On time-to-event outcomes in evidence syntheses of randomized controlled trials: methods, challenges and guidance

Inaugural Dissertation

zur

Erlangung des Doktorgrades
philosophiae doctor (PhD)
der Medizinischen Fakultät
der Universität zu Köln

vorgelegt von

Marius Goldkuhle

aus Essen

Hundt Druck GmbH, Köln

2024

Betreuerin:	Prof. Dr. Nicole Skoetz
Referenten:	Prof. Dr. Thorsten Simon
	Prof. Dr. Achim Tresch
	Prof. Dr. Martin Hellmich
Datum der Mündlichen Prüfung:	26.09.2024

Für Andrea,
Christian,
Cornelius,
Marcus
&
Theresa, die mich immer begleiten.

“*Since all models are wrong the scientist must be alert to what is importantly wrong.*
It is inappropriate to be concerned about mice when there are tigers abroad.”

Box, G.E.P. 1976.
Science and Statistics. Journal of the American Statistical Association. 71:791-99.

“*Good reporting is not an optional extra; it is an essential component of research.*”

Altman D.G., Simera I., Hoey J., Moher D., Schulz K. 2008.
EQUATOR: reporting guidelines for health research. Open Medicine. 2(2):e49-50.

Table of contents

Table of figures

Table of equations

Table of abbreviations

CI	Confidence interval
E	Expected events
GDT	Guideline Development Tool
GRADE	Grading of Recommendations Assessment, Development and Evaluation
HR	Hazard ratio
O	Observed events
PRISMA	Preferred Reporting Items for Systematic reviews and Meta-Analyses
RCT	Randomized controlled trial
RMST	Restricted mean survival time

1. Zusammenfassung

Time-to-event Analysen, oder Ereigniszeitanalysen, sind methodologisch anspruchsvoll, unterliegen spezifischen Grundannahmen und ihre Effektschätzer, insbesondere das Hazard Radio (HR), sind nicht immer einfach zu interpretieren. Frühere Untersuchungen wiesen auf deutliche Einschränkungen bei der Berichterstattung zu Ereigniszeitanalysen in Studienpublikationen hin. Autoren von Evidenzsynthesen stehen also vor besonderen Herausforderungen, wenn sie Ereigniszeitendpunkte aus randomisierten Studien (RCTs) in Meta-Analysen einschließen. Ziel dieses Promotionsprojekts war es, diese Herausforderungen anhand von drei meta-epidemiologischen Studien erstmalig zu charakterisieren und gezielte Handlungsanleitungen zu entwickeln, um den Umgang mit ihnen zu erleichtern.

Eine meta-epidemiologische Studie demonstrierte deutliche Unterschiede in den Eigenschaften, Methoden und der Berichterstattung von 217 Meta-Analysen von Ereigniszeitendpunkten in den insgesamt 100 eingeschlossenen systematischen Reviews. Große Einschränkungen zeigten sich bei der Berichterstattung zu den untersuchten Endpunkten sowie zu allgemeiner und Ereigniszeit-spezifischer Reviewmethodik. Besonders selten berücksichtigten die Reviewautoren jene RCT-Merkmale, welche Einfluss auf die Zuverlässigkeit von Ereigniszeitanalysen nehmen können, etwa die informative Zensierung und nicht-proportionale Hazards.
Eine zweite meta-epidemiologische Studie untersuchte systematisch die Eigenschaften, Methoden und Berichterstattung von 315 Ereigniszeitanalysen in 235 RCTs, die in Meta-Analysen von Ereigniszeitendpunkten eingeschlossen gewesen sind. Diese Arbeit demonstrierte erhebliche Variabilität in der Methodik der RCTs, zum Beispiel bei den verfügbaren Ereigniszeitdaten, wie HRs, Plots und P-Werten. Besonders große Einschränkungen wurden hier zum Beispiel bei der Berichterstattung zu Endpunkten, Zensierung und dem Follow-Up deutlich. Auch in den RCT-Publikationen selbst wurden Studieneigenschaften mit Einfluss auf die Zuverlässigkeit der Ereigniszeitanalysen nur selten berücksichtigt.
Schließlich untersuchte eine dritte meta-epidemiologische Studie die Berechnung absoluter Effektmaße (z.B. natürliche Häufigkeiten oder die Number-needed-to-treat) von Meta-Analysen zu Ereigniszeitendpunkten in 96 Cochrane Reviews. Sie zeigte, dass diese Schätzer häufig falsch bezeichnet und teilweise so berechnet wurden, dass sich die Effektrichtung des zugrundeliegenden, gepoolten HRs umkehrte.
Um den identifizierten Problemen entgegenzuwirken, wurden zwei systematische Leitlinien nach den Standards der Grading of Recommendations Assessment, Development and Evaluation (GRADE) Working Group entwickelt. Eine dieser Leitlinien befasst sich mit dem Umgang mit informativer Zensierung in Evidenzsynthesen, stellt Methoden zur Erkennung eines entsprechen Verzerrungsrisikos in Studienpublikationen vor und erklärt, wie die Beurteilungen für einzeln Studien in einen Evidenzkörper übersetzt werden können. Die zweite GRADE Leitlinie erörtert Alternativen zur Berechnung absoluter Effekte für Time-to-event Endpunkte und diskutiert die Auswahl geeigneter Variablen sowie mögliche Einschränkungen der einzelnen Berechnungsansätze.

Die Arbeiten dieser Dissertation zeigen erhebliche Probleme in Meta-Analysen von Ereigniszeitendpunkten, den Publikationen ihrer eingeschlossenen RCTs und der Kommunikation ihrer Ergebnisse. Gezielte Leitlinien sollen nun einigen der festgestellten Mängel entgegenwirken, um so die Durchführung und die Berichterstattung von Evidenzsynthesen zu Ereigniszeitendpunkten und schließlich auch die Entscheidungen, die auf ihnen basieren, zu verbessern.

2. Summary

Time-to-event analyses are methodologically challenging, subject to distinct assumptions and their associated effect estimators, in particular the hazard ratio (HR), are not always straightforward to interpret. Previous studies suggest considerable limitations in the reporting of time-to-event analyses in study publications. Authors of evidence syntheses therefore face certain challenges when they include time-to-event analyses from randomized trials (RCTs) in their meta-analyses.

The aim of the projects composing this dissertation was to characterize these challenges for the first time by means of three meta-epidemiological studies and to develop targeted methodological guidance to support authors of evidence syntheses in dealing with them.

A meta-epidemiological study of 217 meta-analyses in 100 systematic reviews revealed significant differences in the characteristics, methods and reporting of time-to-event outcomes between the assessed reviews. Major limitations were found, for example, in the reporting of outcome definitions as well as general and time-to-event specific review methods. Review authors rarely considered characteristics of their included RCTs that may affect the reliability of time-to-event analyses, for example, informative censoring and non-proportional hazards.

A second meta-epidemiological study systematically assessed the characteristics, methods and reporting of 315 time-to-event analyses in 235 RCTs included in meta-analyses of time-to-event outcomes. It demonstrated considerable variability in the methodology of the RCTs, for example, in the available types of time-to-event data, such as HRs, plots and P-values. Major limitations were evident, for instance, in the reporting of outcome definitions, censoring and follow-up data. In the RCT publications, trial characteristics affecting the reliability of time-to-event analyses were rarely considered and discussed.

A third meta-epidemiological study examined the calculation of absolute effects (e.g., natural frequencies, number-needed-to-treat) to present the results of meta-analyses of time-to-event outcomes from a total of 96 Cochrane reviews. The study revealed that the corresponding estimates were often mislabeled and frequently calculated in a way that the direction of the effect of the underlying pooled HR was reversed.

To address the identified problems, two systematic guidelines were developed according to the standards of the Grading of Recommendations Assessment, Development and Evaluation (GRADE) Working Group.

One of these guidelines illustrates how to deal with informative censoring in evidence syntheses, presents methods for determining a corresponding risk of bias in trial publications and explains how judgements for individual trials can be translated to a body of evidence.

The second GRADE guideline presents alternatives for calculating absolute effects for time-to-event outcomes, discusses the selection of appropriate variables and possible limitations of each calculation approach.

Based on three meta-epidemiological studies, this dissertation revealed significant problems with time-to-event analyses in meta-analyses of current evidence syntheses, in the publications of their included RCTs and in the communication of their results. Targeted guidance should now counteract some of the identified shortcomings in order to improve the conduct and reporting of meta-analyses of time-to-event outcomes in the future and thus ultimately the decisions based on them.

3. Introduction

3.1. Time-to-event outcomes

3.1.1. Time-to-event outcomes and analyses

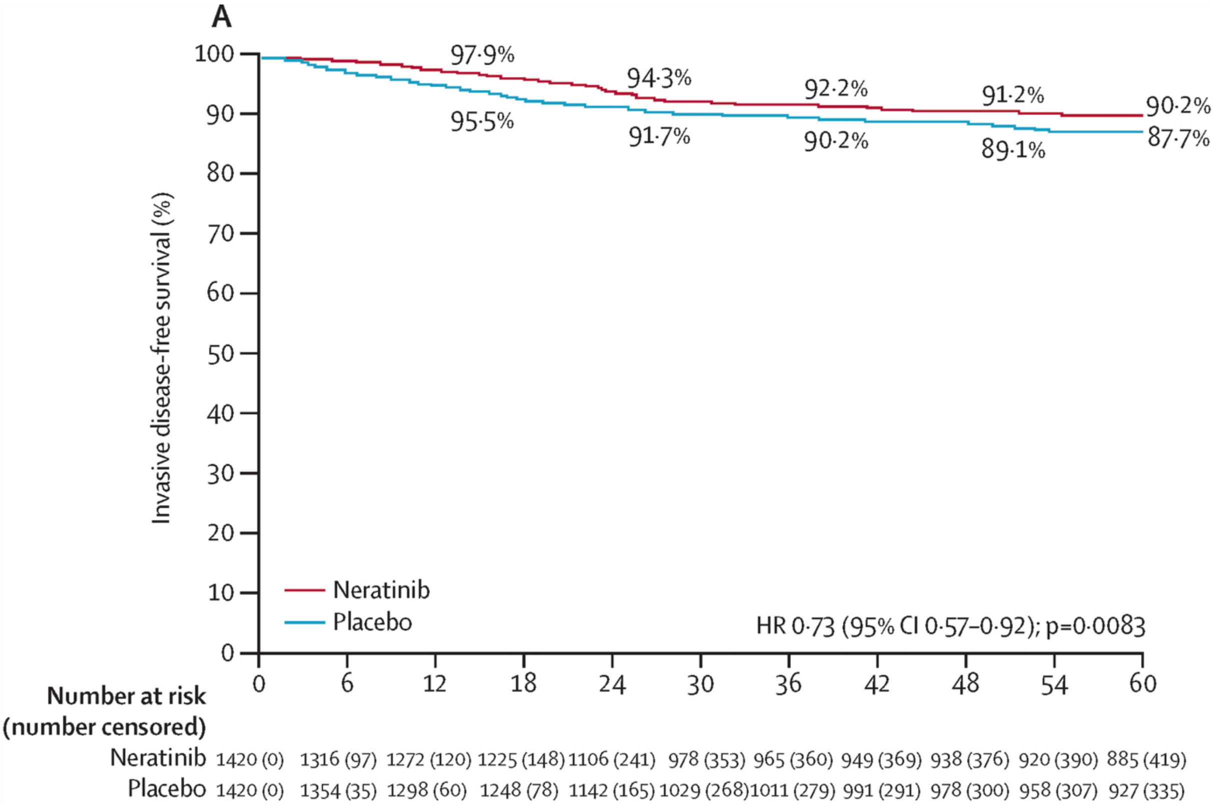
For researchers interested in the effect of an intervention on a dichotomous outcome, such as the occurrence of a certain event, outcome analysis using time-to-event methods offers significant advantages over binary outcome analysis. Unsurprisingly, the number of clinical research articles addressing time-to-event outcomes has steadily increased over the last years (1). Binary outcomes, such as relative risks, compare how often an event (e.g., death) or composite of events (e.g., death or relapse) occurs within a defined period of time. Time-to-event outcomes in addition consider the time from a given starting point until the respective events occur (2-5).

Consider, for example, a comparison of two treatments for individuals suffering from a highly fatal disease. All study participants might have died after a follow-up period of 10 years. Comparing the outcome all-cause death by binary analysis as crude rates after ten years will show no difference between the treatment groups (e.g., there are 100 individuals in each arm: (100/100)/(100/100) = 1). Comparing all-cause death as a time-to-event outcome, however, reveals a number of early deaths in the control group and that only few, if any, individuals lived close to the end of follow-up, whereas in the experimental group the overall probability of dying was lower throughout the entire follow-up period. Even though in both groups all individuals deceased after ten years, treatment in the experimental group appears justifiably beneficial with time-to-event analysis.

With limited observation times in most clinical trials, it is rarely possible to observe the occurrence of a defined outcome event in the entire enrolled population. For some study participants, the exact time point of the event of interest will be unknown. Data from individuals with unknown event times, referred to as censored data, are included by time-to-event analysis by default (2, 6, 7).[1] If, for instance, the outcome under study is death, then time-to-event analysis allows to include the information that some individuals were still alive at the end of follow-up. Through incorporating censored data, time-to-event analysis makes use of all available follow-up data and so increases the power of statistical analyses. It enables trialists to compare data of participants with variable follow-up times in a fair manner, which is particularly relevant for clinical trials with often staggered recruiting times (2, 3, 5, 8, 9).

Although they are often referred to as survival analyses, the target outcome events of time-to-event analyses are not limited to death or its absence. Any dichotomous event or combination of events, negative and positive, constitutes a possible outcome (3). It is critical, however, to consider whether a given time-to-event analysis addresses the occurrence of an event, for example, the probability for an event *P*, or its absence (1–*P*) (10, 11).

[1] The type of censoring most relevant in clinical trials, with a defined starting point of allocation, is right censoring, which I will describe by “censoring” in this dissertation. Right censoring occurs when the exact time point of the event of interest is unknown because the observation has ended (“the event is to occur on the right side of a plot”). Left censoring, referring to unclarity when exactly and whether the event has occurred before an individual’s observation has begun, and interval-censoring, that is unclarity when an event has occurred exactly within a defined interval (e.g., between two points of a clinical assessment) are well elaborated by Kleinbaum and Klein 2012 (3) and Lagakos 1979 (6).



Number at risk (number censored)	0	6	12	18	24	30	36	42	48	54	60
Neratinib	1420 (0)	1316 (97)	1272 (120)	1225 (148)	1106 (241)	978 (353)	965 (360)	949 (369)	938 (376)	920 (390)	885 (419)
Placebo	1420 (0)	1354 (35)	1298 (60)	1248 (78)	1142 (165)	1029 (268)	1011 (279)	991 (291)	978 (300)	958 (307)	927 (335)

Figure 1: Exemplary Kaplan-Meier plot. Survival curve from a RCT comparing the outcome invasive disease-free survival with neratinib or placebo in individuals with advanced renal cell carcinoma. The authors present survival probabilities at different time points and show the number of individuals at risk and the total number of censored individuals up to given follow-up time point below the plot (adapted from Martin et al. 2017 (12), p. 1694, with permission (granted in June 2024), and presented in Goldkuhle et al. 2019 (2), p. 134).

3.1.2. Kaplan-Meier method and statistical tests

Amongst the manifold methods available to draw statistical inference from time-to-event data, the method by Kaplan and Meier is particularly popular (5, 9, 13, 14). It allows to calculate median event times and time point specific event probabilities as well as the presentation of event probabilities over time in form of Kaplan-Meier plots (see figure 1 for an example) (2, 14).

Event probabilities for individuals are defined by the conditional probability of experiencing the event of interest at a certain time or within a time interval, given that the event has not yet occurred (see exemplary equation for a survival function in equation 1) (3, 5, 9, 14). Time-intervals give Kaplan-Meier plots their typical step wise appearance and their property as a “step function”. Each drop represents the occurrence of an event in one or more (resulting in a steeper drop) individuals at risk, while the steps represent the event probability within each interval.[2] The population at risk excludes those who experienced the event or ceased from the observation and were censored up to the time-interval. The intervals continue until no individuals remain under risk, thereby creating a continuous curve (3, 5, 9, 15).

[2] If the outcome of interest is the probability of the absence of an event at a given time, e.g., survival or progression-free survival, the occurrence of an event, e.g., death or disease progression, is indicated by a drop in the curve. This is common, e.g., in oncology research (figure 1). Whereas, if the outcome of interest is the probability of an event at a given time, e.g., mortality or disease progression, the occurrence of events is indicated by a rise. This is common, e.g., in cardiology (see also equation 1).

Illustration of the Kaplan-Meier survival function

T = Survival time
t = Specific value for *T*
S(*t*) = Survival function = P(*T*>*t*) = Probability of being alive at time *t*, e.g., probability to survive 10 years is *T*>*t* = 10
S(*t*-1) = Probability of being alive at time *t*-1
n = Number of individuals known to be event-free before time point *t*, excluding individuals who experienced events or were censored = number of individuals at risk to experience the event at time *t*
d = Number of events at time *t*

$$S(t) = S(t-1)\left(1 - \frac{d}{n}\right)$$

For illustration, the formulation of the Kaplan-Meier function presented here represents a survival function where *S*(*t*=0) = 1 and the probability of *not* having experienced the event up to a certain time point *t* is of interest.
The Kaplan-Meier function can also be expressed for the cumulation of events with a rising probability of experiencing an event of interest and *S*(*t*=0) = 0, e.g., the probability of having experienced a relapse of disease up to time *t*.

Equation 1: Kaplan-Meier survival function (adapted from Kleinbaum and Klein 2012 (3) and Clark et al. 2003 (9)).

The Kaplan-Meier method is non-parametrical and does not require any assumptions about the underlying distributions of the event times, for example, the survival times, in the analyzed groups. Even though adjustment is technically possible, it is predominately performed in a univariate fashion (4). This flexibility makes it attractive for a wide range of situations, particularly for those with unknown or unpredictable event distributions in the population of interest.
Kaplan-Meier estimates can be extended by confidence intervals (CI), which grow wider as the observation time increases, because the sample size, the individuals at risk, continuously reduces through events and censoring (5, 16).

Extending the descriptive nature of Kaplan-Meier estimates, multiple statistical tests allow to compare event time distributions between groups. Particularly common is the log-rank test, which, like the Kaplan-Meier method, is non-parametrical and predominately performed in a univariate fashion (see p-value in figure 1 for an example) (4, 5).
It compares between two or more Kaplan-Meier event time distributions the number of statistically expected events, if there were no difference between the groups, to the number of actually observed events summed over time (5, 9).[3]
Manifold alterations of the log-rank test are available, for example, with adaptive weights or higher weighting of events during earlier or later follow-up (17). The application of such tests depends on the analytical context, but the standard log-rank test is most popular in trial publications and most relevant to meta-analysis authors (9).

[3] The total observed (O) and expected events (E) per group *g* are, in form of $(O_g\text{-}E_g)^2/E_g$, added over all compared groups and compared to a χ^2 distribution with *g*-1 degrees of freedom (5, 9).

3.1.3. Cox proportional hazards regression model and other methods

If adjustment for covariates is required, researchers commonly lean to the Cox proportional hazards regression model (18). In clinical trials, the exposure to treatment and its impact on the distribution of event times, e.g., survival times, is often the most important covariate. Other relevant covariates, such as age, disease stage or performance status, can be included in the model depending on the question of interest (19).

The Cox model's central advantage for between group comparisons is that it allows to produce a quantitative estimate of the difference in effects between two groups in form of the HR (see figure 1) (3, 4).

