically significant hazard ratio in favor of the intervention group, the number of patients at risk is similar for both treatment arms at months 5 and 10. This suggests that a similar number of individuals who died in the control arm must have been censored in the intervention arm. This severe imbalance, despite randomization of the participants, introduces high risk to bias due to censoring of participants with missing follow-up data. In a hypothetical scenario, where individuals lost to follow-up are more likely than those who were not lost to follow up to die shortly after censoring, the survival benefit shown by the hazard ratio in the study is likely inflated and possibly inexistent. Here, we would suspect a high risk of bias and, in a situation where only one study is included in the body of evidence or other included studies have similar imbalances, we would consider rating down due to study limitations for overall survival.

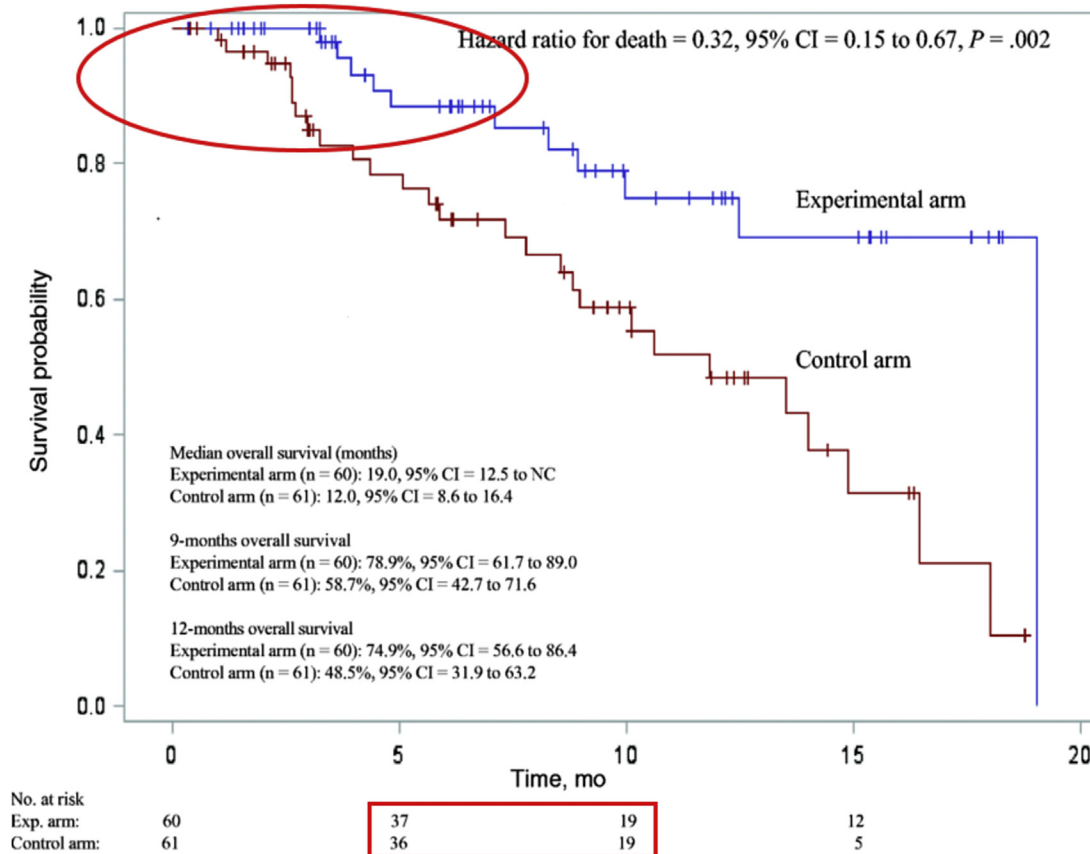


Fig. 2. Kaplan–Meier curve for the outcome overall survival from the study by Denis et al. [32]. The vertical lines crossing the curves mark censored events. The elliptical form indicates that the number of early censored individuals is higher in the experimental arm than the control arm. The rectangular form shows that the number of participants at risk to experience the event for certain time points is reported below the curves for each study arm and are similar for both groups at 5 and 10 months of follow-up, despite a more favorable survival probability in the experimental arm [32]. Adapted from “Randomized Trial Comparing a Web-Mediated Follow-up With Routine Surveillance in Lung Cancer Patients” by Denis et al., 2017, Journal of the National Cancer Institute, 109(9), p. 6. Copyright 2017 by Oxford University Press. Adapted with permission.

we reconstructed from the published survival curve was 0.32 (95% CI 0.15 to 0.65) showing that our reconstructed data set is nearly identical to the original one. The original analysis indicates a substantial survival advantage for participants in the experimental arm under the questionable assumption of independent censoring.