The HR expresses the ratio of the hazard functions of the compared groups. Hazard functions can be derived from survival functions (see equation 2). They can be interpreted as the momentaneous probability of a defined event given that the event has not yet occurred up to that time point or time-interval, e.g., death right at a certain time point in individuals who are still alive at that time. Although their precise interpretation is rather probabilistic, for simplicity, hazard rates have been explained as incidence rates or as velocities, which represent the momentaneous speed of the rate at which the outcome events occur (3, 8, 20).[4]

Definition of the hazard function and its relation to the Kaplan-Meier survival function

T = Survival time

t = Specific value for T

S(*t*) = Survival function

h(*t*) = Hazard function = Hazard at time interval *t* = Momentaneous probability to experience the event at time *t*, if it has not occurred to that time point

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq \mathrm{T} < t + \Delta t \,|\mathrm{T} \geq t)}{\Delta\, t} = -\frac{d}{dt}[\log S(t)]$$

Note that *d* in this equation does not represent the number of events at time *t* (as in equation 1), but that *d/dt* represents the derivative of log*S*(*t*) with regard to time *t*.

Equation 2: Definition of the hazard function (adapted from Kleinbaum and Klein 2012 (3) and Clark et al. 2003 (9)).

The HR is interpreted as the relative difference in the momentaneous probability to experience the event of interest throughout follow-up. If a HR < 1 expresses a larger rate of events in the comparator group in the denominator, a HR of 0.5, for example, represents a reduction of the risk of the event throughout follow-up in the experimental group by half (21).

The HR is not always straightforward to interpret and often simplified as a risk ratio (9). Hazard ratios and risk ratios are equal only when the event times and the censoring rates are not included in the calculation of the HR. This would be the case, for example, in a trial situation where all individuals begin their observation interval at the same point in time and provide complete follow-up data for their foreseen observation period (2).

[4] The correct interpretation of hazards, according to Spruance et al. 2004 (8), is not one of velocity or "*speed*", as distance per time, but rather the "*rate of events per person-time*" (8) which "*can only be inferred in a probabilistic sense from the occurrence of events in a population of at-risk individuals during a follow-up time interval*" (8). Accordingly, HRs are not directly interpretable as a relative difference in the velocity of an event occurring after or at a time point, but rather as the odds that an event will occur in one group before it will occur in the other group. A HR of 2 is thereafter interpreted as a 67% chance that the event will first occur in group in the numerator of the HR, rather than in the group in the denominator of the HR (8).

By its properties, the Cox model is semiparametric and consists of a baseline hazard that does not follow a fixed distribution of event times in the compared groups (3-5). This allows to derive HRs, as log HRs, from the model's coefficients without the necessity to determine the baseline hazard function, as visible in equations 3 (4, 22).
This flexibility makes the Cox model the predominant choice of regression analysis in current clinical trials to address time-to-event outcomes (8, 23-25). Alternative regressions models that produce HRs, namely the regression model proposed by Fine and Gray (26), are relevant under certain situations and are mentioned in chapter 3.2.3 of this dissertation.

Cox proportional hazards regression model structure, example and proportional hazards assumption

h(*t*) = Hazard function
h$_0$ as baseline hazard = Hazard if all covariates (X_i)=0,
b_1, b_2, … b_i = Coefficients measuring the effect of the covariates
X_1, X_2, …, X_i = Covariates, e.g., treatment groups, age groups, etc.
θ = Constant over *t*

1. Definition of the Cox model for a single individual or group:

$$h(t) = h_0(t) \times exp^{b_1X_1+b_2X_2+\cdots+b_iX_i}$$

*2. Definition of the Cox model hazard ratio (HR) for two individuals or groups X** and *X:*

$$\mathrm{HR} = \frac{h(t,X^*)}{h(t,X)} = exp^{b_1(X_1^*-X_1)\,+\,b_2(X_2^*-X_2)\,+\,\ldots\,+\,b_i(X_i^*-X_i)}$$

3. Example with two treatment groups (X = experimental, X = comparator) and a single binary predictor variable, e.g., a treatment indicator (experimental with X_1* = 1, comparator with X_1 = 0):*

$$\mathrm{HR} = \frac{h_{experimental}(t)}{h_{comparator}(t)} = exp^{b(experimental-comparator)}$$

4. Definition of the proportional hazards assumption of the Cox model HR

$$\frac{\mathrm{h}(t,X^*)}{\mathrm{h}(t,X)} = \theta$$

And

$$\mathrm{h}(t,X^*) = \theta h(t,X)$$

Equations 3: Definition of the Cox model, the Cox model hazard ratio, example and definition of the proportional hazards assumption (adapted from Kleinbaum and Klein 2012 (3), Bradburn et al. 2003 (19) and Clark et al. 2003 (9)).

Finally, besides non-parametric methods and the semi-parametric Cox model, time-to-event outcome analysis can be conducted in a parametrical fashion, using models which also allow to calculate HRs. In such cases, the hazard functions are assumed to follow a defined statistical distribution (4, 19, 27, 28).[5]

[5] Prominent distributions in the parametrical modelling of event functions include, for example, exponential distributions and log-Normal distributions as well as the Weibull and the Gompertz distributions,

Possible, alternative analysis options for time-to-event data that do not produce HRs are accelerated failure time models or the estimation of the restricted mean survival time (RMST) (4, 28, 30, 31).[6] Their estimates are not frequently presented in current clinical trials and available meta-analyses are based, almost univocally, on Cox model HRs. Particularly the RMST might become increasingly relevant in meta-analyses for certain situations, as discussed later in this dissertation.

3.2. Methodological challenges of time-to-event analyses

Despite their attractiveness in terms of statistical power and flexibility as well as their widespread application clinical research, time-to-event analyses are associated with methodological difficulties.

3.2.1. Censoring

The central assumption to consider censored data in time-to-event analytic methods is that censoring occurs non-informative. Non-informative describes that the censoring time point of an individual must not hold any information regarding her or his event time (3, 6, 7, 32, 33).
More practically relevant are the assumptions of random and independent censoring (3).[7] Censoring is considered random and independent when the ideal, but hypothetical distribution function of event times, if all individuals were observed until their events, and the hypothetical censoring distribution, if all individuals were censored, are independent (2, 3, 6).
Potentially more straightforward: censoring must occur in such a way that the event probability of those whose observation time is censored is representative for those who remain under observation until the event of interest or until administrative closure of the trial (2).

which incorporate characteristics that make them particularly attractive for modelling survival times (4, 9, 29). Parametrical modelling of time-to-event outcomes is not further discussed in this dissertation for several reasons: first, because it is of lesser importance in clinical trials, e.g., because of limited flexibility (parametrical models require assumptions about the underlying distributions and endanger misspecification) as well as a lack of standardisation and routine in clinical trial methodology (13, 27). Second, because pooling of estimates from parametrical survival models in meta-analyses is not common practice. Nevertheless, if the chosen distribution for a parametrical model well fits the target distributions, the resulting estimates are more informative, precise and more powerful than corresponding non-parametrical estimates (4).

[6] The restricted mean survival time (RMST) uses the area under a Kaplan-Meier curve, from beginning of the observation time up to a defined time point, the "*restriction time*", as estimate (28). For a survival curve, where the outcome is death, e.g., the RMST presents the cumulative survival of a group from beginning of an observation to the restriction time point. A difference between groups can be presented as difference in RMST, which is interpretable as the cumulative duration of expected survival that is gained (or lost) with one treatment or the other.

[7] The three assumptions of censoring are conceptually challenging and their interpretation as well as their relation differ in the literature. Very often the terms are used interchangeably. Kleinbaum and Klein (3) provide definitions and examples in their book. Clark et al. 2003 (32) describe informative censoring simply as "*(…) when individuals are lost to follow-up for reasons that may relate to their (unknown) outcome*".

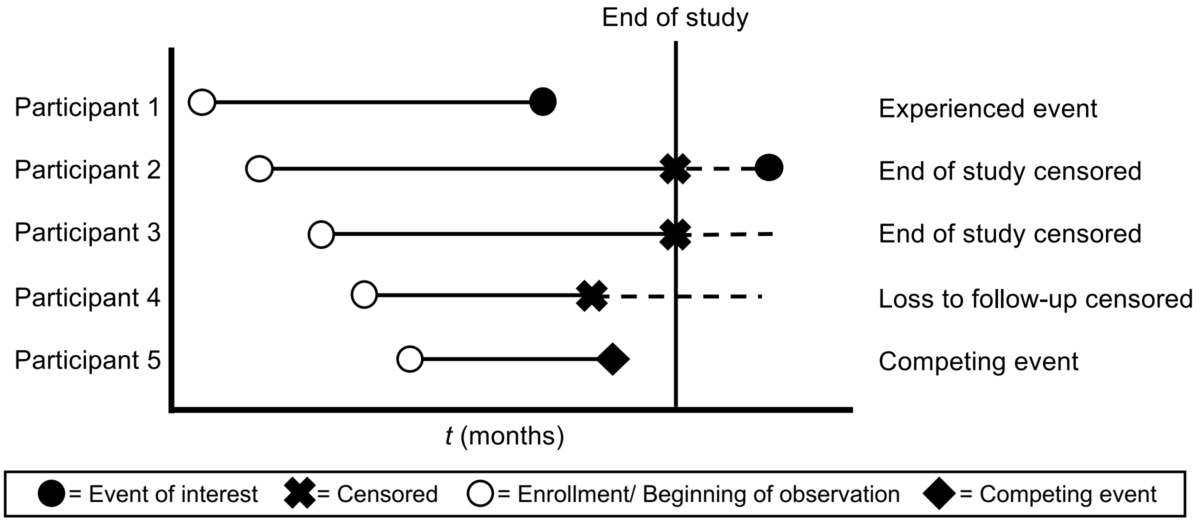


Figure 2: Different mechanisms of censoring. Individuals enter trial follow-up at different time points due to staggered trial recruiting. "End-of-study censored" individuals might or might not experience the outcome event of interest beyond their observations. "Loss to follow-up censored" individuals are censored before the administratively set end of study follow-up. A competing event precludes the observation of the outcome event of interest (adapted from Goldkuhle et al. 2019 (2), p. 130).

Non-problematic censoring can almost ever be assumed in case of administrative censoring, for example, with an administratively defined end of follow-up due to trial closure. Staggered enrollment of trial participants leads to variable observation times between individuals and variable censoring times of those who did not experience the outcome event during the observation period (2). Such censoring is often regarded as *end of study censoring* (see figure 2) and will not result in bias, because there is no association between censoring and the probability of the outcome event (7, 33).

Loss to follow-up censoring, on the other hand, introduces a potential risk for biased trial results (2, 7, 33-35). Loss to follow-up censoring summarizes the manyfold reasons for censoring that occur before the administratively set ending of trial follow-up (2, 7). That is because censoring for withdrawal or exclusion is more likely related to the censored individual's risk of the outcome event (35).

Consider, for example, individuals who experienced severe side effects during a clinical trial. It can be reasonably assumed that individuals with a serious adverse event are generally more susceptible to the event death, for example, because of general frailty or age. Censoring individuals for severe side effects in an analysis of all-cause death is therefore associated with an individual's increased probability to experience the event of interest, death.

Such censoring constitutes a potential source of bias and will, for example, if occurring to a considerable degree in one trial arm, lead to overestimation of the survival in the respective arm (2). Similar problems arise if trial participants become so severely ill that they are no longer able to attend follow-up appointments, drop out from the trial and are censored (36).

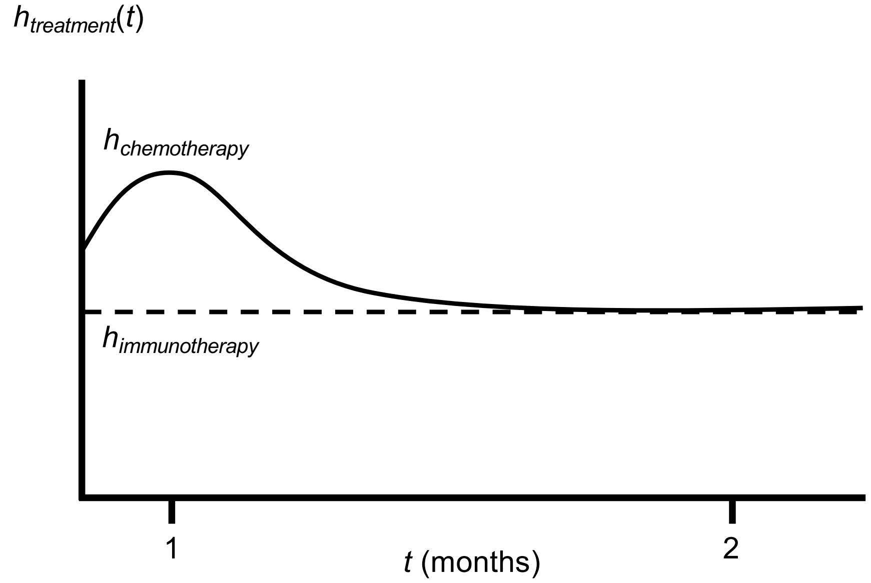
On the other hand, individuals who benefit substantially from a treatment may decide to discontinue relevant follow-up dates and withdraw. In this case, the probability of events such as death will likely be decreased for censored individuals, resulting in bias towards the opposite trial arm (2, 35).

Informative censoring has been classified as a form of selection bias that occurs when the individuals included in the analysis differ from the eligible population. In case of informative censoring, selection takes place after random allocation, e.g., because of differential loss to follow-up between trial arms (36). Unfortunately, as censoring results from absence of outcome data and the event time of affected individuals is effectively non-detectable, the actual impact of censoring on outcome analyses remains unobservable (2, 7). Simulation studies have, however, demonstrated the potential bias through informative censoring on outcome estimates and that the extend of bias depends primarily on the overall degree and early time points for censoring (2, 35, 37-39).

3.2.2. Proportional hazards

Another peculiarity of time-to-event analysis results from the properties of the HR (see chapter 3.1.3). Hazard ratios do not include a function of time. As relative effect measures, they constitute an average ratio of the hazard rates of the compared groups over the entire follow-up duration (4, 40). Consequently, they do not account for the distribution of outcome events within the assessed treatment arms (3, 20).

To draw valid conclusions, the relation of the hazard functions that comprise the HR are required to be proportional and constant over time – an assumption referred to as the proportional hazards assumption of the HR (see equation 3.4) (5, 41). The Cox model is generally referred to as the Cox proportional hazards model (5).



$$\frac{h_{immunotherapy}(t = 1)}{h_{chemotherapy}(t = 1)} < 1 \qquad \frac{h_{immunotherapy}(t = 2)}{h_{chemotherapy}(t = 2)} \approx 1$$

Figure 3: Illustrative example of non-proportional hazards. Non-proportional and not constant relationship of the hazard function of two hypothetical treatment groups, chemotherapy and immunotherapy, in comparison. Note that the curves represent the hazard (*h*) of each treatment arm over time *t* (roughly adapted from Kleinbaum and Klein 2012 (3), p. 125).

Common situations in clinical trials that leave the proportional hazards assumption in question occur, for example, when intervention effects change over time (13, 17, 21, 41).

Consider, as a hypothetical example, two oncological treatments, chemotherapy and a novel immunotherapy, that are compared in their effects on overall survival (see figure 3). Suppose that chemotherapy is associated with earlier, treatment-associated side effects and mortality, while it leads to constant survival of some individuals. The hazard function, the momentaneous probability to die at a time point for individuals alive up to that time point, of the chemotherapy

group will change over time. The hazard will be larger right after treatment due to treatment associated and potentially fatal side effects but will then decrease and eventually reach a constant rate in later follow-up because treatment leads to constant survival of a subpopulation of individuals. Immunotherapy, on the other hand, is not associated with immediate, potentially fatal adverse events, but with a linear distribution of the event death. With immunotherapy, the hazard rate remains fixed throughout follow-up.
The relation of both hazard rates is not by any means constant and changes depending on the time point of follow-up. While comparing treatments after a short follow-up period (time point 1 in figure 3) might show a large benefit with immunotherapy, a later comparison (time point 2 in figure 3) might indicate a lesser difference or even no difference at all (17).

In general, possible deviations from the proportional hazards assumption frequently occur when treatments with different therapeutic mechanisms are compared, for example, surgical versus non-surgical interventions (13, 42, 43). Yet, they might also arise simply from post-baseline selection, for instance, when high-risk individuals in one treatment arm remain free of a negative event due to the efficacy of the treatment, while in the arm with less efficacious treatment they experience the event early during follow-up (40, 41, 43). Other factors that are associated with non-proportional hazards emphasized in the literature include factors related to trial design, such as the duration of follow-up and the use of composite outcomes (13, 44, 45).

The consequence of a deviation from the proportional hazards assumption is primarily the time dependency of the HR (20, 43, 44, 46). Since the HR is most straightforwardly interpreted as an average over a follow-up period, it is strongly depending on the length of this period (20, 44, 47, 48). Although its interpretation and transferability are difficult, a HR under non-proportional hazards can still be useful for comparisons strictly considering its time-dependency (20, 47).[8]

There are several methods available to assess the proportional hazards assumption, differing in their complexity (3). The simplest assessment is by visual inspection of available survival curves. Crossing curves at early or mid-follow-up, when a considerable proportion of the population remains under risk, will inevitably show that the relation of the hazard rates between the compared groups changes over time and that the HR is not constant (HR > 1 → HR < 1).
Statistical assessments of proportional hazards, including log-log event plots and tests based on Schoenfeld residuals, are limited to situations where individual participant data is available (44, 50).

For clinical trials with outcomes disagreeing with the proportional hazards assumption, alternative analysis strategies have been suggested, including the RMST, alterations of the Cox model, for example, with time-dependent covariates, and splitting of follow-up into individual segments during which the assumption holds (3, 45-48).

[8] Because the log-rank test is conceptually testing the null hypothesis that the HR between the compared groups indicates no difference (H_0: HR = 1), it also requires proportionality of hazards and loses in power in case of deviations (9, 17, 43, 44, 47, 49).

3.2.3. Competing events

The methodological issues that are previously discussed can be considered entirely specific to time-to-event analysis. Yet, there are several problems with influence on the validity of time-to-event estimates that are not limited to, but particularly relevant for time-to-event outcomes. That is, inter alia, because the longitudinal nature of time-to-event data allows specific analytical procedures, for example, alternative models or sensitivity analyses, which are not available for other types of outcome analyses.

Such an issue are competing events, or competing risks. Competing events preclude or severely alter the probability of the occurrence of the outcome event(s) under study (see figure 2) (3, 28, 51-53). As a consequence, individuals cannot experience only the event of interest alone but might experience one of multiple different events: the event of interest or one or more potential competing events (22, 53, 54).

Examples for competing events include the death of a participant in the analysis of a non-fatal outcome event, for example, relapse of disease in oncology or myocardial infarction in cardiovascular research, or death from any other cause where cause-specific death is under study, for example, cardiovascular death if cancer-related death is studied (3, 22, 52, 55). Competing events are not always fatal, but can also involve events such as hospital discharge in surgical trials or vaccination when time to infection is studied (56).

A frequently encountered strategy to avoid issues with competing events in clinical trials are combined outcomes incorporating potential competing events (57). Prominent examples are progression-free survival and time-to-treatment failure in the oncological literature or major adverse cardiovascular events (MACE), which include all-cause death by definition (57).

Even though competing events are principally relevant for all types of outcomes, they take a significant role in time-to-event analysis because they are often dealt with using time-to-event analytic techniques (3, 52).