Appendix A4 Figures 1 and 2 illustrate that a positive correlation between early censoring and the experience of the event of interest leads to an overestimation of the survival probability in both study arms. As more participants in the intervention arm are censored before 7 months compared with the control arm (26 participants vs. 19 participants), the hazard ratio increases to 0.69 (95% CI 0.44 to 1.07) in the hypothetical scenario. This illustrates that the effect estimation is biased if there is a positive correlation between early censoring and the experience of the event of interest and additionally a higher proportion of censored participants in the intervention arm. Therefore, there is a loss of certainty in the results of survival analyses in the case of substantial censoring, particularly throughout the early periods of follow-up and where no information is available on the reasons for censoring.

6. Suggestions to assess risk of bias resulting from censoring in an individual study

6.1. Identifying risk of bias due to censoring in individual studies

To appropriately assess the potential bias for the study results emerging from dependent censoring of participants for whom follow-up data are missing, reasons why individual participants were censored for each outcome would be helpful. When information regarding the number of censored individuals with reasons together with the time point of censoring are available, imputation procedures based on assumptions, similar to those described in the GRADE guidance article for missing outcome data within binary data analysis, could be applied to assess the robustness of effect measures to loss to follow-up [7].

Unfortunately, it is unlikely that review authors will be able to obtain data on the reasons and time points for censoring for study participants and the reporting of information on missing data [34]. Nevertheless, before assessing a potential bias, gathering all available information on possible mechanisms for censoring, if possible from the

primary study investigators themselves, is likely to be helpful.

For an informed judgment of risk of bias resulting from censoring of participants because of missing follow-up data, both the degree and the distribution of censoring among the study groups over time should be available. In randomized trials with a valid randomization process, censoring events resulting from treatment independent covariates (independent censoring) should have a similar distribution over time in both treatment arms. An unequivocal difference in the distribution of individuals lost to follow-up over time, for example, a high number of early censoring in one arm vs. late censoring in the other, is likely to indicate dependence of these censoring events.

Differences in early censoring are especially relevant because they can be more easily associated with missing follow-up data than “end-of-study censoring.” In the absence of individual patient data, investigators will need to rely on information about the study participants throughout the course of the study that is available from the reports. Most informative are survival curves and the number of reported individuals at risk to experience the outcome event across the study period.

It is a good practice, even though not consistently performed, to indicate in the survival curves the time points at which individuals were censored [16,22]. This is often performed by study authors by marking censoring time points on the survival curves, for example, as vertical lines or as number of participants censored between given time points displayed along the number of participants at risk for these time points. This information then allows an assessment of whether censoring happened early or late throughout the observation period and to assess differences in this distribution between study arms.

Fig. 2 presents an example in which considerably more participants are censored in the intervention arm during the first month of the study as indicated by the vertical lines crossing the survival curves of the treatment arms. Box 1 presents a detailed description of the example (see also section 5.3).

If only a survival curve and the number at risk for particular time points are available and direct information on the distribution of censoring is not presented (e.g., no censoring marks on the curves) or assessable (e.g., single marks for censoring not distinguishable on the curve due to high degree of censoring), it is sometimes possible to estimate the degree of participants censored for a certain time point by comparing the visible survival benefits in the curves and the number at risk for the reported time points [20]. In Fig. 2, for example, at five and 10 months of follow-up, the same or a similar number of participants at risk are reported in both treatment arms (5 months: 37 vs. 36; 10 months: 19 vs. 19). Comparing this information with the visible differences in survival probabilities in the curves, noticeably favoring the experimental arm, allows

the conclusion that substantially more participants have been lost to follow-up in the experimental than in the control arm. This is because after five and 10 months of follow-up, approximately the same number of individuals who experienced the event (death) in the control arm must have been lost to follow-up in the experimental arm. Box 1 presents a detailed description of the example.