Regular Kaplan-Meier analysis allows to include individuals who experience competing events by treating them as censored observations (3). Previous research has identified that “naïve” censoring of competing events is relatively prominent in current clinical trials, for example, in oncology or trials in high-impact journals (52, 58-60). Recall that censoring occurs under the presumption that the event time in an individual cannot be observed, but the event might still occur at some point in time. Competing events are fundamentally different: an individual who dies from a stroke cannot die from a myocardial infarction anymore, irrespective of whether follow-up would have been extended (22, 52).[9]

Furthermore, since the probability of competing events such as death or a serious adverse events is often associated with the outcome event of interest, censoring for competing events will inevitably be informative and result in an overestimation or underestimation of relative effects, depending on the circumstances (22, 52, 54, 60).[10]

[9] Conceptually, because of the non-informative censoring assumption, the Kaplan Meier method with censoring individuals for competing events estimates the event probability without the possibility of the competing event to occur (22).

[10] Assuming the independence of the distribution of the event of interest and the distribution(s) of the competing event(s), Kaplan-Meier estimates with censored competing events are sometime interpreted as describing the probability of the event in a hypothetical world (or “population” (56)), where the competing event does not exist. That is, because through censoring and its underlying assumption of independence, individuals who are censored are theoretically still able to experience the event of interest. Such an interpretation is seldom of use in clinical trials (56, 60). The resulting estimates of competing event censored Kaplan-Meier analyses are sometimes referred to as marginal event (or survival) estimates (60).

The suggested alternative to the Kaplan-Meier method for estimating the absolute risk of the event of interest at a given time point in presence of competing events are cumulative incidence functions (52, 56, 60). They allow to produce the probability of the event of interest of individuals who have, until a defined time point, not experienced the event or a competing event and who are therefore still able to experience an event of any type (22, 52). Because they depend on both the probability of the event of interest as well as the probability of the competing event(s), cumulative incidence functions estimate the event probability while taking into account the possible occurrence of competing events (22, 53). This is contrast to Kaplan-Meier estimates, which only take into account the probability of the event of interest (53). Cumulative incidence curves allow to present event curves similar to Kaplan-Meier curves as well as median event times and time point specific event probabilities (3, 52, 53). As an alternative to the log-rank test in competing event settings, Grey's test has been suggested to test the *"equality of cumulative incidence curves between groups"* (52).

Cox models with censored competing events are referred to as cause-specific hazard models (60). When individuals with competing events are censored, hazard functions describe the momentaneous rate of the event of interest for those who have not yet experienced the event of interest *or* the respective competing event(s), for example, the momentaneous rate of disease relapse in individuals who have, until that point in time, not yet relapsed or died (22).[11]

To estimate event incidences under competing events an alternative proportional hazards regression model by Fine and Gray has been suggested (22, 52). Fine and Gray's model makes estimations on the subdistributional hazards function. While the Cox models cause-specific hazard describes the rate of events in individuals who have neither experienced the event of interest nor the competing events, the subdistributional hazard describes the rate of events in individuals who have not experienced the event of interest, while still including those with competing events amongst the individuals under risk (22).[12] The subdistributional hazard describes, for example, the momentaneous rate of the outcome event disease relapse in individuals who have previously died and those have not yet relapsed from disease.

Previous research has indicated that adequate analytical techniques for the analysis of competing events are underused in current clinical trials and that trial outcomes with potential competing events are often not adequately interpreted (52, 56, 58-60).

3.2.4. Treatment switching

Treatment switching describes the situation when individuals allocated to one trial arm receive the intervention that is by randomization designated for another treatment arm. For instance, participants receive or have access to an experimental treatment although they had been assigned to the control arm or vice versa (62-67). Like competing events, it is common in clinical

[11] A more technical explanation is provided by Koller et al. 2012 (60), who explain that, under competing events, the Cox model constitutes a *"multi state model(s) with initial state 0 and two absorbing states 1 (event of interest) and 2 (competing event), with the transition intensity from the initial state to either of the two absorbing states determined by the cause-specific hazards (hazard)$_1$(t) and (hazard)$_2$(t). The cause-specific hazard for 1, the event of interest, can then be interpreted as the momentary force that draws a subject out of state 0 into state 1"* (60).

[12] Due to its relation to the cumulative incidence function, which is analogous to the relation of the Kaplan-Meier function and the Cox models cause specific hazard, subdistributional hazard functions depend on the hazard for the event of interest as well as the hazards of potential competing events (22, 52). A technical explanation of the Fine and Gray regression model, including its relation to the cumulative incidence function, is provided by Austin et al. 2016 (61).

trials and an issue not unique to time-to-event analysis. Yet, time-to-event analysis allows for specific adjustment and sensitivity analyses, which are possible because of the longitudinal nature of time-to-event data.

Treatment switching can be observed in oncology (here also commonly termed “cross-over”), where participants are offered the experimental anti-cancer drug upon disease progression (see figure 4) (65, 67), as contamination in screening studies, where participants in the control group might seek screening outside of the trial (68), or in surgery, where participants in the experimental group undergo the standard procedure based on the surgeons’ decision (62, 64). It is also transferable to the issues of rescue medication, dose (de-)escalation or alteration, cessation from treatment and reception of third treatments in trial arms (69).
Such occurrences impose a challenge for the interpretation of trial comparisons and are sometimes interpreted as a limitation in RCT design or execution and as a risk of bias (70-72).

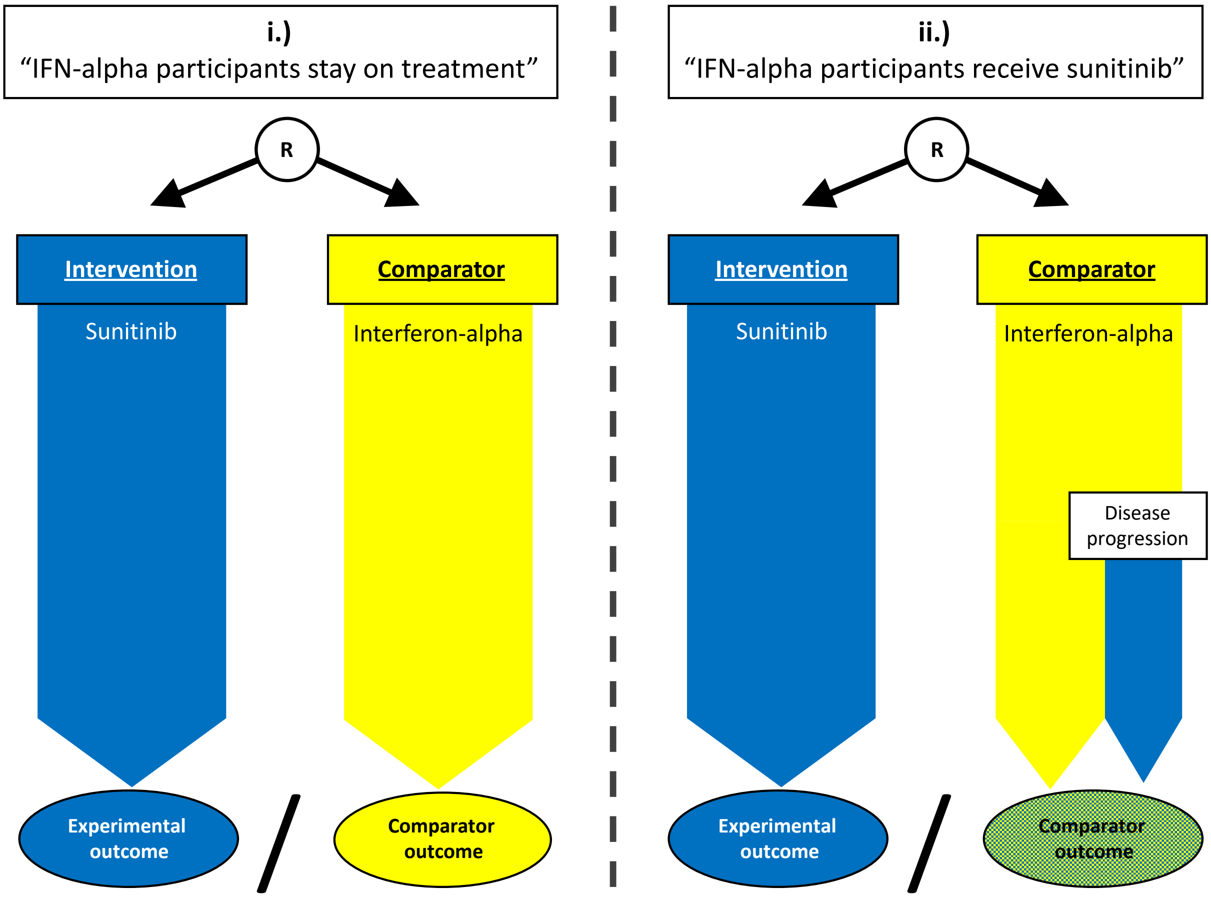


Figure 4: Example for issues with treatment switching in oncology. When interpreting a hypothetical trial comparing sunitinib (blue) to interferon-alpha (yellow) for the treatment of renal cell carcinoma two different outcome effects might be of interest: a comparison effect had all individuals stayed on their initially allocated treatment (i.) or a comparison effect where individuals who relapse under interferon-alpha have the option to receive sunitinib (ii.). A trial performed in accordance with situation ii. will not provide direct evidence for situation i., but may still provide useful evidence under certain considerations (adapted from Goldkuhle et al. 2023 (63), p. 43).

How treatment switching in a trial affects its results depends primarily on the trialists, who, corresponding (or eventually not corresponding) with their question of interest, select a principle of analysis: intention-to-treat or per protocol (63, 71, 73).[13] This question of interest has also been described as the type of causal treatment effect of interest to those conducting and analyzing a trial (70).

Trialists interested in the "effect of assignment to the treatments in comparison", irrespective of what occurs past the assignment, will lean to the intention-to-treat (or as randomized) principle (71, 73). Trial participants are analyzed in their randomly allocated trial arms, irrespective of their degree of adherence or which treatment they finally received throughout the trial (75, 76). Intention-to-treat is particularly relevant for those addressing the effects of treatments on the population level and under routine conditions (75).

In a situation where individuals randomized to the control arm receive the experimental trial intervention at some point of follow-up, for example, because of disease progression in oncological trials, such individuals will be analyzed in the control arm (see situation ii. in figure 4). The proceeding is robust, because the randomization is kept, yet the actual effect of the interventions will remain unknown. If the experimental intervention is beneficial or harmful in both groups alike, a potentially visible group difference tends to diminish towards the null (70). An opposing effect of the intervention in both groups, on the other hand, will result in an under- or overestimation of the treatment effect (70).

The interpretation of trial estimates from intention-to-treat analyses depends strongly on the course of treatment of the individuals after randomization, including alternative interventions and the degree of adherence. The applicability of such estimates to clinical decision making depends on the similarity of the trial treatment courses and the target situation (63, 75).

Trial authors interested in the effect of an intervention when all individuals received and adhered to treatment in accordance with the foreseen treatment protocol, also referred to as the "effect of treatment", lean to the per protocol principle (70, 71, 73, 75, 76). More illustratively, for an individual, per protocol analyses address the question "*What effect of treatment can I, as a patient, expect, if I take treatment perfectly according to the treatment plan?*".

Naïve per protocol analyses include only trial participants that received treatment in accordance with their initially allocated group and exclude individuals who deviate from the protocol to a certain degree, for example, who received the comparator intervention (such as in situation i. in figure 4). (75). The exclusion of participants from the analysis, however, reduces statistical power and requires that the protocol violations leading to exclusion occur completely at random, which is not plausible in most situations and will consequently result in biased estimates (74, 75).

With time-to-event data, it is possible to censor observations of individuals who received a treatment initially allocated to a comparator trial arm. Switching of participants to a different treatment, for example, because of disease progression, will almost inevitably be associated with their outcome probability and may lead to issues of informative censoring (74).

To address the constraints of per protocol analyses, multiple procedures are available to adjust for prognostic factors that differ between those who complete the study in accordance with the treatment protocol and those who don't (65, 67, 73-75, 77-83). Such procedures can be applied for all sources of censoring (discussed in chapter 7.4.2.2). In the context of treatment

[13] For an introduction to issue of treatment switching in clinical trials, particularly oncological trials, I suggest the article by Köhler et al. 2018 (74). For details on the interpretation of treatment switching in clinical trials under different conditions, I recommend the article by Mansourina et al. 2017 (70) and the elaboration of the associated questions or causal effects of interest accompanying the Risk of Bias 2.0 tool by Sterne et al. 2021 (71).

switching, their application is particularly common, for example, in the oncological literature, to model a relative effect that would have been observed, if all individuals had remained on their allocated treatment (77-82, 84).[14] The procedures require the availability of individual participant time-to-event data and underly strong assumptions (74, 77).

3.3. Evidence syntheses, systematic reviews and meta-analyses

3.3.1. Evidence syntheses and systematic reviews

Particularly in health sciences, with an ever-increasing body of literature, decision makers, health care providers and patients alike depend on syntheses of the available research to derive adequate evidence-based decisions (85, 86). Elliott et al. 2021 (87) define evidence synthesis as the "*process of identifying and combining data across studies to create a clear understanding of a body of research*".

Evidence syntheses take a wide range of functions in health care decision making, e.g., as ground for individual health care questions or to provide an evidence base for clinical guidelines, health economic evaluations and decision analyses (87, 88). The most common source of evidence in this context are research data originating from primary studies (88, 89).

Amongst evidence syntheses, the term "systematic review" is sometimes used to characterize an individual research design, which according to authors or the Cochrane Handbook, "*[…] collate evidence that fits pre-specified eligibility criteria in order to answer a specific research question. They aim to minimize bias by using explicit, systematic methods documented in advance with a protocol*" (90).[15] To date, a diverse landscape of systematic reviews is available, amongst which reviews addressing intervention effects are most common. Adapted methods exist, for example, for reviews on questions of exposure, prevalence, prognosis or diagnostic test accuracy (85, 86, 90). While standard systematic review methods synthesize RCT results, multiple adaptations also exist for data from a range of primary study designs, including non-randomized or observational studies, diagnostic test accuracy studies, prognostic model and factor studies as well as preclinical studies (85, 90, 91). The methods and principles underlying systematic reviews are generally applicable to all forms of evidence syntheses which rely on study data (88).

This dissertation addresses time-to-event analysis specific issues that predominately arise in systematic reviews of RCTs. Yet, many of the issues touched are transferable to other review types which involve time-to-event outcomes, especially prognostic reviews and reviews based on observational studies. Furthermore, many of the issues assessed in the following work are applicable to other types of evidence syntheses which originate from systematic review methods, such as clinical guidelines, health economic evaluations and decision analytic models.

[14] For a detailed description of methods such as inverse probability censoring weighting, rank preserving structural failure time and iterative parameter estimation, which are not limited to adjustment for treatment switching but are available for various situations with informative censoring, please see the articles by Ishak et al. 2014 (77), Henshall et al. 2016 (67) and Köhler et al. 2018 (74). Otherwise, I refer to chapter 7.4.2.2 of this dissertation.

[15] The Cochrane Handbook is the most renowned resource of evidence synthesis methodology. Cochrane, formerly the Cochrane Collaboration, as an organization is considered, as of today, a leader in methodological advancement of systematic reviews. Systematic reviews produced by Cochrane have been acknowledged as gold standard (88). For detailed information on systematic reviews and their methods, I suggest the Cochrane Handbook, which can be accessed through: www.training.cochrane.org/handbook (89).

Systematic reviews of interventions address a clearly defined clinical question formulated according to the established People/Participants-Interventions-Comparators-Outcomes (PICO) scheme (92):

- People/ Participants: Who are the individuals that are addressed, for example, adults suffering from early-stage Hodgkin lymphoma (93)?
- Interventions: What is the experimental treatment that is compared, including its way of application and other treatment related factors, for example, radiotherapy in combination with chemotherapy (93)?
- Comparators: What is the control treatment that is compared, including its way of application and other treatment related factors, for example, chemotherapy alone (93)?
- Outcomes: Which outcomes are of interest, including their definition, for example, overall survival or all-cause death, progression-free survival and adverse events (93)?

Central features of all systematic reviews are a systematic search for relevant literature from medical databases or other data sources, clearly defined in- and-exclusion criteria, a systematic selection process and a summary of the identified information (86, 88). In addition, the process involves a critical evaluation of the identified data, for instance, in form of Risk of Bias and Grading of Recommendations Assessment, Development and Evaluation (GRADE) assessments (as discussed in chapters 3.3.3 and 3.3.4).

Quantitative core of evidence syntheses are meta-analyses, which, under given circumstances, combine the outcome effect estimates of individual trials to a common estimate (see the next chapter 3.3.2).

If performed rigorously and under consideration of possible restraints, systematic reviews with meta-analyses of RCTs have frequently been considered the most reliable type of evidence for treatment comparisons (94, 95).

A central source of systematic reviews is Cochrane, or the Cochrane Collaboration, which publishes highly standardized and methodologically elaborate systematic reviews. Besides renowned systematic reviews, Cochrane produces widely utilized guidance, such as the Cochrane Handbook, review tools, such as the Risk of Bias tool, and continuously develops all steps of systematic review production in various working groups of international methodologists (71, 72, 89).

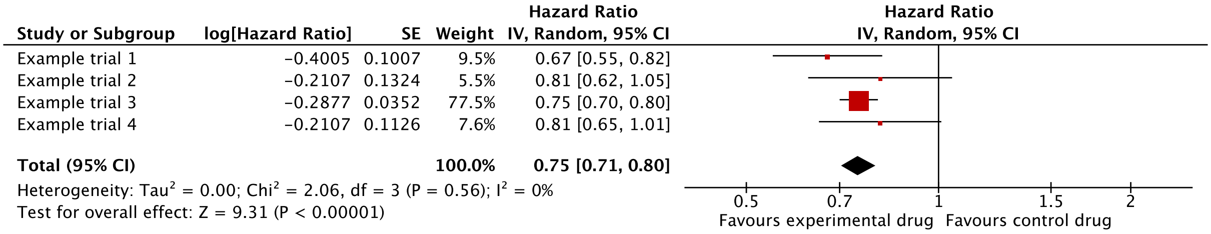
Precondition for the reliability and usefulness of systematic reviews and meta-analyses is their transparent reporting. An established and generally accepted reporting standard exists in form of the Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) guideline, which comprises 27 individual items over the entire review process (85).

Multiple adaptions of the PRISMA guideline for diverse types of systematic reviews have been developed, for example, for reviews based on diagnostic test accuracy studies, individual participant data meta-analyses and scoping reviews (96-98). A guideline for systematic reviews with meta-analyses of time-to-event outcomes is currently lacking (99).

3.3.2. Meta-analysis

Meta-analyses pool data from two or more separately analyzed populations in order to draw more confident and elaborate conclusions (85). Quantitative synthesis of trial results increases the sample size and results in higher statistical power and precision of estimates and tests (88, 90). Furthermore, pooling data from variable populations can increase the generalizability of findings and allows to assess the consistency of intervention effects across populations, or occasionally the lack thereof (88). These advances require the rigorous performance the underlying systematic reviews, the meta-analyses as well as the quality and reliability of any underlying data (90).

Meta-analysis can be performed based on individual participant data or with already aggregated trial result estimates. Even though individual participant data meta-analyses are considered as the gold standard, conducting them requires access to primary study data, greater statistical expertise and, usually, more resources (100, 101).[16] This is why most meta-analyses published today rely on aggregate data (86). Compared to meta-analyses of individual participant data, aggregate data meta-analyses tend to favor the experimental intervention, but perform similar particularly with larger population sizes and/ or higher rates of outcome events (101).