When authors report the number of individuals for several time points together with the survival curves, established methods to reconstruct summary time-to-event data also allow approximations of the number of individuals censored within certain time intervals [11,35]. When authors provide the number of individuals at risk for a sufficient number of time points, such procedures may also conclusively support an assessment of the distribution of censoring in the study arms over time. Considerable variation in the overall difference and a difference in the distribution in terms of early versus late censoring between arms can then confirm a high risk of bias and a critical limitation to the effect estimator of a time-to-event outcome of an individual study allowing guideline authors to carefully and transparently justify their decisions. Box 2 and Fig. 3 provide an additional illustrative example.

6.2. What to do when individual studies do not provide the distribution of censoring over time

Review authors often find themselves in situations in which they must assess potential risk of bias through censoring of participants because of missing follow-up data based on only very limited information [16–19]. When the distribution of censoring over time in individual studies is not clear, but there are serious imbalances in the number of individuals for whom data are missing (e.g., individuals lost to follow-up summarized in a study flow diagram) in the study arms or the reasons for the absence of follow-up data differ among arms (e.g., provided in a study flow diagram), we suggest, in accordance with the risk of bias 2.0 tool, concern for a high risk of bias (“probably yes”) for an individual study outcome [23,24]. To derive a decision, the instructions for risk of bias due to loss to follow-up in binary data analysis from the GRADE guideline on study limitations (risk of bias) should be considered [6]. For time-to-event analyses from individual studies that do not report information regarding the distribution of censoring over time, its degree, and reasons, we suggest explicitly stating that a judgment was not possible because the required information was absent.

6.3. Individual participant data would be desirable to assess the risk of bias

Within-study sensitivity analyses for censoring, such as best/worst-case scenarios and other imputation procedures, require individual participant data. If data on individual failure and censoring times and reasons are available,

Box 2 Example 2: Martin et al. [36].

The randomized, double-blind, placebo-controlled ExteNET study compared adjuvant neratinib and placebo in patients with HER2-positive breast cancer after standard locoregional treatment, trastuzumab, and chemotherapy. The 5-year analysis of the primary end point invasive disease-free survival which was defined as time from randomization to first occurrence of invasive disease and recurrences or all cause death showed a significant benefit for the intervention. Hazard ratios were derived from a Cox proportional hazards model, and individuals were censored for the primary end point when they did not consent for additional follow-up at the date of their last physical examination, if disease recurrence did not occur within the 2 years of follow-up in this study or if they did not have a disease-free survival event within the relevant time frame (5.6 months). In each treatment arm, 1,420 participants were randomized and included in the intention-to-treat analysis.

While the study publication did not specify the proportion of censored individuals and the respective reasons for censoring, the survival curve for the primary outcome (Fig. 3) shows severe imbalances in the number of censored individuals. The number of censored participants between the time points is reported together with the number of participants at risk to experience the event for certain time points below the curves and for each study arm, respectively. The percentages present the proportion of participants who are event-free for the respective time points. The number of censored individuals in the experimental arm is substantially higher than in the placebo arm, especially in the early observational period. This results in a lower number of individuals is at risk, excluding those who have experienced the event of interest or were censored, at any time point thereafter in the favored experimental arm. Assessing the times for the beginning of accrual (July 9, 2009), the ending of accrual (October 24, 2011), and the end of the 5 year follow-up (March 1, 2017), one can be certain that the early censoring events were due to loss to follow-up, and not to “end-of-follow-up,” because the minimum complete observation time was at least 5.4 years (from October 24, 2011 to March 1, 2017). Given the information outlined previously, a judgment of high risk of bias for this study due to censoring of participants because of missing follow-up data is justifiable. In a hypothetical situation, where a body of evidence for a certain outcome consists solely of this example, we would consider rating down for study limitations.

lead to decisions to rate down the certainty of evidence. Available statistical tests for the independence assumption also require additional data [37] and are usually impossible to perform, when conducting a standard systematic review. Simple quantification measures for the completeness of follow-up in survival analyses also exist but are usually not included in the study reports.

6.4. Rating the risk of bias resulting from censoring of participants because of missing follow-up data and deriving an overall judgment for an individual study

A judgment on the risk of bias associated with missing data for time-to-event outcomes within GRADE should be based on the principles outlined in previous guidelines for rating the quality of the evidence addressing study limitations (GRADE guideline 4), particularly with regard to the risk of bias associated with missing participant outcome data in a body of evidence for both binary and continuous outcomes (GRADE guideline 17) [6,7]. The assessment criteria specified in this guidance allow integration of time-to-event—specific differences (e.g., censoring of individuals for whom data are missing and those who ended follow-up appropriately) and to support a decision on the presence of a risk of bias.

Table 1 provides considerations that reviewers can use to estimate the extent of the risk of bias introduced by censoring of participants because of missing data in an individual study. To derive a decision on the impact of missing follow-up data on the overall risk of bias for an outcome in an individual study reviewers must consider all other potential study limitations including lack of allocation concealment or the lack of blinding following which they can judge the risk of bias can follow usual GRADE principles [6]. A crucial limitation in one risk of bias criterion, which may include substantial differences in the degree and distribution in the amount of early and late censoring, or several criteria with some limitations, which may include considerable difference in the overall degree of censoring, may be sufficient to merit a judgment of a serious limitation. A crucial limitation for one or more criteria would result in a judgment of a very serious limitation for the outcome of an individual study [6]. These judgments should then inform an overall rating of the GRADE risk of bias domain for a body of evidence.

7. Making an overall judgment for a body of evidence

To derive a judgment for the risk of bias domain across studies in a body of evidence, reviewers should apply the usual GRADE principles for study imitations [6]: no serious limitations (do not rate down), if evidence comes largely from studies at low risk of bias; serious limitations (rate down one level), if evidence comes largely from studies at high risk of bias; very serious limitations (rate

individual patient data meta-analyses for time-to-event outcomes would allow for a more elaborate assessment of the sensitivity of results to missing data issues.

For example, such analyses may be possible when data for individuals lost to follow-up can be imputed based on plausible assumptions for individuals for whom data are missing [7]. Significant changes in the estimates could then