Study or Subgroup	log[Hazard Ratio]	SE	Weight	Hazard Ratio IV, Random, 95% CI
Example trial 1	−0.4005	0.1007	9.5%	0.67 [0.55, 0.82]
Example trial 2	−0.2107	0.1324	5.5%	0.81 [0.62, 1.05]
Example trial 3	−0.2877	0.0352	77.5%	0.75 [0.70, 0.80]
Example trial 4	−0.2107	0.1126	7.6%	0.81 [0.65, 1.01]
Total (95% CI)			**100.0%**	**0.75 [0.71, 0.80]**

Heterogeneity: Tau^2 = 0.00; Chi^2 = 2.06, df = 3 (P = 0.56); I^2 = 0%
Test for overall effect: Z = 9.31 (P < 0.00001)

Figure 5: Exemplary forest-plot. Forest-plot of a meta-analysis addressing the outcome “overall survival” in a meta-analysis. The experimental treatment is clearly favored by the pooled HR (<1), as indicated by the diamond at the bottom of the plot. A HR <1 indicates a lower rate of deaths during follow-up in the experimental group (trial data hypothetical; created with RevMan 5 (102)).

Aggregate data meta-analyses combine effect estimates from multiple trials to an average weighted by the variability of the individual estimates. Trials with a lower variability, in form of a lower variance, for instance, with larger populations and/or numbers of events, are assigned higher weights (see the standard error of example trial 3 compared to the remaining trials in figure 5).

Which variability is considered in the analysis depends on the meta-analysis model that is used. Random effects models consider the variability within individuals of the population under study (the within study (error) variance) and between the assessed trials (the between-study variance). Fixed effects models only consider the variability within individuals of the studied population (103, 104). In consequence, random effects models acknowledge that results can differ between the considered trials and can be randomly distributed amongst them. Meta-analysis estimates the mean of the distribution of the effect sizes (105). Fixed effects models (or common effects models), on the other hand, estimate a common effect across studies and assume that the available data represent this common, true effect.

Both give similar pooled results if there is no variability (or heterogeneity) between the trials in the meta-analysis. However, random effects models generally tend to produce wider CIs and

[16] Meta-analyses of individual participant data assemble primary data from multiple trials and allow to perform primary analytic procedures to the resulting data set, while respecting cluster effects of the data. Particularly for time-to-event analyses such data is useful as it allows, besides the performance of one’s own survival models, a wide range of additional analytic options, e.g., competing event analyses and adjustment for censoring. For more information on individual participant data meta-analyses, I recommend the PRISMA-IPD Statement (97).

assign relatively higher weights to smaller studies (those with greater variability) than fixed effects models (103, 104).[17]

The choice of the meta-analytic procedure depends on the type of effect estimate under study. Time-to-event data meta-analyses are usually performed by pooling individual trial HRs to a single HR weighted across trials. The most prominent method to pool HRs is the inverse variance method, which is also implemented in the popular systematic review software package RevMan (see figure 1) (102, 104). The method requires for each eligible trial a HR or its logarithm (log(HR)) as well as the variance or standard error of either the HR, which can also be derived from reported CIs, or its logarithm (106). Inverse variance methods are available for random and fixed effects models (104, 107, 108).
Alternative methods for pooling of time-to-event data include the fixed effects model suggested by Peto or the random effects meta-analysis by the Hartung-Knapp-Sidik-Jonkman method (104, 109, 110).[18] To account for the between-study variance, random effects meta-analyses require an estimate of this variance, for which several methods are available (103, 109, 111).

Necessary for the adequate interpretation of results from aggregate data meta-analyses is an assessment of the degree and potential sources of the between-study variance. The variability of results between trials in this context is also often referred to as statistical heterogeneity, resulting from the differences in the populations, interventions and/ or outcomes (clinical heterogeneity) and the methodological variation between trials (methodological heterogeneity) (104).
As available statistical test, the Chi^2 test allows to assess whether the visible variability between trials of a sample is compatible with chance alone (104). The I^2 value, which can be derived from the Chi^2 test statistic, allows to quantify the degree of between-study variability that is attributable to statistical heterogeneity rather than chance (104). Both procedures are frequently used, but underly certain limitations, are only feasible in defined situations and require adequate interpretation.[19]

In presence of heterogeneity between studies, but also in case of further interest in potential subgroup effects, subgroup analyses with aggregate data meta-analyses are possible by stratifying an identified trial sample by participant characteristics, such as age or severity of disease, or intervention related characteristics, such as application or dosing differences. Individual meta-analysis of the resulting trial strata then allows to explore potential effect modification between subgroups (104, 112, 113). Likewise, sensitivity analyses are possible by stratifying the included studies according to trial-related characteristics, such as study quality or publication dates, to assess potential methodological heterogeneity. This requires a sufficient number of included trials and differences amongst them which allow stratification (104).
Statistical tests for subgroup differences are available (104). If the set of included trials is large, meta-regression allows to assess the influence of certain trial characteristics on the pooled HR (104).

[17] In their article, Borenstein et al. 2010 (103) give a more technical introduction to random and fixed effects models as well as their influence on inverse variance meta-analyses.
[18] For a detailed elaboration of the methods of meta-analysis, I suggest "*Chapter 10: Analysing data and undertaking meta-analyses*" of the current version of the Cochrane Handbook (89).
[19] For a detailed explanation of the role assessment of heterogeneity between studies in meta-analyses I recommend the article by Higgins et al. 2002 (112).

3.3.3. Risk of Bias

As previously highlighted, a central feature of every evidence synthesis is the critical assessment of the evidence under consideration. The validity of evidence syntheses is particularly susceptible to misleading trial results resulting from bias, which refers to systematically distorted results (70, 114, 115). Bias in RCTs might arise due to limitations in trial planning and conduct and/ or from failure to establish safeguards against bias, and leads to over- or underestimation of effects (114, 115). Previous research has shown, for example, that trials with inadequate generation of the randomization sequence, failure in allocation of randomization concealment or lack of blinding tend to overestimate treatment effects as compared to trials with implemented safeguards against bias (116, 117).

Bias in trial results must be distinguished from statistical or random error, which decreases with higher statistical precision, for instance, due larger sample sizes or higher numbers of events, and is reflected in the widths of an estimates associated CI (115).

Multiple tools have been developed to formalize the risk of bias assessment of evidence from RCTs, amongst which the tools provided by Cochrane are most widely used (71, 72, 86, 114, 118). Cochrane provides both a traditional tool (Risk of bias) and an updated version (Risk of Bias 2.0). Irrespective of slight differences in terminology and conceptualization, both tools focus on five domains that might introduce bias in trial outcomes (71, 72):[20]

- *Selection bias* (or "*Bias arising from the randomization process*") results from a systematic difference in the characteristics and the distribution of prognostic factors in individuals of the compared groups (36, 70, 89, 115).
- *Performance bias* ("*Bias due to deviations from the intended interventions*") results from a systematic difference in the access or exposure to care or other associated factors between groups (89, 115).
- *Attrition bias* ("*Bias due to missing outcome data*") results from a systematic difference in individuals who are lost-to-follow-up, withdraw or are otherwise excluded from the trial between groups (89, 115).
- *Detection bias* ("*Bias in measurement of the outcome*") results from a systematic difference of measurement and detection of the outcomes between the compared groups (89, 115).
- *Selective outcome reporting bias* ("*Bias in selection of reported results*") results from a systematic difference in outcomes of a trial that were a-priori specified and/ or assessed and those that were published (89, 115).
- Other factors that could introduce bias, for example, stopping trials early for benefit or conflicts of interest (89, 114, 115).

Based on these potential sources of bias, conductors of evidence syntheses derive structured judgements of a low, high or unclear risk of bias for an associated domain of the risk of bias tools and subsequently a study outcome (71, 89).

Various potential sources of bias due to time-to-event specific study limitations exist (36, 39, 56, 58, 62, 70, 77, 119, 120). As described previously, informative censoring constitutes a possible source of selection bias resulting, for example, from loss to follow up or censoring for

[20] Mansournia et al. 2017 (70) provide a conceptual elaboration of the respective bias domains using causal inference methods.

competing events or treatment switching (discussed in chapter 3.2.1. and 3.4.3) (2, 36, 63, 71).

3.3.4. Certainty of evidence (GRADE)

The GRADE approach is widely used in evidence syntheses, systematic reviews and international clinical guidelines to communicate the results of meta-analyses and to determine the certainty of an overall body of evidence for a defined PICO question (121, 122). It is mandatory for reviews published by Cochrane and implemented in guidelines, for example, by the World Health Organization, the European Commission and the German Arbeitsgemeinschaft der Wissenschaftlichen Medizinischen Fachgesellschaften (123).

For systematic reviews, the GRADE approach provides a formal rating of the certainty that a pooled effect estimate is correct and/ or located relative to defined effect thresholds (122, 124, 125). For clinical guideline questions, it provides a rating that the available evidence is appropriate to determine the strengths of a defined guideline recommendation (122, 124, 125).
To rate the certainty in a body of evidence consisting of RCTs, the approach distinguishes five domains that reduce one' certainty:[21]

- *Study limitations* (*risk of bias,* as discussed in chapter 3.3.3) refer to limitations in the design or conduct in the underlying trials resulting in a systematic deviation of trial results from the truth (114, 126).
- *Indirectness* refers to a deviation of the population, interventions or outcomes that are studied in the trials of a body of evidence from the target situation to which the evidence is to be applied (63, 127).
- *Imprecision* refers to the degree of random or statistical error associated with an effect estimate. In GRADE, judgements of imprecision have standardly been based on the 95% CIs. The CIs boundaries are compared to defined effect thresholds, such as the null effect, if one is interested in the question "is there any effect at all", or the so called minimally important difference, which is defined by patient preference and addresses the question "is the visible effect important". If the CI crosses the defined threshold(s), the estimate is considered imprecise regarding a judgement of the location of the effect relative to the defined threshold(s) (128-130).
- *Inconsistency* refers to the consistency of the magnitude of relative intervention effects across the included trials in a body of evidence and is analogous to the concept of between-study heterogeneity. If the differences in effects between trials cannot be explained, for example, by differences in the population or interventions (assessed through subgroup analyses) or factors such as trial characteristics, quality or setting (assessed through sensitivity analyses), this results in reduced certainty or might require the exclusion certain trials from the body of evidence (113).
- *Publication bias* is based on the observation that trials with positive results are published more often, earlier and more accessible (131, 132). When negative studies are missed, a pooled effect estimate for a body of evidence will be skewed, often favoring the intervention (131, 132). The available GRADE guidance suggests critical assessment of certain trial characteristics, i.e., caution in case of large numbers of smaller

[21] Detailed guidance for the application of the GRADE approach as well as its underlying concepts is available in form of the GRADE guidelines series which is published in the Journal of Clinical Epidemiology: www.jclinepi.com/content/jce-GRADE-Series.

trials and commercial funding, as well as the use of statistical and visual methods, such as the funnel plot (132).[22]

The approach also recognizes factors that may increase the certainty in a body of evidence, given that it has not been reduced previously. These factors include particularly large effects (e.g., a relative risk >2 or <0.5), evidence for a dose response relationship and residual confounding that plausibly leads to an underestimation of the visible effects (134).

Following a structured approach, GRADE users derive judgements in and across the above-described domains and so rate their certainty in an individual outcome of a body of evidence. These judgements range from high certainty (further research is unlikely to alter the confidence in the respective estimate) to very low certainty (any estimate of effect is very uncertain) (124).

To ensure the comprehensible and transparent communication of the evidence in relation to its associated certainty, systematic reviews and clinical guidelines include GRADE Summary of Findings tables and related formats (e.g., Evidence Profiles) (135-137). Key elements of these formats include for each individual outcome:

- The certainty of evidence rating derived via the GRADE approach
- The pooled relative effect that was derived through meta-analysis
- A measure of absolute risk, which is occasionally calculated from the pooled relative effect and a measure of the population baseline or control risk[23]

Additional information that is presented includes the precisely formulated PICO question, definitions of each presented outcome, the underlying sample sizes and numbers of trials and footnotes that elaborate the presented values and judgements (135-137).
Multiple studies have indicated the benefits of Summary of Findings tables and related formats to communicate the results of evidence syntheses (139-141). They are mandatory for reviews published by Cochrane.

The GRADE approach and its associated products, including Summary of Findings tables and Evidence Profiles, are developed and constantly adapted within the GRADE Working Group. The working group is formed by international experts, including, but not limited to, renowned scientists from the field of clinical-epidemiology, medicine, epidemiology and biostatistics.
All GRADE related publications, which include GRADE guidance and concept articles, must be accepted through a highly formalized process, that includes small- and large-group discussions at official GRADE Working Group meetings and votes during which the content of a novel

[22] A funnel plot relates the magnitude of the individual effect estimates of trials in a body of evidence to their precision. Asymmetries in the distribution of trials on the plot (which should have the form of a reverse funnel), particularly in absence of smaller, less precise trials with negative effects, can suggest a potential risk of publication bias in the evidence at hand. That is because in absence of publication bias, with increasing precision, trials should distribute closer around the pooled effect estimate. Statistical tests for asymmetry exist. Yet, interpretation of the funnel plot remains subjective and is, according to the Cochrane Handbook, not suggested if less than ten trials are included. Manyfold alternative statistical approaches are available and are, for example, described in the Cochrane Handbook (138) and the corresponding GRADE guideline (132, 133).

[23] The presentation of absolute effects estimates, such as natural frequencies or risk differences, in addition to a pooled relative effect ensures adequate interpretation of the magnitude of effects in relation to a population and/ or control risk, allows risk stratification according to different baseline risk groups and prevents common problems that are associated with relative effect estimates, in particular their overinterpretation (138).

article must be accepted by at least 80% of the attending GRADE Working Group members (142, 143).

3.4. Challenges of time-to-event analyses in evidence syntheses

3.4.1. Recalculation of data from trial publications

As previously described, meta-analyses of time-to-event outcomes are commonly performed by pooling HRs from multiple individual trials to derive a common HR. Established software packages for meta-analysis require for each trial a HR or log(HR) and its variance or CI (102). Unfortunately, such data is not always presented in trial publications with time-to-event analyses (23-25, 144).

In difference to binary relative effect measures, i.e., risk ratios or odds ratios, which can be straightforwardly calculated based on the number of events and the total number of analyzed individuals in each trial arm, time-to-event summary measures require more elaborate recalculation methods. These methods are based on the relationships between the HR/ log(HR), the hazard function, the survival function, in form of Kaplan-Meier curves and estimates, and the log-rank test, with its O–E events, log-rank variance and p-value, which are discussed in chapter 3.1 (106, 145-156).

A renowned resource for the recalculation of time-to-event summary data from trial publications is provided by Tierney et al. 2007 (106).[24] The authors distinguish approaches that are based on reported information from approaches based on Kaplan-Meier curves. The different methods differ in their required data, their complexity as well as in their validity (106, 153, 154, 157). Approaches based on reported information make use of available summary statistics from trial time-to-event analyses with or without additional information. Direct methods use, as previously reported, summary statistics such as the HR, the log HR, the O–E events and their variances. They are associated with less uncertainty than indirect methods which rely, for example, on the log-rank p-value, the number of events and/ or the number of analyzed individuals in each trial group (101, 151, 157, 158). Factors that affect the validity of indirect recalculation methods include the trial size, the relative treatment effect as well as the pattern of censoring (157, 158).

Methods for recalculation of time-to-event data based on Kaplan-Meier plots utilize the presented event probabilities to either derive a summary HR or to approximate individual participant data, which can then be included in one's own time-to-event analyses.
The prominent Kaplan-Meier plot-based approach described by Parmar et al. 1998 (151) and Tierney et al. 2007 (106) allows to calculate a summary effect measure by combining individual HRs derived for separate time-intervals of a Kaplan Meier curve. Interval-specific HRs and variances are estimated from each interval's number of event-free individuals at the interval start, number of censored individuals, number of individuals at risk and number of events, which can be obtained from the information available in the plot.
If a number of individuals at risk for specific follow-up time points is reported along the Kaplan-Meier plot (see figure 1), the necessary information to be derived from the curves to calculate

[24] The article provided by Tierney et al. 2007 (106) provides a well-established collection of recalculation methods, which I recommend for a practical overview. It is published together with an Excel spreadsheet that helps to perform each method, e.g., the Kaplan-Meier plot-based approaches.

interval-specific HRs reduces to the number at risk during the interval as well as the number of events, assuming constant censoring within the interval (106).
The approach is similar to the method proposed by Williamson et al. 2002 (156), allowing a more liberate selection of intervals, not limited to intervals where a number of individuals at risk reported in the plot.

Other complex methods, such as the approach suggested by Guyot et al. 2012 (145), allow to approximate event times for individual participants throughout follow-up. This method relies on digitization of a presented curve as well as on information such as the number of individuals at risk at certain time points and the overall number of events. With this information, the approach seeks to derive the underlying time-to-event data parameters of the Kaplan-Meier equations that are graphically represented in the plot (145).
There are several additional sophisticated approaches, such as incorporating information from censoring marks on published curves or approaches that extend and improve available methods by fitting curves that correspond to the recalculated data (146-149, 152, 153, 155, 159).
Besides calculation of summary effect estimates, the secondarily derived individual participant data enables to assess the number of censored individuals in each arm over time as well as sensitivity analyses and statistical tests for the proportional hazards assumption (see chapter 3.2.2), which are not possible with aggregate data (44).
A critical prerequisite of all Kaplan-Meier plot-based approaches is, however, the availability of curves in appropriate resolution that allows digitization. In addition, they substantially rely on additional information, such as the number of individuals at risk at sufficient time points and the reporting or marking of censored individuals along the curves (106, 145).

These methods allow authors of evidence syntheses to recalculate summary data from various sources and thus to substantially extend the data available for their meta-analyses. It remains unexplored to date, which of the described approaches find uptake in current systematic reviews and which primary trial time-to-event data is actually used.

3.4.2. Reporting issues in time-to-event analyses of studies and trials

Authors of evidence syntheses, in particular those who perform aggregate data meta-analyses, critically rely on published information to obtain data for their analyses, interpret their findings and critically assess the available evidence. Previous research has indicated that the reporting of studies and trials which include time-to-event analyses occurs with certain limitations (2, 23-25, 144, 160-162). Is has been shown, for example, that general methodological information, such as the definitions for the analyzed outcomes, the intervals under observations (i.e., their start and end points), relevant information on events and follow-up as well as information that is relevant to assess the methods of statistical analysis (e.g., model building and validation) are often not or only inadequately presented.
The same applies for time-to-event outcome specific information. Study and trial publications often lack important details on censoring and time-to-event analysis specific assumptions and show deficient or conflicting result data as estimates and in Kaplan-Meier curves (2, 23-25, 144, 160-162).

Despite the currently available evidence, it is yet to be clarified which challenges authors of evidence syntheses face in clinical trial publications that they include for meta-analysis. Furthermore, it is currently unknown how review authors might deal with these limitations when they synthesize, report and interpret their results (162).

3.4.3. Issues that affect the certainty of time-to-event meta-analyses

The methodological hardships that are associated with time-to-event analysis, informative censoring, proportional hazards and issues such as competing events and treatment switching, discussed in previous chapters of this dissertation, are extensively studied in the currently available literature, but only for trial analyses.

How they translate to the evidence synthesis level, how evidence synthesis authors deal with them and how they should be ideally dealt with is, however, largely unclear. Particularly for authors of meta-analyses that are based aggregate time-to-event data, these issues might cause considerable problems. That is because, in the absence of individual participant data, they must rely on published information and will rarely be able to perform distinct analyses, for example, by imputing data under reasonable assumptions (2, 126).

Informative censoring, as previously discussed, has been reported as a potential source of "*Bias due to missing outcome data*" in the explanation of the Risk of bias 2.0 tool provided by Cochrane. For the detection of informative censoring in trial publications the tool specifies: "*Either differences in rates of censoring or differing reasons for censoring may provide evidence that censoring was informative*" (Sterne et al. 2019 (71)).[25]

Yet, if not reported explicitly by trialists in their publications, it is unclear how to identify and consider this time-to-event specific risk of bias. A particular problem arises because censoring is a regular component of the analytical procedure. Individuals that are censored for informative reasons, in particular loss to follow-up censoring, may in most situations be reported as included in the analysis. In survival plots, given that they are reported and censoring is marked on the plot, end of study and loss to follow-up censoring cannot be distinguished (2, 71). Even though in the exemplary survival curve in figure 1, the number of censored individuals is explicitly reported together with the number of individuals at risk, it is unclear for which reasons the respective individuals were censored.