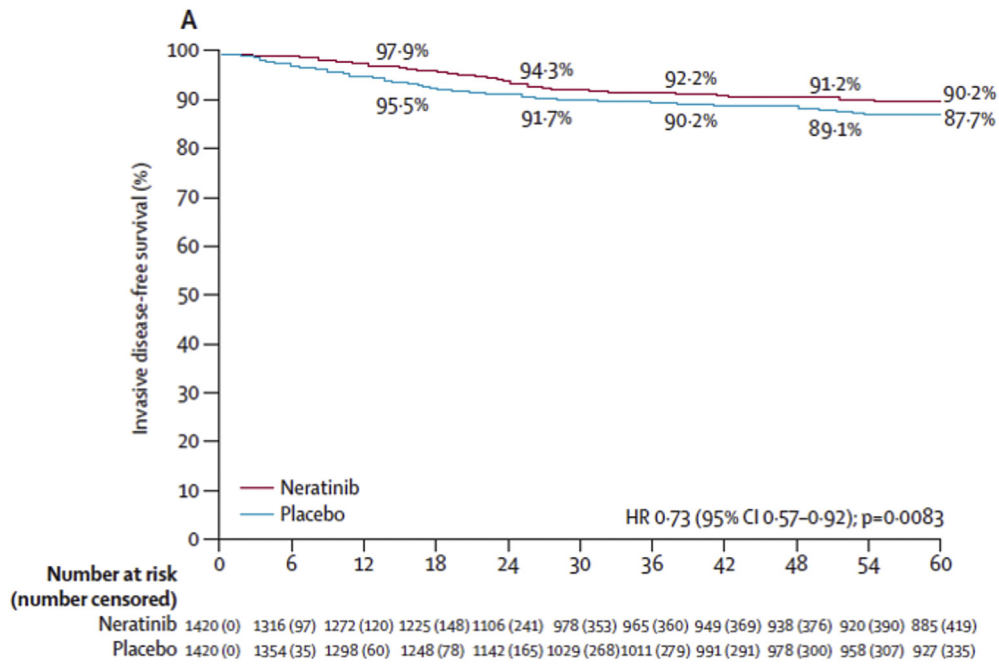


Fig. 3. Kaplan–Meier curve for the outcome invasive disease-free survival from the study by Martin et al. [36] (see Box 2). The number of individuals censored up to the respective time-points of follow-up is reported along the number of individuals at risk to experience the outcome at this time point. The number of censored individuals is substantially higher in the neratinib arm throughout the follow-up period. The number of individuals at risk (excluding those who experienced the event or were censored) in the placebo arm is substantially higher than the number of individuals at risk in the neratinib arm. Nonetheless, the neratinib arm is shown to be beneficial by the HR (<1). Adapted from “neratinib after trastuzumab-based adjuvant therapy in HER2-positive breast cancer (ExteNET): 5-year analysis of a randomized, double-blind, placebo-controlled, phase 3 trial” by Martin et al., 2017, The Lancet Oncology, 18(12), p. 1694. Copyright 2017 by Elsevier. Reprinted with permission. HR, hazard ratio.

down two levels), if evidence comes largely from studies at very high risk of bias.

If studies vary in their risk of bias, and the results differ in high and low risk of bias studies, reviewers may base best evidence summaries on the lower risk of bias studies [6]. In particular, in an appropriately large set of studies, when the potential risk of bias due to censoring of participants with missing lost to follow-up data differs across studies, reviewers can conduct sensitivity analysis to determine whether results differ in high and low risk of bias studies. When results differ, reviewers should present best estimates from only low risk of bias studies.

8. Discussion and further guidance for the assessment of time-to-event evidence

For this guide, we chose the prior outlined definitions and concepts, but they are not unassailable. Well-known resources for the conduct of systematic reviews focus on the hazard ratio as a relative effect measure to include time-to-event data in meta-analyses [38]. Therefore, our guidance focuses on the hazard ratio as the relative effect measure for time-to-event analysis. In time-to-event analysis, certain competing risk analyses require censoring of competing events, meaning single or multiple events precluding the occurrence of the event of interest [13,39].

Table 1. Decision support for judgments of a risk of bias through inappropriate censoring in an individual study

Indicators	Considerations for the risk of bias through censoring of participants with missing follow-up data assessment in individual studies
Time point of censoring considerably different in both arms (early versus late censoring)	Critical concern for high risk of bias as early censoring is more likely to be due to missing data (e.g., loss to follow-up) as opposed to the end of study censoring.
Censoring degree among arms diverging (overall number of censored patients reported but distribution over time not known)	A high risk of bias is more likely as a different degree and differing reasons for censoring are contradicting with a valid randomization process and thus imply that missingness may depend on the received intervention [23]
If reasons for censoring are reported (e.g., summarized in a study flow diagram): Different reasons why data for individuals were missing (e.g., were lost to follow-up) and different degree between arms.	

Nevertheless, such analyses remain susceptible to bias due to censoring of participants because of missing follow-up data when individuals are excluded from follow-up and censored for other reasons. An exception occurs when study authors applied competing risk analysis methods to account for the particular reasons data are absent, e.g., loss to follow-up, in their primary analysis.

To illustrate the issues outlined in this guidance, we present examples from randomized trials; some considerations are, however, also applicable to nonrandomized studies with control arms. In the absence of randomization, confounders may introduce bias because of an association between censoring time and the outcome of interest and the control of such confounders plays a critical role [40]. We acknowledge possible subsequent progress of the field and will adapt this guidance as necessary.

A great variety of additional approaches to analyze time-to-event data applies less frequently for primary analyses and rarely finds their way into meta-analyses. Investigators have proposed numerous analytic techniques to test the sensitivity of single trial results to the dependence of censoring, several of which are based on multiple imputation and account for the dependence of follow-up, taking the distribution of survival events into account.

These approaches are not solely applicable to the Cox model, but address Kaplan-Meier estimators, parametric proportional hazards models, and other analysis techniques. Practical applications of the methods show substantial bias when the survival expectation of the censored individuals alters in a negative or positive manner from the expectation of the individuals remaining on the study [41–49]. Computationally, more advanced methods, including approaches that explicitly allow for adjustment of dependent censoring are based on strict assumptions, require detailed data and are currently used only for exploratory purposes. When the results of such procedures are available, they can support a judgment on the consequences of censoring [50–52].

Because the occurrence of adverse events is usually carried out as binary data analysis in contingency tables, censoring is an important threat to the validity of safety analyses. However, when comparing adverse events among study arms, all individuals should be observed for a similar time period to allow a fair comparison of interventions. Censoring of participants from individual study arms, for example, because of competing events such as switching treatment after disease progression, the results in varying observation times among participants and subsequently in diverging average times are at risk for adverse events. Bender et al. [53] pointed out specific situations in which the risk of bias due to inadequate analysis of adverse events led to significant reductions of the certainty in the evidence in evaluations to inform reimbursement decisions for new drugs by relevant authorities in Germany as “greater harm could not be excluded with sufficient certainty.” Analysis of safety end points by means of appropriate time-to-event analysis techniques should be common practice [54].

CRedit authorship contribution statement

Marius Goldkuhle: Writing - original draft, Methodology, Writing - review & editing, Conceptualization, Project administration. **Ralf Bender:** Writing - original draft, Writing - review & editing, Methodology, Conceptualization. **Elie A. Akl:** Writing - original draft, Writing - review & editing, Methodology, Conceptualization. **Elvira C. van Dalen:** Writing - original draft, Writing - review & editing, Methodology, Conceptualization. **Sarah Nevitt:** Writing - original draft, Writing - review & editing, Methodology, Conceptualization. **Reem A. Mustafa:** Writing - original draft, Writing - review & editing, Methodology, Conceptualization. **Gordon H. Guyatt:** Writing - original draft, Writing - review & editing, Methodology, Conceptualization. **Marielene Trivella:** Writing - original draft, Writing - review & editing, Methodology, Conceptualization. **Benjamin Djulbegovic:** Writing - original draft, Writing - review & editing, Methodology, Conceptualization. **Holger Schünemann:** Writing - original draft, Writing - review & editing, Methodology, Conceptualization. **Michela Cinquini:** Writing - original draft, Writing - review & editing, Methodology, Conceptualization. **Nina Kreuzberger:** Writing - original draft, Writing - review & editing, Methodology, Conceptualization. **Nicole Skoetz:** Writing - original draft, Writing - review & editing, Methodology, Conceptualization, Supervision.

Acknowledgments

All authors are members of the GRADE working group.

Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.jclinepi.2020.09.017>.

References

- [1] Guyatt GH, Oxman AD, Kunz R, Brozek J, Alonso-Coello P, Rind D, et al. GRADE guidelines 6. Rating the quality of evidence—imprecision. *J Clin Epidemiol* 2011;64:1283–93.
- [2] Guyatt GH, Oxman AD, Kunz R, Woodcock J, Brozek J, Helfand M, et al. GRADE guidelines: 8. Rating the quality of evidence—indirectness. *J Clin Epidemiol* 2011;64:1303–10.
- [3] Guyatt GH, Oxman AD, Kunz R, Woodcock J, Brozek J, Helfand M, et al. GRADE guidelines: 7. Rating the quality of evidence—inconsistency. *J Clin Epidemiol* 2011;64:1294–302.
- [4] Guyatt GH, Oxman AD, Montori V, Vist G, Kunz R, Brozek J, et al. GRADE guidelines: 5. Rating the quality of evidence—publication bias. *J Clin Epidemiol* 2011;64:1277–82.
- [5] Guyatt GH, Oxman AD, Sultan S, Glasziou P, Akl EA, Alonso-Coello P, et al. GRADE guidelines: 9. Rating up the quality of evidence. *J Clin Epidemiol* 2011;64:1311–6.
- [6] Guyatt GH, Oxman AD, Vist G, Kunz R, Brozek J, Alonso-Coello P, et al. GRADE guidelines: 4. Rating the quality of evidence—study limitations (risk of bias). *J Clin Epidemiol* 2011;64:407–15.