In absence of sufficient guidance, it is questionable whether authors of current evidence syntheses are aware of this time-to-event specific source of bias and, if so, how they identify, judge and report a potential bias in their publications (2).

Methodological research has indicated that the assumption of proportional hazards might fail frequently in clinical trials (44). It is currently unclear, however, how such a failure in one or more trials should be dealt with in meta-analyses and how it affects the certainty in pooled estimates.

As described in chapter 3.2.2, a failure of the assumption of proportional hazards will lead to a time-specificity of the respective trial HR (41). While a HR from a single trial might still be interpretable as restricted to the respective follow-up time, interpreting a pooled HR from individual trials with variable follow-up durations is substantially complicated.

Assume, for example, a hypothetical meta-analysis of trials that assess immunotherapy versus chemotherapy in individuals who suffer from a certain cancer, a question that is also used as an example in chapter 3.2.2. Some trials in that body of evidence might have ended their follow-up earlier and, as previously described, their results indicate a clear benefit of immunotherapy with regard to mortality. Other trials might have followed-up their included individuals for a longer duration and therefore show that the benefit was less clear. The differing effect estimates between the trials will inevitably lead to between-study heterogeneity in the meta-

[25] The statement stems from the online explanation of the tool, available from: drive.google.com/file/d/19R9savfPdCHC8XLz2iiMvL_71lPJERWK/view (p. 42; published 22.08.2019; last accessed: 24.05.2024)

analysis. How to interpret this heterogeneity and how to incorporate a failure of the proportional hazards assumption into judgements about the certainty of meta-analysis results remains, however, unclear to date. Exploring failures of the proportional hazards assumption and variable follow-up between trials as a cause of heterogeneity by sensitivity analyses might constitute a plausible option. Others have argued that non-proportional hazards in one or more trials of a body of evidence will result in over- or underestimation of meta-analysis results and thus creates a risk of bias (44).

Additional challenges arise because the assessment of proportional hazards for individual trials in a meta-analysis is often hampered, if such an assessment is not clearly reported in trial publications. A rough visual inference on the assumption can be drawn from the course of the Kaplan-Meier curves. As reported in section 2.2.2, crossing curves, for example, could indicate a failure of the assumption under certain circumstances. Such an assessment is, however, not always applicable or conclusive. More complex statistical methods to assess the proportional hazards assumption in underlying trials require recalculating individual participant data under application of the afore mentioned procedures (section 2.4.1) (44, 145). This demands, however, the availability of sufficient data in trial publications.

At this point, it remains unclear whether and how review authors test the assumption of proportional hazards in their included trial time-to-event analyses and which consequences they draw for their analyses if, for instance, they perform sensitivity analyses or utilize alternative effect measures.

As highlighted previously, empirical evidence for the prevalence of competing events and treatment switching in current clinical trials and their effect on trial estimates exists (52, 58, 59, 62, 67, 70, 78). Yet, literature on problems with competing events or treatment switching, their effect and interpretation in meta-analyses based on aggregate data is currently absent or conflicting. While competing events are not mentioned in established resources for review authors today, the handling of treatment switching simply as a risk of bias has been drawn into question (63, 71).

Furthermore, as discussed in chapter 3.2.3 and 3.2.4, if standard time-to-event analytical procedures, in particular Cox models, are applied to derive effect estimates in trials that are affected by either of these issues, the resulting estimates require a distinct interpretation: hazard ratios under competing events are best interpreted as cause-specific hazards, while the interpretation of treatment switching largely depends on the effect of interest to the trialists, the intention-to-treat or the per protocol effect.

Given these complications, the lack of empirical investigation and guidance on competing events and treatment switching in meta-analyses of aggregate data could critically affect the quality of current evidence syntheses.

Finally, additional time-to-event specific problems in evidence syntheses might exist that are of particular interest to authors but have not been highlighted yet. Even in absence of methodological literature, it is plausible that systematic review authors mention such issues in their own publications, which has, however, not been assessed to date.

3.4.4. Interpretation of hazard ratios from meta-analyses

To this point, the discussed issues around time-to-event analyses in evidence syntheses have primarily focused on analytical hardships. Yet, there are additional challenges particularly associated with the quantitative outputs of meta-analyses, especially with their adequate interpretation and communication.

As previously indicated, the HR as a relative effect measure may lead to overinterpretation of visible treatment effects (138, 163, 164). Leading organizations in health care, including GRADE and Cochrane, therefore suggest absolute effect measures, for example, risk differences or a number-needed-to-treat, in addition to relative effects to communicate the results of quantitative analyses (136, 137, 165, 166).

In absence of individual participant data, absolute effect measures for time-to-event outcomes are usually calculated based on the pooled HR, which is combined with an absolute population or control group risk. Because the pooled HR is not time point specific, authors of evidence syntheses must select adequate time points for their calculation and must navigate certain pitfalls (166).

Furthermore, as outlined in chapter 3.1.3, the HR itself is not always straightforward to interpret. Additional challenges arise if it is utilized to assess outcomes such as overall survival and progression-free survival (10, 11). In difference to all-cause mortality and/or relapse of disease, overall survival and progression-free survival are not referred to as their representing event, but rather the absence of them. Because the HR is usually calculated based on the hazard functions for events in the compared groups, this might lead to confusion and endangers flawed interpretation. Issues might arise, for example, when a pooled HR is used to calculate an associated absolute effect. That is, because the baseline or control-group risk included in the calculation can either represent the risk of an event in the population or its absence (10, 11).

Despite these possible pitfalls, the interpretation of the HR and the calculation of absolute effects in evidence syntheses with aggregate data meta-analyses of time-to-event outcomes have not yet been assessed.

4. Objectives

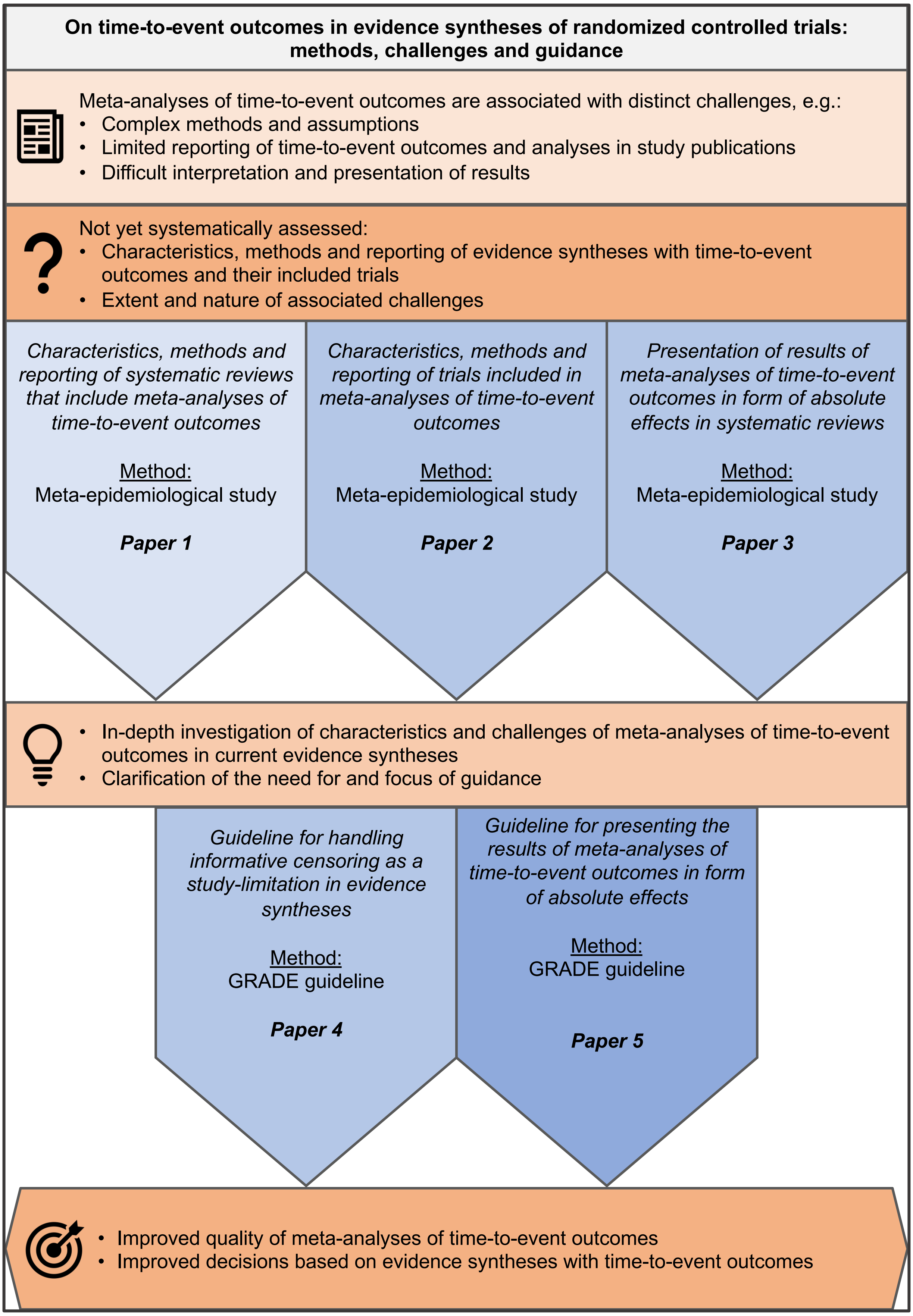
Authors of evidence syntheses who are interested in the effects of interventions on time-to-event outcomes face a great variety of difficulties when meta-analyzing data, interpreting results, formulating conclusions and communicating their findings.

These difficulties have, however, not been systematically assessed to date and concrete guidance on how to identify, assess and interpret them is lacking.

The objectives of this project were therefore:

1) To systematically explore how the authors of current evidence syntheses in form of systematic reviews perform meta-analyses of time-to-event outcomes based on aggregate data, which challenges they face and how they interpret and report their results (papers 1, 2 & 3).

2) To develop targeted guidance for the inclusion of time-to-event outcomes in evidence syntheses to support review authors in adequately reporting their proceeding and their results, and in optimally communicating their certainty in them (papers 4 & 5).



On time-to-event outcomes in evidence syntheses of randomized controlled trials: methods, challenges and guidance

Meta-analyses of time-to-event outcomes are associated with distinct challenges, e.g.:

- Complex methods and assumptions
- Limited reporting of time-to-event outcomes and analyses in study publications
- Difficult interpretation and presentation of results

Not yet systematically assessed:

- Characteristics, methods and reporting of evidence syntheses with time-to-event outcomes and their included trials
- Extent and nature of associated challenges

Characteristics, methods and reporting of systematic reviews that include meta-analyses of time-to-event outcomes

Method:
Meta-epidemiological study

Paper 1

Characteristics, methods and reporting of trials included in meta-analyses of time-to-event outcomes

Method:
Meta-epidemiological study

Paper 2

Presentation of results of meta-analyses of time-to-event outcomes in form of absolute effects in systematic reviews

Method:
Meta-epidemiological study

Paper 3

- In-depth investigation of characteristics and challenges of meta-analyses of time-to-event outcomes in current evidence syntheses
- Clarification of the need for and focus of guidance

Guideline for handling informative censoring as a study-limitation in evidence syntheses

Method:
GRADE guideline

Paper 4

Guideline for presenting the results of meta-analyses of time-to-event outcomes in form of absolute effects

Method:
GRADE guideline

Paper 5

- Improved quality of meta-analyses of time-to-event outcomes
- Improved decisions based on evidence syntheses with time-to-event outcomes

Figure 6: Graphical presentation and connection of the papers included in this dissertation.

5. Description of the interconnected papers

As shown in figure 6, the papers included in this dissertation can be divided into two groups: meta-epidemiological studies and systematic guidance articles that build on the insights generated during these studies.

To determine how authors currently perform meta-analyses of time-to-event outcomes, which specific challenges they might face, how they deal with these challenges and how they report their results and evaluations, requires a general exploration of current evidence syntheses.
Following a meta-epidemiological approach, a large sample of methodologically up-to-date systematic reviews that include meta-analyses of time-to-event outcomes based on aggregate data from RCTs was systematically identified and assessed. Besides general characteristics of the reviews and their review methodology, a particular interest was in all time-to-event specific methods, analyses and additional evaluations, for instance, sensitivity analyses as well as risk of bias and certainty of the evidence ratings. Target of this sub-project was to characterize the current proceeding with and reporting of meta-analyses of time-to-event outcomes in order to establish an evidence base for further investigation. A second target was to identify potential negative as well as best-practice examples that inform the development of guidance.

Conclusive judgements about the challenges that authors of evidence syntheses face when they conduct their meta-analyses of time-to-event outcomes require the clarification of the conditions in their included trials.
For this reason, an in-depth assessment of general and time-to-event specific methods and reporting of RCTs included in meta-analyses of a random sub-sample of the afore selected systematic reviews was performed. This assessment included trial and trial outcome specific data, for instance, general characteristics and methods, outcome definitions, time-to-event specific methods, available data as well as trial characteristics that might affect the validity of time-to-event analyses. Furthermore, relevant data on the handling and reporting of these particular trial characteristics in review publications were included, for example, which data the review authors included from trials, whether they mentioned potential shortcomings and how they addressed them. Together with the results of the first sub-project, this approach allowed a comprehensive view on the conditions in RCTs and their uptake in systematic reviews. As in the previous work step, the sub-project builds an evidence base for targeted guidance.

Finally, as previously emphasized, not only the methodological complexities of time-to-event analyses in meta-analyses, which were assessed in the previous two subprojects, require a critical assessment, but also how review authors communicate the findings of their analyses. This includes the adequate interpretation of the HR, i.e., the direction of its effect, and the correct calculation of absolute effect estimates.
Therefore, a comprehensive assessment of a large sample of oncological Cochrane reviews was performed. The sample consisted of methodologically up-to-date systematic reviews with a high prevalence of time-to-event outcomes, which, because of the mandatory presentation of GRADE Summary of Findings tables in Cochrane reviews, should present absolute effect estimates based on their pooled HRs. It was assessed whether the absolute effect estimates were calculated and reported correctly and whether the review authors provided additional details on their calculations. Besides conclusions on the frequency of correct and incorrectly calculated absolute effects for time-to-event outcomes, the results of this sub-project allowed to elicit reasons for potential problems and to build a basis for targeted guidance that supports

review authors in adequately interpreting time-to-event outcomes and calculating the associated absolute effects.

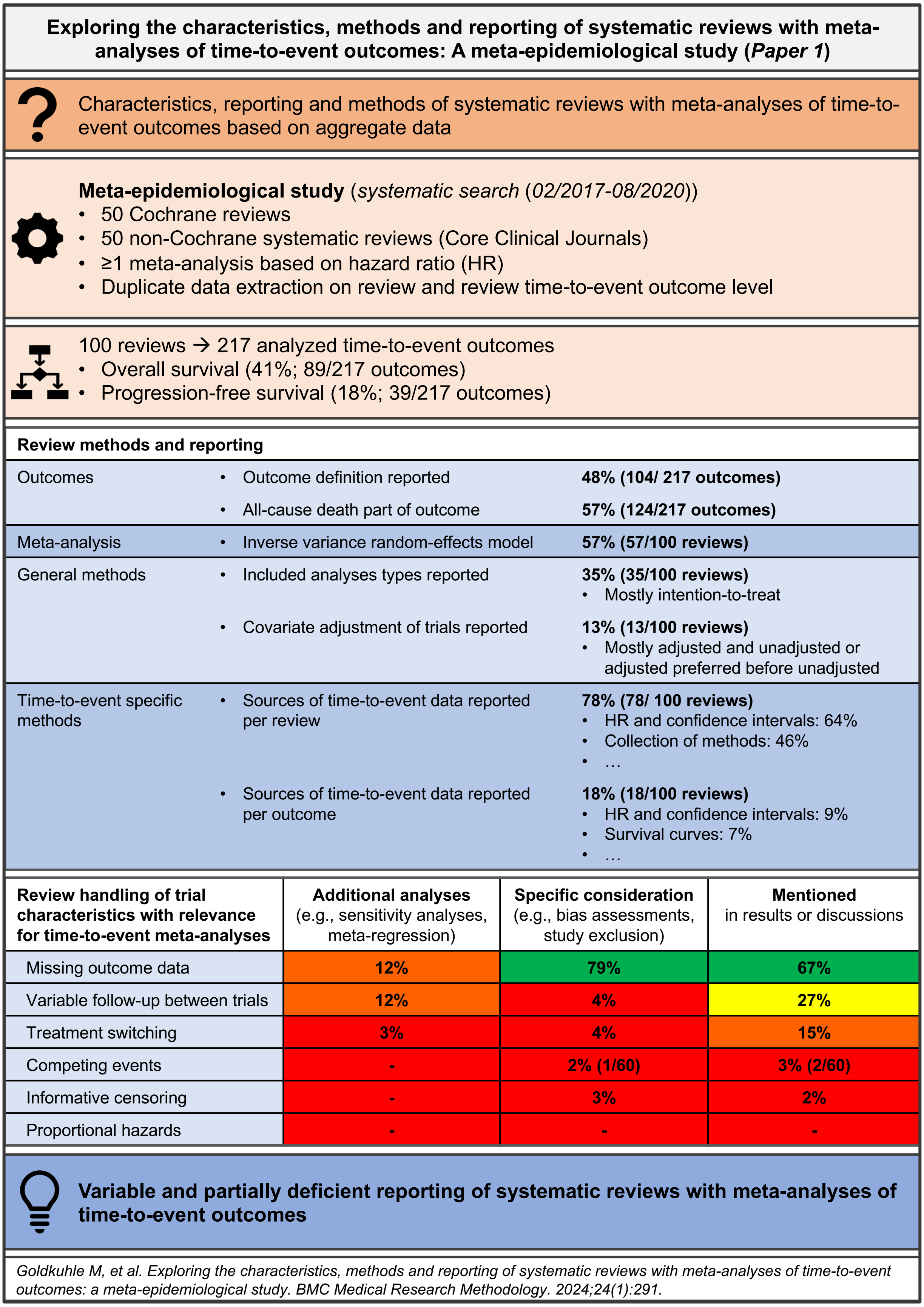
The previous three meta-epidemiological studies provide a thorough investigation of the characteristics, challenges and reporting of current meta-analyses of time-to-event outcomes in evidence syntheses. They elicit the potential as well as possible examples for additional guidance.

Two time-to-event specific issues in evidence syntheses were found most striking and were backed by practical examples and sufficient methodological literature to inform the development of guideline articles: informative censoring and the calculation absolute effects.

For these two issues, targeted guideline articles were developed within the GRADE Working Group. As highlighted in chapter 3.3.4, the GRADE Working Group is formed by a group of international experts in clinical epidemiology and GRADE guidance is developed systematically with a highly structured approach. The development of guidance articles is embedded in continuous discourse of experts within iterative small- and large group discussions, which ensures the high quality and applicability of GRADE guidance (126). The two guidance articles form the remaining sub-projects of this dissertation.

The consortium of sub-projects that constitute this dissertation should now achieve the overreaching project goal: to improve the quality of evidence syntheses with meta-analyses of time-to-event outcomes and to positively influence the decisions that are informed by them.