- [7] Guyatt GH, Ebrahim S, Alonso-Coello P, Johnston BC, Mathioudakis AG, Briel M, et al. GRADE guidelines 17: assessing the risk of bias associated with missing participant outcome data in a body of evidence. *J Clin Epidemiol* 2017;87:14–22.
- [8] Kahale LA, Diab B, Brignardello-Petersen R, Agarwal A, Mustafa RA, Kwong J, et al. Systematic reviews do not adequately report or address missing outcome data in their analyses: a methodological survey. *J Clin Epidemiol* 2018;99:14–23.
- [9] Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. *J Am Stat Assoc* 1958;53:457–81.
- [10] Cox DR. Regression models and life-tables. *J R Stat Soc Ser B Methodol* 1972;34(2):187–220.
- [11] Tierney JF, Stewart LA, Ghersi D, Burdett S, Sydes MR. Practical methods for incorporating summary time-to-event data into meta-analysis. *Trials* 2007;8:16.
- [12] Leung K-M, Elashoff RM, Afifi AA. Censoring issues IN survival analysis. *Annu Rev Public Health* 1997;18(1):83–104.
- [13] Kleinbaum DG, Klein M. *Survival Analysis*. 3 ed. New York: Springer-Verlag; 2012.
- [14] Putter H, Fiocco M, Geskus RB. Tutorial in biostatistics: competing risks and multi-state models. *Stat Med* 2007;26:2389–430.
- [15] Lagakos SW. General right censoring and its impact on the analysis of survival data. *Biometrics* 1979;35:139–56.
- [16] Batson S, Greenall G, Hudson P. Review of the reporting of survival analyses within randomised controlled trials and the implications for meta-analysis. *PLoS One* 2016;11:e0154870.
- [17] Abaira V, Muriel A, Emparanza JI, Pijoan JI, Royuela A, Plana MN, et al. Reporting quality of survival analyses in medical journals still needs improvement. A minimal requirements proposal. *J Clin Epidemiol* 2013;66:1340–1346.e5.
- [18] Altman DG, De Stavola BL, Love SB, Stepniowska KA. Review of survival analyses published in cancer journals. *Br J Cancer* 1995;72:511–8.
- [19] Mathoulin-Pelissier S, Gourgou-Bourgade S, Bonnetain F, Kramar A. Survival end point reporting in randomized cancer clinical trials: a review of major journals. *J Clin Oncol* 2008;26:3721–6.
- [20] Vervölgyi E, Kromp M, Skipka G, Bender R, Kaiser T. Reporting of loss to follow-up information in randomised controlled trials with time-to-event outcomes: a literature survey. *BMC Med Res Methodol* 2011;11:130.
- [21] Altman DG. *Practical Statistics for Medical Research*. London: Chapman & Hall; 1999: ISBN 0-412-27630-5.
- [22] Pocock SJ, Clayton TC, Altman DG. Survival plots of time-to-event outcomes in clinical trials: good practice and pitfalls. *Lancet* 2002;359(9318):1686–9.
- [23] Higgins JPT, Sterne JAC, Savović J, Page MJ, Hróbjartsson A, Boutron I, et al. A revised tool for assessing risk of bias in randomized trials 2019. Available at <https://sites.google.com/site/riskofbiastool/welcome/rob-2-0-tool>. Accessed August 6, 2019.
- [24] Sterne J, Savović J, Page M, Elbers R, Blencowe N, Boutron I, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. *BMJ* 2019;366:14898.
- [25] Cochrane Community. Glossary: The Cochrane Collaboration. 2019. Available at <https://community.cochrane.org/glossary>. Accessed August 6, 2019.
- [26] Kahale LA, Guyatt GH, Agoritsas T, Briel M, Busse JW, Carrasco-Labra A, et al. A guidance was developed to identify participants with missing outcome data in randomized controlled trials. *J Clin Epidemiol*.
- [27] Montori VM, Guyatt GH. Intention-to-treat principle. *CMAJ* 2001;165(10):1339–41.
- [28] Brinkhof MWG, Pujades-Rodriguez M, Egger M. Mortality of patients lost to follow-up in antiretroviral treatment programmes in resource-limited settings: systematic review and meta-analysis. *PLOS ONE* 2009;4:e5790.
- [29] Brinkhof MWG, Spycher BD, Yiannoutsos C, Weigel R, Wood R, Messou E, et al. Adjusting mortality for loss to follow-up: analysis of five ART programmes in sub-Saharan Africa. *PLoS One* 2010;5:e14149.
- [30] Campigotto F, Weller E. Impact of informative censoring on the Kaplan-Meier estimate of progression-free survival in phase II clinical trials. *J Clin Oncol* 2014;32:3068–74.