6. Papers included in this dissertation



Exploring the characteristics, methods and reporting of systematic reviews with meta-analyses of time-to-event outcomes: A meta-epidemiological study (*Paper 1*)

Characteristics, reporting and methods of systematic reviews with meta-analyses of time-to-event outcomes based on aggregate data

Meta-epidemiological study (*systematic search* (*02/2017-08/2020*))

- 50 Cochrane reviews
- 50 non-Cochrane systematic reviews (Core Clinical Journals)
- ≥1 meta-analysis based on hazard ratio (HR)
- Duplicate data extraction on review and review time-to-event outcome level

100 reviews → 217 analyzed time-to-event outcomes

- Overall survival (41%; 89/217 outcomes)
- Progression-free survival (18%; 39/217 outcomes)

Review methods and reporting		
Outcomes	• Outcome definition reported	**48% (104/ 217 outcomes)**
	• All-cause death part of outcome	**57% (124/217 outcomes)**
Meta-analysis	• Inverse variance random-effects model	**57% (57/100 reviews)**
General methods	• Included analyses types reported	**35% (35/100 reviews)** • Mostly intention-to-treat
	• Covariate adjustment of trials reported	**13% (13/100 reviews)** • Mostly adjusted and unadjusted or adjusted preferred before unadjusted
Time-to-event specific methods	• Sources of time-to-event data reported per review	**78% (78/ 100 reviews)** • HR and confidence intervals: 64% • Collection of methods: 46% • …
	• Sources of time-to-event data reported per outcome	**18% (18/100 reviews)** • HR and confidence intervals: 9% • Survival curves: 7% • …

Review handling of trial characteristics with relevance for time-to-event meta-analyses	**Additional analyses** (e.g., sensitivity analyses, meta-regression)	**Specific consideration** (e.g., bias assessments, study exclusion)	**Mentioned** in results or discussions
Missing outcome data	**12%**	**79%**	**67%**
Variable follow-up between trials	**12%**	**4%**	**27%**
Treatment switching	**3%**	**4%**	**15%**
Competing events	**-**	**2% (1/60)**	**3% (2/60)**
Informative censoring	**-**	**3%**	**2%**
Proportional hazards	**-**	**-**	**-**

Variable and partially deficient reporting of systematic reviews with meta-analyses of time-to-event outcomes

Goldkuhle M, et al. Exploring the characteristics, methods and reporting of systematic reviews with meta-analyses of time-to-event outcomes: a meta-epidemiological study. BMC Medical Research Methodology. 2024;24(1):291.

Figure 7: Graphical abstract for paper 1 (167).

6.1. Characteristics, methods and reporting of systematic reviews that include meta-analyses of time-to-event outcomes *(Paper 1)*

6.1.1. Publication status

This article was published in October 2024 in the journal *BMC Medical Research Methodology* (167).
The work shares of the individual participants and their involvement in this paper are detailed in appendix 11.1.1.

Results of this article were presented at the 24th Congress of the German Network for Evidence-based Medicine in Potsdam (168) and at the 27th Cochrane Colloquium in London, UK, (169) in 2023 (chapter 9.4).

6.1.2. Synopsis (167-169)

Building on the implications elaborated in the introduction, a meta-epidemiological study was conducted to systematically assess the characteristics, methods and reporting of meta-analyses of time-to-event outcomes in current systematic reviews (167-169).

This systematic assessment was performed according to an a-priori published protocol (osf.io/5qxbd). It included a sample of 100 systematic reviews with at least one pairwise meta-analysis of aggregate data from RCTs that addressed a time-to-event outcome based on the HR. Methodologically specialized review types, such as network meta-analyses, were excluded. Fifty eligible Cochrane reviews were identified from the Cochrane Database of Systematic Reviews and systematically selected in a backwards fashion starting with the most recent publication (08/2020). This ensured a methodologically advanced and up to date sample of systematic reviews which could also provide possible best-practice examples. A corresponding sample of systematic reviews published in Core Clinical Journals, a MEDLINE filter limiting the search to the most relevant journals to physicians according to the U.S. National Library of Medicine (170), was identified through a systematic search on Medline (08/02/2021). From the collection of reviews published in Core Clinical Journals, 50 additional reviews were randomly selected, stratified by publication years, so that the final sample for the assessment comprised 100 systematic reviews.

For the included reviews and each of their individual time-to-event outcomes, data on review characteristics (e.g., publication dates, populations, interventions, comparisons, outcomes and trials), the analyzed time-to-event outcomes (e.g., definitions, composition, and presentation), general review methodology (e.g., analysis types, adjustment for covariates, meta-analysis methods), time-to-event-specific methods (data basis and recalculation of outcome data) and handling of specific trial characteristics relevant to time-to-event analysis (e.g., variable follow-up, competing events, informative censoring, proportional hazards and their interpretation) were extracted. Data on the results of the analyses (e.g., relative effects and their interpretation) were also extracted as well as their reporting and the presentation of the results as absolute effects. In agreement with methodological standards, the selection of reviews and the extraction of data were performed in duplicate by two researchers (171). The statistical analysis was performed descriptively.

The 100 eligible systematic reviews included a total of 217 individual time-to-event outcomes for which meta-analyses were performed. Most of the reviews assessed oncological research

questions, followed by questions on cardiovascular diseases, whilst the most frequently encountered review outcomes were overall survival and progression-free survival.
The results of the systematic assessment demonstrate that the reporting of systematic reviews with meta-analyses on time-to-event outcomes occurs inconsistent and occasionally inadequate. Outcome definitions, for example, were available for less than half of the total 217 assessed review outcomes. General review methodology, such as the types of trial analyses included in reviews (e.g., by intention-to-treat, per protocol) or the covariate adjustment status of included trial estimates, were reported in only few of the 100 reviews. Time-to-event data were obtained from trial publications with great variability, both in the number of different data sources and in the methodological depth of the respective recalculation procedures, i.e., ranging from the inclusion of reported HRs to the recalculation of individual patient data from Kaplan-Meier plots.
Particularly deficient were the discussion and consideration of trial characteristics that affect the validity of time-to-event analyses, such as variable follow-up between studies, informative censoring, competing events, treatment switching and proportional hazards. With few exceptions, the respective characteristics were only seldomly included in additional assessments, such as sensitivity analyses or risk of bias assessments, or discussions by the review authors.

This work is the first to systematically assess the reporting of time-to-event analyses at the review level. A central implication of its findings is the need to increase the reporting quality of the corresponding evidence syntheses (see chapter 7.4.11). Furthermore, it raises the importance of an assessment of the conditions that review authors face with regard to time-to-event analyses in the trial publications they include in their syntheses (see paper 2).

6.1.3. Full manuscript (167)

The supplementary material accompanying this manuscript is provided in appendix 11.1.2.

Goldkuhle *et al. BMC Medical Research Methodology* (2024) 24:291
https://doi.org/10.1186/s12874-024-02401-4

BMC Medical Research
Methodology

RESEARCH

Open Access

Exploring the characteristics, methods and reporting of systematic reviews with meta-analyses of time-to-event outcomes: a meta-epidemiological study

Marius Goldkuhle[1*], Caroline Hirsch[1], Claire Iannizzi[1], Ana-Mihaela Zorger[1], Ralf Bender[2], Elvira C. van Dalen[3], Lars G. Hemkens[4,5,6,7], Ina Monsef[1], Nina Kreuzberger[1†] and Nicole Skoetz[1†]

Abstract

Background Time-to-event analysis is associated with methodological complexities. Previous research identified flaws in the reporting of time-to-event analyses in randomized trial publications. These hardships impose challenges for meta-analyses of time-to-event outcomes based on aggregate data. We examined the characteristics, reporting and methods of systematic reviews including such analyses.

Methods Through a systematic search (02/2017-08/2020), we identified 50 Cochrane Reviews with ≥ 1 meta-analysis based on the hazard ratio (HR) and a corresponding random sample ($n = 50$) from core clinical journals (Medline; 08/02/2021). Data was extracted in duplicate and included outcome definitions, general and time-to-event specific methods and handling of time-to-event relevant trial characteristics.

Results The included reviews analyzed 217 time-to-event outcomes (Median: 2; IQR 1–2), most frequently overall survival (41%). Outcome definitions were provided for less than half of time-to-event outcomes (48%). Few reviews specified general methods, e.g., included analysis types (intention-to-treat, per protocol) (35%) and adjustment of effect estimates (12%). Sources that review authors used for retrieval of time-to-event summary data from publications varied substantially. Most frequently reported were direct inclusion of HRs (64%) and reference to established guidance without further specification (46%). Study characteristics important to time-to-event analysis, such as variable follow-up, informative censoring or proportional hazards, were rarely reported. If presented, complementary absolute effect estimates calculated based on the pooled HR were incorrectly calculated (14%) or correct but falsely labeled (11%) in several reviews.

[†] Nina Kreuzberger and Nicole Skoetz contributed equally to this work.

*Correspondence:
Marius Goldkuhle
marius.goldkuhle@uk-koeln.de

Full list of author information is available at the end of the article

Conclusions Our findings indicate that limitations in reporting of trial time-to-event analyses translate to the review level as well. Inconsistent reporting of meta-analyses of time-to-event outcomes necessitates additional reporting standards.

Keywords Systematic review, Meta-analysis, Time-to-event outcomes, Survival analysis, Reporting quality, Quantitative analysis

Background

Systematic reviews of time-to-event analyses, also referred to as survival analyses, provide fundamental evidence in many fields of research [1]. In randomized controlled trials (RCTs) and non-randomized studies alike, time-to-event outcomes combine the occurrence of events with information about how long they took to occur, considering the observation time of participants with and without an event (censored observations). Particularly relevant for medical research, e.g., in oncology or cardiology, this allows analyzing longer-term outcomes or outcomes which might occur in all participants at some point, such as death [2, 3]. Prominent measures of time-to-event data include Kaplan and Meier survival plots and probabilities, and the hazard ratio (HR) for between-group comparisons [1, 4]. For trial-level data, HRs are commonly calculated using the Cox proportional hazards regression model, which allows the consideration of relevant covariates [4]. Aggregate data meta-analyses pool HRs from individual RCTs.

Common methods of time-to-event analysis assume non-informative censoring, requiring that the distributions of event times and censoring times do not provide any information about each other [4]. In many circumstances, this is most vividly explained by the less restrictive assumption of random censoring: censoring should occur as if those censored were randomly drawn, which is questionable, for instance, when censoring occurs for adverse events or other reasons related to the intervention [2, 4, 5]. The Cox model assumes at least approximate proportionality of the hazards of the compared groups over time [4, 6, 7]. If violated, the HR is best interpreted as an average over the observed period, which might affect the between-study heterogeneity in meta-analyses of trials with variable follow-up durations [8]. Other challenges include competing events (like death) that may preclude the event of interest [9, 10], treatment switching [11, 12] and the poor reporting in trial publications [13–18]. Furthermore, the HR as relative effect measure is not always straightforward to interpret and might endanger exaggerated interpretation of treatment effects [19, 20]. This is why leading organizations in evidence synthesis recommend complementary absolute effect estimates, e.g., natural frequencies, risk differences or the number-needed-to-treat, calculated based on the pooled HR, as additional presentation of results [21–24].

To foster prioritization of remedial actions and further research, this meta-epidemiological analysis explored characteristics, reporting and methods of systematic reviews including time-to-event meta-analyses per review and per analyzed time-to-event outcome.

Methods

We followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) checklist´s adaption to meta-epidemiological research [25]. The project is registered under: osf.io/5qxbd.

Eligibility criteria

We included systematic reviews of interventions with at least one pairwise meta-analysis based on aggregate data from RCTs. Reviews must have meta-analyzed a minimum of two trials evaluating a health-related time-to-event outcome using the HR. We considered any peer-reviewed full-text review published in English on any intervention type, medical field or setting. We excluded network meta-analyses, previous versions of updated reviews and co-publications of Cochrane reviews (CR).

Identification and selection of reviews and studies

We chose an overall sample size of 100 reviews, 50 CR and 50 Non-Cochrane reviews (nCR), to represent a diverse and methodologically up-to-date landscape of systematic reviews. We systematically selected CR according to our inclusion criteria in a backwards fashion starting with publications in August 2020 from the Cochrane Database of Systematic Reviews. To identify a representative sample of nCR published during the publication timeframe corresponding to all eligible CR (28/02/2017 to 18/08/2020), an experienced information specialist (IM) developed a search strategy for Medline/Ovid (Table A1, Additional File 1; search date: 08/02/2021). To ensure the relevance of the included nCR, we limited our search to systematic reviews published in Core Clinical Journals, as defined by the U.S National Library of Medicine [26].

Titles and abstracts were screened and potentially eligible reviews selected based on their full text publications by two authors (MG, NK) independently. Discrepancies were resolved by discussion and, if necessary, by involving a third author (NS). We identified more potentially eligible nCR and selected, stratified by publication year, a random sample of 50 nCR published within the publication time of the CR. The random selection was performed by a third author not involved in the selection of reviews and extraction of data by using a random numbers generator.

Our assessment was limited to meta-analyses of outcomes under the main review comparison, which was in case of multiple comparisons the first comparison reported in the abstract. We did not assess any subgroup or sensitivity analyses. When the total and subtotal results for an individual eligible meta-analysis were reported, we extracted data on the total summary measure only. For feasibility, we excluded comparisons with one or more time-to-event meta-analyses that included more than 20 trials.

Data extraction

All extractions were performed in duplicate by two authors (NK, CI, CH, AB, MG) (25 reviews extracted independently, 75 reviews double-checked after sufficient agreement). The data extraction sheet was developed a priori, piloted, and implemented on the data management platform Ninox (ninox.com). It was developed based on experience from previous methodological surveys of systematic reviews and based on published proposals for the reporting of time-to-event analyses in study publications [14, 15, 20, 27]. Any potential discrepancies were resolved by discussion and, where necessary, by involving a third author (NS). We provide our complete extraction sheets, including all items, their descriptions and response options in the appendix (Table A2, Additional File 1).

Statistical analysis

Data are analyzed descriptively by means of absolute and relative frequencies for categorical data and medians, means and variability measures for continuous data.

We anticipated considerable differences in the handling of time-to-event specific trial characteristics between systematic review types (CR vs. nCR) and between different outcomes. Cochrane reviews are published under strict accordance with the Cochrane Handbook and without wording restrictions [22]. Because of these distinctions, we report our findings stratified by CR and nCR, even though our primary intent is not to compare review types. In addition, we present our findings stratified into three outcome categories in appendix tables that extend the data included in the text: Overall survival/all-cause mortality, composite outcomes including all-cause mortality and outcomes not including all-cause mortality, because of their different susceptibility to competing events, measurement robustness and relative importance.

Results

We provide a flow diagram (A3), a description of our search results (A4) and a list of all included reviews (A5) in the appendix (Additional File 1).

Characteristics of included reviews

The majority of reviews were published in 2017 and most frequently addressed oncological research questions in adults at advanced clinical stages and compared biologics/drugs to biologics/drugs (Table 1; extended data in appendix-table A6, Additional File 1).

They included a median of five studies (interquartile range (IQR) 4–8) with a median of four studies (IQR 3–7) in time-to-event outcome analyses. The median overall population was 1722 participants (IQR 978–4390) with 811 participants (IQR 308–2876) in time-to-event meta-analyses.

Analyzed time-to-event outcomes

Reviews reported overall a median of five (IQR 4–8) outcomes, containing a median of two (IQR 2–2) time-to-event outcomes in their methods sections. Median five (IQR 3–6) outcomes and a median (IQR 1–2) of two time-to-event outcomes were analyzed quantitatively in meta-analyses. In several reviews, not all time-to-event outcomes mentioned in the methods were analyzed, primarily because the included trials did not assess the respective outcome. The most frequently planned time-to-event outcomes were overall survival, all-cause death or death from any cause in 89% and progression-free survival in 39% of reviews. They were also the most commonly analyzed outcomes. The majority of reviews (69% (69/100 reviews); CR: 92% (46/50 reviews), nCR: 46% (23/50 reviews)) selected a time-to-event outcome as primary outcome, which commonly was overall survival, all-cause mortality or death from any cause (appendix-table A7, Additional File 1).

Overall, the reviews analyzed 217 individual time-to-event outcomes (Table 2; extended data in appendix-table A7, Additional File 1). If grouped by outcome type, 41% (89/217 outcomes) of analyzed outcomes solely addressed the outcome event all-cause death, i.e., overall survival or all-cause mortality, 29% (63/217 outcomes) were composite outcomes that included all-cause death as outcome event, e.g., progression-free survival, and 30% (65/217 outcomes) were outcomes that did not include all-cause death. Outcome definitions were provided for less than half (48% (104/217 outcomes); CR: 83% (77/93 outcomes), nCR: 22% (27/124 outcomes)) of all review time-to-event outcomes. 22% (47/217 outcomes) of outcomes were composite outcomes for which composing events were commonly reported. Starting points of follow-up for individual outcomes were not

Table 1 Summary of review characteristics

Reviews		**Overall (*N* = 100)**	**Cochrane (*n* = 50)**	**Non-Cochrane (*n* = 50)**
Publication				
Publication year	2017	36% (36)	36% (18)	36% (18)
	2018	28% (28)	28% (14)	28% (14)
	2019	18% (18)	18% (9)	18% (9)
	2020	18% (18)	18% (9)	18% (9)
Journal Impact Factor 2021	Median (IQR)		11.87*	4.41 (3.33–6.18)
	Mean (Range)		11.87*	6.37 (1.817–35.86)
Update of previous review		27% (27)	50% (25)	4% (2)
Multiple comparisons		28% (28)	44% (22)	12% (6)
Population				
Medical field	Neoplasms	82% (82)	86% (43)	78% (39)
	Circulatory system	11% (11)	4% (2)	18% (9)
	Other	7% (7)	10% (5)	4% (2)
Medical condition	Breast cancer	13% (13)	10% (5)	16% (8)
	Colorectal cancer	9% (9)	8% (4)	10% (5)
	Non-small cell lung cancer	8% (8)	6% (3)	10% (5)
	Prostate cancer	6% (6)	12% (6)	0% (0)
	Other	7% (7)	62% (31)	64% (32)
Age group[#]	Adults	96% (96)	98% (49)	94% (47)
Comparisons				
	Biologics/ drug vs. Biologics/ drug	37% (37)	24% (12)	50% (25)
	Surgical procedure vs. Surgical procedure	7% (7)	8% (4)	6% (3)
	Biologics/ drug vs. Supportive/ Optimal care	4% (4)	4% (2)	4% (2)
	Biologics/ drug vs. Observation	4% (4)	0% (0)	8% (4)
	Other	48% (48)	64% (32)	32% (16)
Outcomes				
Planned outcomes	Median (IQR)	5 (4–8)	7 (5–8)	4 (3–5)
	Mean (Range)	5.79 (1–15)	6.82 (3–12)	4.67 (1–15)
TTE outcomes in methods	Median (IQR)	2 (2–2)	2 (2–3)	2 (2–2)
	Mean (Range)	2.39 (1–12)	2.17 (1–4)	2.62 (1–12)
	OS, ACM or death from any cause	89% (89)	88% (44)	90% (45)
	Progression-free survival	44% (44)	36% (18)	52% (26)
	Disease-free survival	13% (13)	16% (8)	10% (5)
	Myocardial infarction	5% (5)	0% (0)	10% (5)
	Stroke	5% (5)	0% (0)	10% (5)
	Other[§]	72% (72)	74% (37)	70% (35)
	Unclear/ Not reported	5% (5)	2% (2)	3% (3)
Outcomes analyzed	Median (IQR)	5 (3–6)	5 (4–6,75)	4 (2,25–5)
	Mean (Range)	4.83 (1–12)	5.34 (1–12)	4.32 (1–12)
TTE outcomes analyzed	Median (IQR)	2 (1–2)	2 (1–2)	2 (2–2)
	Mean (Range)	2.23 (1–12)	1.92 (1–4)	2.54 (1–12)
	OS, ACM or death from any cause	89% (89)	84% (42)	94% (47)
	Progression-free survival	39% (39)	26% (13)	52% (26)
	Disease-free survival	10% (10)	10% (5)	10% (5)
	Myocardial infarction	6% (6)	0% (0)	12% (6)
	Stroke	5% (5)	0% (0)	10% (5)
	Other	78% (78)	70% (35)	86% (43)
TTE outcome(s) in methods not among analyzed		11% (11)	20% (10)	2% (1)
Reasons for difference	Outcome not in trial(s)	7% (7)	14% (7)	0% (0)
	No TTE data in trial(s)	2% (2)	2% (1)	0% (0)
	Not pooled due to heterogeneity	1% (1)	2% (1)	0% (0)
	Not reported	1% (1)	0% (0)	2% (1)

Table 1 (continued)

Reviews		**Overall** **(*N* = 100)**	**Cochrane** **(*n* = 50)**	**Non-Cochrane** **(*n* = 50)**
	Not applicable	89% (89)	80% (40)	98% (49)
Sample size				
Included studies	Median (IQR)	5 (4–8)	5 (3–10)	6 (4–8)
	Mean (Range)	6.69 (2–24)	7.12 (2–24)	6.26 (2–19)
Total population	Median (IQR)	1722 (97–4390)	1415 (572–4022)	1866 (1395–4526)
	Mean (Range)	3877 (170–56004)	2795 (170–13216)	4911 (343–56004)
	Not reported	12% (12)	14% (7)	10% (5)
Studies in TTE-MA	Median (IQR)	4 (3–7)	4 (2–6)	5 (4–7)
	Mean (Range)	5.25 (2–19)	5.05 (2–19)	5.40 (2–19)
Population in TTE-MA	Median (IQR)	811 (308–2876)	711 (177–2327)	1042 (698–3173)
	Mean (Range)	5745 (181–38723)	2656 (181–13949)[§]	8985 (482–38723)
	Not reported	23% (49)	8% (7)	34% (42)

Abbreviations IQR = interquartile range, MA = meta-analysis, TTE = time-to-event

[*] All CR share the impact factor of the Cochrane Database of Systematic Reviews; [#] One CR included children and adults, for three nCR no information was provided. [§] None of the analyzed outcomes under "Other" were used in more than four reviews. [§] This number exceeds the upper range of the total review population, since the review including the respective analysis did not report a total review population)

Table 2 Characteristics of time-to-event outcomes in reviews

		Reviews			**TTE outcomes**		
		Overall (*N* = 100)	Cochrane (*n* = 50)	Non-Cochrane (*n* = 50)	Overall (*N* = 217)	Cochrane (*n* = 93)	Non-Cochrane (*n* = 124)
TTE outcome definitions provided		55% (55)	80% (40)	30% (15)	48% (104)	83% (77)	22% (27)
Composite TTE outcomes	Yes	39% (39)	42% (21)	36% (18)	22% (47)	24% (22)	20% (25)
	Unclear/ Not reported	28% (28)	12% (6)	44% (22)	14% (30)	8% (7)	19% (23)
Composite events described for composite outcome		87% (34)	90% (19)	83% (15)	87% (41)	91% (20)	84% (21)
All-cause death part of outcome	Yes	91% (91)	86% (43)	96% (48)	57% (124)	65% (60)	52% (64)
	Unclear	31% (31)	14% (7)	48% (24)	14% (31)	8% (7)	19% (24)
Death as competing event possible	Yes	29% (29)	40% (20)	18% (9)	29% (64)	28% (26)	31% (38)
	Unclear	31% (31)	16% (8)	46% (23)	15% (32)	10% (9)	19% (23)
Outcomes reporting as events or absence	Absence of event only	61% (61)	48% (24)	74% (37)	54% (118)	53% (49)	56% (69)
	Event only	24% (24)	26% (13)	22% (11)	37% (81)	30% (28)	43% (53)
	Both (with reasoning)	5% (5)	10% (5)	0% (0)	5% (10)	11% (10)	0% (0)
	Both (without reasoning)	8% (8)	12% (6)	4% (2)	4% (8)	6% (6)	2% (2)
Follow-up start in outcome definitions	Randomization	32% (32)	48% (24)	16% (8)	32% (70)	57% (53)	14% (17)
	Allocated treatment	2% (2)	2% (1)	2% (1)	2% (4)	2% (2)	2% (2)
	Enrollment	2% (2)	2% (1)	0% (0)	3% (6)	6% (6)	0% (0)
	Not applicable (e.g., start of follow-up not reported)	62% (62)	46% (23)	82% (41)	62% (135)	32% (30)	85% (105)

Abbreviations ACM = all-cause mortality, HR = hazard ratio, IPD = individual participant data, OS = overall survival, TTE = time to event

regularly defined, but if they were, it most often was randomization.

Results of time-to-event meta-analyses were typically (54% (118/217 outcomes)) reported as absence of event only, e.g., survival, as opposed to reporting of results as events, e.g., death. Some review authors reported the same outcome inconsistently as event and absence thereof, few explained their decision. Outcomes not including all-cause death were most frequently reported as events (appendix-table A7, Additional File 1).

General methodological characteristics of reviews and their time-to-event outcome analyses

Less than half (43% (43/100 reviews); CR: 72% (36/50 reviews), nCR: 14% (7/50 reviews)) of reviews reported the primary trial analysis (e.g., intention-to-treat, per protocol, etc.) that was preferred for meta-analysis (appendix-table A8, Additional File 1). Review authors preferred analyses according to the intention-to-treat approach in all cases. The actually included analysis principles, also commonly per intention-to-treat, were reported in 38% (38/100 reviews; CR: 54% (27/50 reviews), nCR: 22% (11/50 reviews)) of reviews. Few reviews (15% (15/100 reviews); CR: 26% (13/50 reviews), nCR: 4% (2/100 reviews)) reported if they used adjusted or unadjusted effect estimates from trials for meta-analyses. The most frequent options were eligibility of both, analyses adjusted for covariates or unadjusted analyses only. Handling divergently adjusted effects was sporadically reported: e.g., as potential source of heterogeneity between trials or as combined in the same analysis, and discrepancies in adjustment of effects were seldomly mentioned. None of the reviews compared covariate adjusted and unadjusted trial estimates.

To address statistical heterogeneity between trials, review authors commonly reported subgroup analyses or a random-effects meta-analysis (Table 3). Time-to-event data was most frequently pooled with random-effects models and the inverse variance method.

Table 3 General methodological characteristics of time-to-event outcome meta-analyses in reviews

Reviews		**Overall** **(*N* = 100)**	**Cochrane** **(*n* = 50)**	**Non-Cochrane** **(*n* = 50)**
Methods to pool time-to-event data specified in review methods				
Random-effects model	Inverse variance	38% (38)	48% (24)	28% (14)
	Hartung-Knapp-Sidik-Jonkman	2% (2)	0 (0%)	2% (2)
	Not reported	6% (6)	4% (2)	8% (4)
Fixed-effect model	Inverse variance	7% (7)	12% (6)	2% (1)
	Other (Peto or Mantel-Haenszel)	3% (3)	4% (2)	2% (1)
	Not reported	1% (1)	0% (0)	2% (1)
Either, depending on heterogenity	Inverse variance	15% (15)	18% (9)	12% (6)
	Mantel-Haenszel or inverse variance	3% (3)	0% (0)	6% (3)
	Not reported	22% (22)	22% (11)	22% (11)
Both (e.g., one as sensitivity analysis)	Inverse variance	5% (5)	4% (2)	6% (3)
	Other*	2% (2)	0% (0)	4% (2)
	Not reported	5% (5)	6% (3)	4% (2)
Model and method not reported		2% (2)	2% (1)	2% (1)
Heterogeneity parameter specification[#]	DerSimonian-Laird	31% (27)	22% (9)	39% (18)
	Paule-Mandel	2% (2)	0%	4% (2)
	Not reported	66% (58)	78% (32)	57% (26)
Methods to pool time-to-event data used in results (e.g., indicated in forest plot)				
Random effects	Inverse variance	57% (57)	72% (36)	42% (21)
	Not reported	11% (11)	0% (0)	22% (11)
Fixed-effect	Inverse variance	24% (24)	24% (12)	24% (12)
	Peto	2% (2)	2% (1)	2% (1)
	Not reported	4% (4)	0% (0)	8% (4)
Both (e.g., one as sensitivity analysis)	Inverse variance	4% (4)	2% (1)	6% (3)
	Not reported	2% (2)	0% (0)	4% (2)
Model and method not reported		1% (1)	2% (1)	0% (0)
Handling of heterogeneity				
Heterogeneity handling per review	Subgroup analyses	59% (59)	80% (40)	38% (19)
	Random-effects meta-analysis performed	55% (55)	54% (27)	56% (28)
	No pooling if too heterogeneous	12% (12)	24% (12)	0% (0)
	Meta-regression	7% (7)	4% (2)	10% (5)
	Unclear/ Not reported	15% (15)	2% (2)	26% (13)
Heterogeneity handling per outcome	Subgroup analyses	9% (9)	16% (8)	2% (1)
	Subgroup analyses and meta-regression	2% (2)	4% (2)	0% (0)
	Meta-regression	1% (1)	0% (0)	2% (1)
	Shared frailty survival model (based on IPD)	1% (1)	0% (0)	2% (1)
	Not reported per outcome	87% (86)	80% (40)	94% (47)

Abbreviations HR = hazard ratio IPD = individual participant data, MA = meta-analysis

* Other includes one review that reported the use of the inverse variance method under a random effects model and, in case of available data (O-E and variance) in trial publications, the Peto's method for a fixed-effects model. It includes a second review that reported the use of the inverse variance method for random-effects meta-analysis and the Mantel-Haenzel method for a separate fixed-effects meta-analysis; [#] Applies to 87 reviews (CR: 41; nCR: 46) which reported to perform a random-effects model in their methods, either alone, depending on the degree of heterogeneity or together with a fixed effects model

Time-to-event specific methods used in the included systematic reviews

Methods to obtain time-to-event summary data from included trials varied substantially (Table 4; extended data in appendix-table A9, Additional File 1). In 36% (18/50) of nCR no approaches to retrieve time-to-event data were reported. The most frequently used methods were direct inclusion of available HRs in 64% (64/100 reviews), particular sets of methods, e.g., those collected by Tierney [1] or Parmar [28], in 46% (46/100 reviews), and available log(HR)s with a standard error in 16% (16/100 reviews) of reviews. Data sources, if reported for individual outcomes, were most frequently HRs and confidence intervals, P-values with other information and survival curves. In total, 63% (63/100) of reviews reported at least one approach to recalculate summary time-to-event data from trial publications, in contrast to 16% (16/100) reviews that reported only HRs and confidence intervals. For individual review outcomes, 18% (18/100) of reviews reported any approach to recalculate summary time-to-event data. Review authors who reported direct inclusion of a HR or log(HR), most often did not provide further specifications of the eligible HR. Sporadic specifications include HRs from Cox models, log-rank tests or survival curves or HRs recalculated from median survival times.

Risk of bias assessments were predominately performed with the Cochrane Risk of Bias 1 tool on study

Table 4 Time-to-event specific methods applied in reviews

		Reviews			**TTE outcomes**		
		Overall (*N* = 100)	Cochrane (*n* = 50)	Non-Cochrane (*n* = 50)	Overall (*N* = 217)	Cochrane (*n* = 93)	Non-Co-chrane (*n* = 124)
HR type eligible in reviews	HR/ log(HR) not further specified	91% (91)	94% (47)	88% (44)	NA	NA	NA
	HR/ log(H)R from Cox model	2% (2)	4% (2)	0% (0)	NA	NA	NA
	HR / log(HR) from Cox model, log-rank test and Kaplan Meier Curve	1% (1)	2% (1)	0% (0)	NA	NA	NA
	Not reported	6% (6)	0% (0)	12% (6)	NA	NA	NA
HR types eligible per outcome	HR/ log(HR) from Cox model	1% (1)	2% (1)	0% (0)	0% (1)	1% (1)	0% (0)
	HR / log(HR) from median survival times and confidence intervals	1% (1)	2% (1)	0% (0)	0% (1)	1% (1)	0% (0)
	Unclear/ Not reported	98% (98)	96% (48)	100% (50)	99% (215)	98% (91)	100% (124)
Methods to obtain TTE data per review	HR and confidence intervals	64% (64)	66% (33)	62% (31)	NA	NA	NA
	Specified set of methods (e.g., Tierney 2007, Cochrane Handbook, etc.)	46% (46)	76% (38)	16% (8)	NA	NA	NA
	log(HR) and standard error	16% (16)	26% (13)	6% (3)	NA	NA	NA
	Survival curves	13% (13)	14% (7)	12% (6)	NA	NA	NA
	HR with other information (e.g., events per arm, total events, etc.)	10% (10)	16% (8)	4% (2)	NA	NA	NA
	P-value with additional information (e.g., total events, etc.)	8% (8)	8% (4)	8% (4)	NA	NA	NA
	Other*	11% (11)	14% (7)	8% (4)	NA	NA	NA
	Unclear/ Not reported	22% (22)	8% (4)	36% (18)	NA	NA	NA
Recalculation of TTE data reported for outcome		18% (18)	34% (17)	2% (1)	NA	NA	NA
Methods to obtain TTE data for outcome	HR and confidence intervals	9% (9)	18% (9)	0% (0)	7% (16)	17% (16)	0% (0)
	P-value together with additional information (e.g., events, etc.)	4% (4)	8% (4)	0% (0)	5% (10)	11% (10)	0% (0)
	Survival curves	7% (7)	12% (6)	2% (1)	5% (10)	10% (9)	1% (1)
	Other#	6% (6)	12% (6)	0% (0)	5% (11)	12% (11)	0% (0)
	Unclear/ Not reported	90% (90)	80% (40)	100% (50)	89% (193)	75% (70)	99% (123)

Abbreviations ACM = all-cause mortality, HR = hazard ratio, IPD = individual participant data, OS = overall survival, TTE = time to event

* Other includes, for example, individual participant data (recalculated or from publication), median survival times and time-point specific survival times; # Other includes, for example, HR with other information (e.g., events per arm, total events, etc.), time-point specific survival times, IPD recalculated or from publication and median survival times

level. Time-to-event specific risk of bias criteria were applied in four CR, addressing informative censoring and competing events. Grading of Recommendations Assessment, Development and Evaluation (GRADE) certainty of the evidence ratings were included in all CR and 14% (7/50 reviews) of nCR, none of which described criteria one could consider time-to-event specific. 45% of all assessed time-to-event outcomes were included in GRADE Summary of Findings tables.

Results of meta-analyses

Table 5 presents the results of time-to-event meta-analyses in included reviews. According to the confidence intervals of individual time-to-event meta-analyses, most reviews included at least one analysis which showed statistically significant or non-significant effects in favor of the intervention as indicated by the review authors. Most analyses showed relative risk reductions between 1 and 0.5 as indicated by their point estimates.

Regarding the directionality of the HR as a relative effect measure, in almost all reviews the pooled HRs were applicable to events, meaning that review authors considered occurrence of the event of interest as basis for their calculation, and a pooled HR≤1 indicated a decreased risk. Inversion of trial HRs in accordance with the direction of the meta-analysis HR was reported for none of the time-to-event outcomes.

Handling of specific trial characteristics with relevance for time-to-event outcome analysis and interpretation

Heterogenous outcome definitions between trials were discussed in eight reviews (appendix-table A10, Additional File 1). Follow-up durations of included trials were frequently not specified. If they were, most often, reviews (5%, 5/100) reported a required minimum duration for all review outcomes or, for individual time-to-event outcomes, reported to use the longest follow-up available for each trial (8% (8/100 reviews)). Few reviews indicated how they dealt with potential varying follow-up between

Table 5 Summary of time-to-event meta-analyses results in the included reviews

		Reviews			**TTE outcomes**		
		Overall (*N* = 100)	Cochrane (*n* = 50)	Non-Cochrane (*n* = 50)	Overall (*N* = 217)	Cochrane (*n* = 93)	Non-Cochrane (*n* = 124)
*Result**	Favourable, statistically significant	52% (52)	40% (20)	64% (32)	40% (86)	32% (30)	45% (56)
	Favourable, statistically non-significant	57% (57)	54% (27)	60% (30)	38% (83)	40% (37)	37% (46)
	Unfavourable, statistically significant	7% (7)	6% (3)	8% (4)	5% (10)	3% (3)	6% (7)
	Unfavourable, statistically non-significant	25% (25)	36% (18)	14% (7)	16% (35)	23% (21)	11% (14)
	Direction of effect undetermined (HR = 1)	3% (3)	4% (2)	2% (1)	1% (3)	2% (2)	1% (1)
Size of HR point estimate	> 2	5% (5)	6% (3)	4% (2)	3% (6)	4% (4)	2% (2)
	< 2, > 1	24% (24)	34% (17)	14% (7)	18% (38)	20% (19)	18% (19)
	1	3% (3)	4% (2)	2% (1)	1% (3)	2% (2)	1% (1)
	< 1, > 0.5	83% (83)	72% (36)	94% (47)	72% (156)	66% (61)	77% (95)
	< 0.5	13% (13)	12% (6)	14% (7)	6% (14)	8% (7)	6% (7)
I^2 value	Median (IQR)	NA	NA	NA	5 (0–48.81)	0 (0–41.935)	20 (0–55.9)
	Mean (range)	NA	NA	NA	25.36 (0–97)	21.22 (0–85.36)	28.55 (0–97)
	Not reported	NA	NA	NA	3% (6)	1% (1)	4% (5)
HR for event or non event	Event	98% (98)	98% (49)	98% (49)	99% (214)	99% (92)	98% (122)
	Unclear	2% (2)	2% (1)	2% (1)	1% (3)	1% (1)	2% (2)
Interpretation of HR < 1[#]	Decreased risk of event	96% (96)	96% (48)	96% (48)	98% (212)	98% (91)	98% (121)
	Increased risk of event	2% (2)	2% (1)	2% (1)	1% (2)	1% (1)	1% (1)
	Unclear	2% (2)	2% (1)	2% (1)	1% (3)	1% (1)	2% (2)
Trial HRs inverted	Not reported	100% (100)	100% (50)	100% (50)	100% (217)	100% (93)	100% (124)

Abbreviations: HR = hazard ratio; IQR = interquartile range

* The designation “favourable/ unfavourable” is based on the review authors definition of the intervention, e.g., specified in Summary of Findings tables, e.g., “favourable”, when they interpreted a HR < 1 as beneficial for their designated intervention. It is based on the point estimate. [#] Interpretation of a HR < 1 as decreased/ increased risk of the event is based on the review authors designated intervention and their interpretation. If in their forest plot a HR < 1 for overall survival was reported as “favors the intervention”, for example, it was clear that the HR represents a decreased the risk of the event (death), even though it is referred to as absence of the event (overall survival) by the review authors

trials, e.g., through sensitivity analyses or meta-regression. Some mentioned varying follow-up times in the results and discussion sections.

Some form of handling of missing outcome data in trials was mentioned in all CR and 48% (24/50 reviews) of nCR. Most review authors included it among risk of bias criteria and it was mentioned in the majority of reviews, but outcome specifically only in two CR. Informative censoring, competing events, in particular deaths as competing events, and treatment switching were reported as risk of bias criteria and otherwise mentioned in the results and discussion sections of singular reviews. Three reviews described sensitivity analyses for treatment switching, e.g., according to the rates of switchers. No review reported an assessment of the proportional hazards assumption of included time-to-event analyses.