- [31] Persson I, Khamis H. Bias of the Cox model hazard ratio. *J Mod Appl Stat Methods* 2005;4(1):90–9.
- [32] Denis F, Lethrosne C, Pourel N, Molinier O, Pointreau Y, Domont J, et al. Randomized trial comparing a web-mediated follow-up with routine surveillance in Lung cancer patients. *J Natl Cancer Inst* 2017;109.
- [33] Guyot P, Ades AE, Ouwens MJNM, Welton NJ. Enhanced secondary analysis of survival data: reconstructing the data from published Kaplan-Meier survival curves. *BMC Med Res Methodol* 2012;12:9.
- [34] Kahale LA, Diab B, Khamis AM, Chang Y, Lopes LC, Agarwal A, et al. Potentially missing data are considerably more frequent than definitely missing data: a methodological survey of 638 randomized controlled trials. *J Clin Epidemiol* 2019;106:18–31.
- [35] Parmar MKB, Torri V, Stewart L. Extracting summary statistics to perform meta-analyses of the published literature for survival endpoints. *Stat Med* 1998;17:2815–34.
- [36] Martin M, Holmes FA, Ejlertsen B, Delaloge S, Moy B, Iwata H, et al. Neratinib after trastuzumab-based adjuvant therapy in HER2-positive breast cancer (ExteNET): 5-year analysis of a randomised, double-blind, placebo-controlled, phase 3 trial. *Lancet Oncol* 2017;18(12):1688–700.
- [37] Lee S-Y, Wolfe RA. A simple test for independent censoring under the proportional hazards model. *Biometrics* 1998;54:1176–82.
- [38] Higgins JPT, Li T, Deeks JJ. Chapter 6: choosing effect measures and computing estimates of effect. Draft version (29 January 2019) for inclusion. In: Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, et al, editors. *Cochrane Handbook for Systematic Reviews of Interventions*.
- [39] 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.
- [40] Hernán MA, Hernández-Díaz S, Robins JM. A structural approach to selection bias. *Epidemiology* 2004;15:615–25.
- [41] Emoto SE, Matthews PC. A weibull model for dependent censoring. *Ann Stat* 1990;18(4):1556–77.
- [42] Jackson D, White IR, Seaman S, Evans H, Baisley K, Carpenter J. Relaxing the independent censoring assumption in the Cox proportional hazards model using multiple imputation. *Stat Med* 2014;33:4681–94.
- [43] Faucett CL, Schenker N, Taylor JM. Survival analysis using auxiliary variables via multiple imputation, with application to AIDS clinical trial data. *Biometrics* 2002;58:37–47.
- [44] Huang X, Wolfe RA. A frailty model for informative censoring. *Biometrics* 2002;58:510–20.
- [45] Kaciroti NA, Raghunathan TE, Taylor JM, Julius S. A Bayesian model for time-to-event data with informative censoring. *Biostatistics (Oxford, England)* 2012;13(2):341–54.
- [46] Hsu C-H, Taylor JMG, Murray S, Commenges D. Survival analysis using auxiliary variables via non-parametric multiple imputation. *Stat Med* 2006;25:3503–17.
- [47] Siannis F. Applications of a parametric model for informative censoring. *Biometrics* 2004;60:704–14.
- [48] Siannis F. Sensitivity analysis for multiple right censoring processes: investigating mortality in psoriatic arthritis. *Stat Med* 2011;30:356–67.
- [49] Siannis F, Copas J, Lu G. Sensitivity analysis for informative censoring in parametric survival models. *Biostatistics (Oxford, England)* 2005;6(1):77–91.
- [50] Robins JM, Finkelstein DM. Correcting for noncompliance and dependent censoring in an AIDS clinical trial with inverse probability of censoring weighted (IPCW) log-rank tests. *Biometrics* 2000;56:779–88.

- [51] Cole SR, Hernán MA. Constructing inverse probability weights for marginal structural models. *Am J Epidemiol* 2008;168:656–64.
- [52] Tsiatis AA, Robins JM. Correcting for non-compliance in randomized trials using rank preserving structural failure time models. *Commun Stat - Theor Methods* 1991;20(8):2609–31.
- [53] Bender R, Beckmann L, Lange S. Biometrical issues in the analysis of adverse events within the benefit assessment of drugs. *Pharm Stat* 2016;15(4):292–6.
- [54] Allignol A, Beyersmann J, Schmoor C. Statistical issues in the analysis of adverse events in time-to-event data. *Pharm Stat* 2016;15(4):297–305.