Absolute effects for communication of findings

44% (44/100 reviews; CR: 80% (40/50 reviews), nCR: 8% (4/50 reviews)) of reviews used complementary absolute effect estimates (e.g., natural frequencies or risk differences) that were calculated based on the pooled HRs to communicate the results of their time-to-event meta-analyses (appendix-table A11, Additional File 1). None of the reviews reported the use of more complex and direct methods to estimate absolute effects for their meta-analyzed time-to-event outcomes, such as separate meta-analyses of the risk difference, simulation or bivariate models incorporating a baseline risk in the meta-analysis [29]. In six reviews, authors did explicitly not calculate absolute effects because outcomes "were time-to-event outcomes". Most frequently used were natural frequencies, e.g., the number of participants experiencing an event until a given time-point among 1000 participants under observation. Baseline risks used to calculate absolute effects were mostly applicable to events. In a third of reviews, the outcome descriptions did not match the absolute effect descriptions (e.g., overall survival used for result reporting throughout the review, but absolute effects reported as all-cause mortality). The absolute effect description was adapted in seven CR with reasoning and in two reviews without any reasoning. Overall, correctly calculated and reported absolute effects were available in about a quarter (22% (22/100 reviews); CR: 40% (20/50 reviews), nCR: 4% (2/50 reviews)) of reviews. Correctly calculated but discrepantly labeled absolute effects were available in 10% (5/50 reviews) of CR. They were incorrect in 12% (6/50 reviews) of CR, where a HR for events was multiplied with a baseline risk for absence of events. In 7% (7/100 reviews) of reviews the interpretation of the absolute effect was unclear, because the direction of the baseline risk was not clear.

Discussion

In summary, the majority of reviews that included meta-analyses of time-to-event outcomes addressed neoplasms and assessed all-cause mortality. Less than half of their analyzed outcomes were time-to event outcomes and outcome definitions for time-to-event outcomes were provided only in half of reviews.

The eligible analyses, e.g., intention-to-treat or per protocol, or unadjusted or adjusted analyses, were not always reported and information regarding the actually included analyses was rarely given. Methods to obtain time-to-event data varied substantially, prominently reported were direct inclusion of the HR and complete sets of recalculation methods, and were seldomly provided for individual outcomes.

In most cases the random-effects model was used for summarizing the estimated study treatment effects. The most prominently used method was the DerSimonian-Laird method [30]. Only in two reviews the Hartung-Knapp-Sidik-Jonkman (HKSJ) method was used, contrary to the recommendation of using HKSJ as standard approach for random-effects meta-analyses in the literature [31].

Although outcomes were most frequently presented in text as absence of event, the meta-analyzed HRs were almost entirely calculated based on rates of events and a HR < 1 indicated a lower risk in experimental arm. The results of the assessed meta-analyses were predominantly favorable (in 78%). Absolute effects based on these results were common in CR, however, in some reviews these were described or calculated incorrectly.

Lastly, trial characteristics with relevance to time-to-event outcome analysis, for example varying follow-up, adjusted/unadjusted effect estimates, informative censoring, competing events, treatment switching and proportional hazards, were only infrequently included in additional assessments (e.g., sensitivity analyses, risk of bias or certainty assessments) and seldomly mentioned throughout review texts at all.

To our knowledge, and in contrast to previous methodological surveys of general systematic review methods and reporting, this is the first assessment of time-to-event specific methods and time-to-event specific reporting of a sample of systematic reviews [27, 32]. For reporting of time-to-event analyses in study publications, several previous assessments found considerable limitations: e.g., infrequent reporting of start and end points of the observation, censoring reasons, follow-up times and calculation, as well as exact numbers of events and censored observations [13–17]. Furthermore, assumptions of the models used, such as the proportional hazards assumption, were often not verified or important details of statistical modelling were withheld [14, 16].

These findings were confirmed by a recent study that assessed trials included in meta-analyses of a random subset of the here analyzed reviews [33]. The study identified shortcomings in the reporting of time-to-event analyses in trial publications that are consistent with the limitations visible in review publications. These shortcomings relate to, for example, outcome definitions and trial characteristics relevant for the validity of time-to-event analyses. Some items, such as the types of analyses (e.g., intention-to-treat) and the adjustment of estimates were more consistently reported in trial publications than in their including reviews. Limitations and discrepancies in reporting of time-to-event analysis in trial publications introduce considerable difficulties for review authors. It seems reasonable to assume that the problematic reporting at trial level has an impact on reporting on review level. However, the limited reporting in trial publications was hardly or not at all addressed in the reviews themselves.

Methodological guidance for including time-to-event data in aggregate data meta-analyses exists and authors should follow these guidelines [1, 22, 28]. Irrespective of whether lack and variability of reporting in many of our assessed items translates into a neglect of the respective issues, limited reporting in review publications endangers misinterpretation of effects and certainty therein. For example, in a few cases the directionality of the pooled HR was unclear, either because it remained unclear whether the HR was representative of events or the absence of events, or because it was unclear whether an HR < 1 favored the intervention or the comparator. At worst, this can lead to a misinterpretation of effects in the opposite direction.

To enhance the quality of systematic reviews including aggregate data meta-analyses of time-to-event outcomes, a central implication of our assessment is the need for reporting standards that extend the currently available general reporting guidance [23]. While reporting standards for study publications that report on time-to-event analyses have been proposed, e.g., by Altman et al. [15] and Abraira et al. [14], such standards are currently not available for systematic reviews. As a product of the here reported methodological study, a respective reporting guideline is currently in development [34]. In addition, focused methodological guidance could increase review conductors' awareness for more specific hardships, e.g., informative censoring, competing events and proportional hazards [5, 24]. Editors of biomedical journals also have a role in assuring the quality of the systematic reviews they publish. Without standards for the analysis and reporting of time-to-event outcomes in systematic reviews, they face certain difficulties. Future guidelines would therefore also be extremely helpful for them.

Variability in reported sources for time-to-event data in included reviews is a core finding of our study. Exclusion of studies because of perceived insufficient data does not only lead to less precise meta-analysis results but might also cause bias. The variety of approaches to recalculate time-to-event summary data from primary trials is therefore a critical predictor of review quality. In addition, the robustness of individual recalculation methods should be considered, and hierarchical selection of approaches could be advised. A recent study, for example, compared survival curve-based approaches to published HR and found that the method to reconstruct individual participant data by Guyot et al. [35] had the smallest bias and largest precision. Should established recalculation methods fail due to absence of data, it must be discussed whether to resort to less conventional methods. In an assessment of Cochrane reviews, Salika et al. [18] found that binary analysis of time-to-event outcomes could be feasible under low event probabilities. They suggest complementary log-log links as possible option for the case that only binary data are available. Particularly for situations with non-proportional hazards and for ease of interpretation, meta-analysis on the restricted mean survival time is an attractive option, which requires, however, recalculation of individual participant data using one of many available approaches [36].

Strengths and limitations

We provide an in-depth view on the current landscape of systematic reviews following rigorous, prespecified methods. Duplicate performance of relevant steps and structured forms warrant robustness despite the considerable extent of data. To deal with the large amount of data for extraction, we chose a two stage process: First, we extracted the data for 25 reviews in duplicate and then, with sufficient agreement (less than < 10% discrepancy), we continued extracting data in a double-checked manner, with one project participant performing the extraction and a second independent project participant checking the extracted data in the review publications.

We aligned our sample on a fixed number of CR to ensure a view on methods and reporting in reviews from a well-established source often considered the gold-standard. This may have led us to underestimate the size of the observed problems in the overall biomedical literature. For inference to the general landscape of systematic reviews, we included a large number of non-Cochrane reviews. We restricted the reviews to those published in Core Clinical Journals to ensure an elevated publication standard of the reviews, thereby also endangering possible underestimation of visible problems in relation to the overall literature. Because we identified a larger number of potentially eligible non-Cochrane reviews and for feasibility of our endeavor under its set goals, we drew a random sample from the non-Cochrane reviews stratified by publication years to meet the number of Cochrane reviews. To limit any potential selection bias, despite the application of the Core Clinical Journals filter, the random selection was performed independently by a project participant not involved in the selection of articles and extraction of data.

We limited our assessment to comparisons of ≤20 trials for feasibility of a separate, independent analysis step, which could impact the results [33]. As the total number of excluded reviews was minimal, 3% CR and 5% nCR during full-text screening, we believe the impact to be minimal. Finally, and relevant for literature-based methodological assessments in general, we must stress that review reporting does not necessarily correspond with review performance, e.g., regarding authors approaches to retrieve time-to-event data. Yet, reporting of analytic steps may reflect their relative importance to authors and might constitute a sufficient surrogate for their performance.

Conclusions

We identified variable and often insufficient reporting of time-to-event specific features in current systematic reviews including aggregate data meta-analyses based on the HR. Review authors should rigorously utilize available methodological guidance for the conduct of their analysis. Reporting standards might improve the quality of respective reviews.

Abbreviations

ACM	All-cause mortality
CR	Cochrane review
GRADE	Grading of Recommendations Assessment, Development and Evaluation
HKSJ	Hartung-Knapp-Sidik-Jonkman
HR	hazard ratio
IPD	Individual participant data
IQR	Interquartile range
MA	Meta-analysis
nCR	Non-Cochrane review
OS	Overall survival
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RCT	Randomized controlled trial
TTE	Time to event

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12874-024-02401-4.

Supplementary Material 1: Search strategy and extending tables

Acknowledgements

Not applicable.

Author contributions

MG: Conceptualization, data curation, formal analysis, investigation, methodology, project administration, writing – original draft, writing - review & editing; CH: Formal analysis, investigation, writing - review & editing; CI: Formal analysis, investigation, writing - review & editing; AMZ: Formal analysis, investigation, writing - review & editing; RB: Conceptualization, writing - review & editing; EvD: Conceptualization, writing - review & editing; LGH: Conceptualization, writing - review & editing; IM: systematic search; NK: Conceptualization, data curation, formal analysis, investigation, methodology; NS: Conceptualization, methodology, supervision, writing - review & editing.

Funding

Open Access funding enabled and organized by Projekt DEAL. Funded by the German Research Foundation (DFG) (project SK 146/2 – 1; https://gepris.dfg.de/gepris/projekt/437030575). The funding bodies played no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Data availability

The data supporting the conclusions of this article are presented in the manuscript and in extending tables in the appendix (Additional File 1). All raw data is available as Excel file upon request.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Registration

This meta-epidemiological study is registered under osf.io/5qxbd.

Competing interests

Two included Cochrane reviews were co-authored by a project participant (NS), who did not appraise data or resolve conflicts for these reviews. The same holds for one non-Cochrane review, which was co-authored by another project participant (LGH). All other authors declare that they have no competing interests.

Author details

[1]Institute of Public Health, Faculty of Medicine and University Hospital Cologne, University of Cologne, Kerpener Str. 62, 50937 Cologne, Germany
[2]Department of Medical Biometry, Institute for Quality and Efficiency in Health Care, Im Mediapark 8, D-50670 Cologne, Germany
[3]Princess Máxima Center for Pediatric Oncology, Heidelberglaan 25, Utrecht 3584CS, The Netherlands
[4]Research Center for Clinical Neuroimmunology and Neuroscience Basel (RC2NB), University Hospital Basel and University of Basel, Basel, Switzerland
[5]Department of Clinical Research, University Hospital Basel, University of Basel, Basel, Switzerland
[6]Meta-Research Innovation Center at Stanford (METRICS), Stanford University, Stanford, CA, USA
[7]Meta-Research Innovation Center Berlin (METRIC-B), Berlin Institute of Health, Berlin, Germany

Received: 17 October 2023 / Accepted: 4 November 2024
Published online: 25 November 2024

References

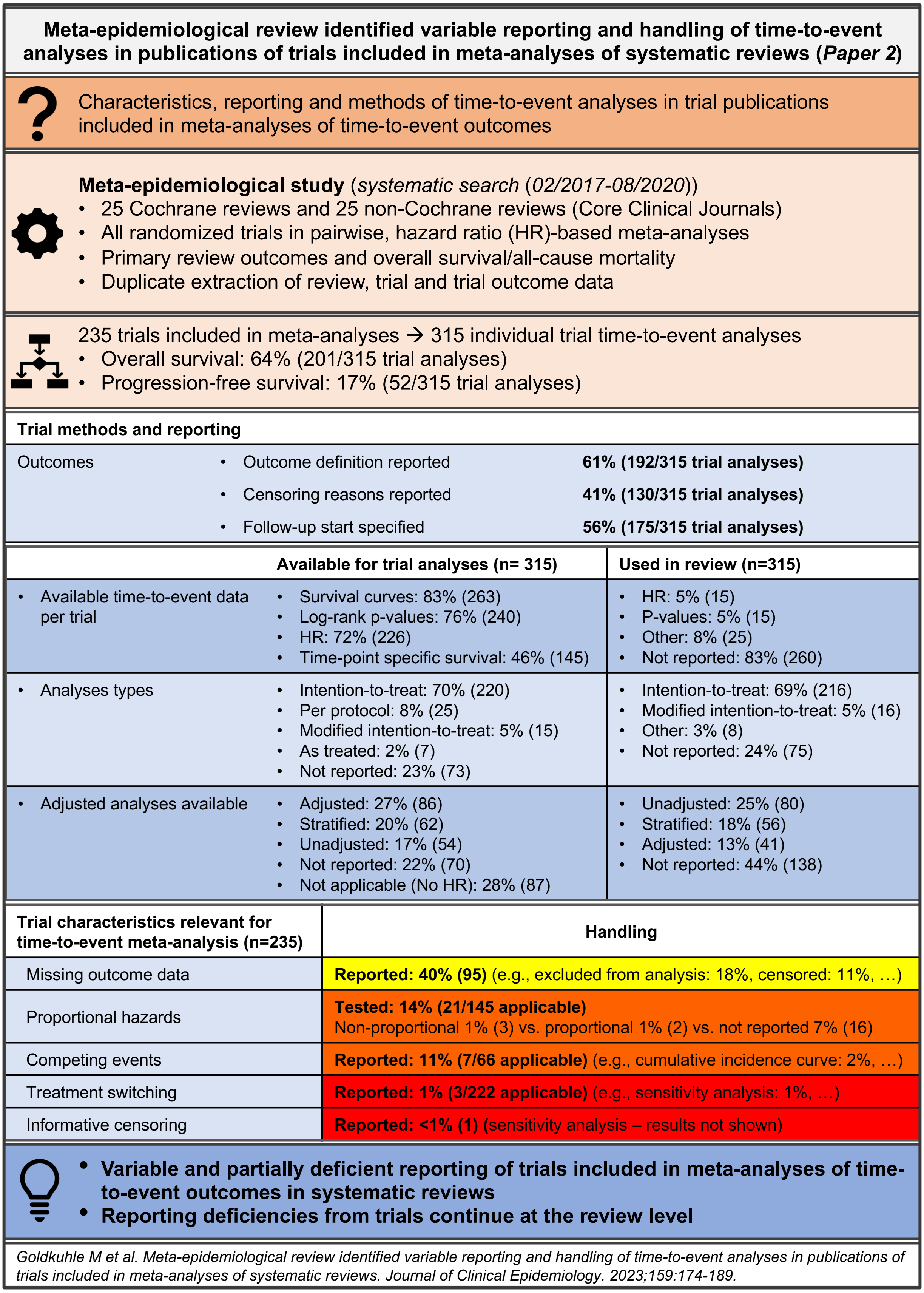
1. Tierney JF, Stewart LA, Ghersi D, Burdett S, Sydes MR. Practical methods for incorporating summary time-to-event data into meta-analysis. Trials. 2007;8:16.
2. Leung K-M, Elashoff RM, Afifi AA. CENSORING ISSUES IN SURVIVAL ANALYSIS. Annu Rev Public Health. 1997;18(1):83–104.
3. Lagakos SW. General right censoring and its impact on the analysis of survival data. Biometrics. 1979;35(1):139–56.
4. Kleinbaum DG, Klein M. Survival analysis. 3 ed. New York: Springer-; 2012.
5. Goldkuhle M, Bender R, Akl EA, van Dalen EC, Nevitt S, Mustafa RA, et al. GRADE guidelines: 29. Rating the certainty in time-to-event outcomes - study

limitations due to censoring of participants with missing data in intervention studies. J Clin Epidemiol. 2021;129:126–37.

6. Stensrud MJ, Hernán MA. Why test for proportional hazards? JAMA. 2020;323(14):1401–2.
7. Rulli E, Ghilotti F, Biagioli E, Porcu L, Marabese M, D'Incalci M, et al. Assessment of proportional hazard assumption in aggregate data: a systematic review on statistical methodology in clinical trials using time-to-event endpoint. Br J Cancer. 2018;119(12):1456–63.
8. Hernán MA. The hazards of hazard ratios. Epidemiol (Cambridge Mass). 2010;21(1):13–5.
9. Austin PC, Fine JP. Accounting for competing risks in randomized controlled trials: a review and recommendations for improvement. Stat Med. 2017;36(8):1203–9.
10. Schumacher M, Ohneberg K, Beyersmann J. Competing risk bias was common in a prominent medical journal. J Clin Epidemiol. 2016;80:135–6.
11. Sullivan TR, Latimer NR, Gray J, Sorich MJ, Salter AB, Karnon J. Adjusting for treatment switching in oncology trials: a systematic review and recommendations for reporting. Value Health. 2020;23(3):388–96.
12. Ishak KJ, Proskorovsky I, Korytowsky B, Sandin R, Faivre S, Valle J. Methods for adjusting for bias due to crossover in oncology trials. PharmacoEconomics. 2014;32(6):533–46.
13. Zhu X, Zhou X, Zhang Y, Sun X, Liu H, Zhang Y. Reporting and methodological quality of survival analysis in articles published in Chinese oncology journals. Med (Baltim). 2017;96(50):e9204.
14. Abraira 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(12):1340–e65.
15. Altman DG, De Stavola BL, Love SB, Stepniewska KA. Review of survival analyses published in cancer journals. Br J Cancer. 1995;72(2):511–8.
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(5):e0154870.
17. 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(22):3721–6.
18. Salika T, Turner RM, Fisher D, Tierney JF, White IR. Implications of analysing time-to-event outcomes as binary in meta-analysis: empirical evidence from the Cochrane database of systematic reviews. BMC Med Res Methodol. 2022;22(1):73.
19. Carling CLL, Kristoffersen DT, Montori VM, Herrin J, Schünemann HJ, Treweek S, et al. The effect of alternative summary statistics for communicating risk reduction on decisions about taking statins: a randomized trial. PLoS Med. 2009;6(8):e1000134.
20. Skoetz N, Goldkuhle M, Weigl A, Dwan K, Labonté V, Dahm P, et al. Methodological review showed correct absolute effect size estimates for time-to-event outcomes in less than one-third of cancer-related systematic reviews. J Clin Epidemiol. 2019;108:1–9.
21. Guyatt G, Oxman AD, Akl EA, Kunz R, Vist G, Brozek J, et al. GRADE guidelines: 1. Introduction—GRADE evidence profiles and summary of findings tables. J Clin Epidemiol. 2011;64(4):383–94.
22. Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ et al. Cochrane Handbook for Systematic Reviews of Interventions: Cochrane; 2020. www.training.cochrane.org/handbook
23. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. Syst Reviews. 2021;10(1):89.
24. Skoetz N, Goldkuhle M, van Dalen EC, Akl EA, Trivella M, Mustafa RA, et al. GRADE guidelines 27: how to calculate absolute effects for time-to-event outcomes in summary of findings tables and evidence profiles. J Clin Epidemiol. 2020;118:124–31.
25. Murad MH, Wang Z. Guidelines for reporting meta-epidemiological methodology research. BMJ Evidence-based Med. 2017;22(4):139–42.
26. Akl EA, Kahale LA, Agarwal A, Al-Matari N, Ebrahim S, Alexander PE, et al. Impact of missing participant data for dichotomous outcomes on pooled effect estimates in systematic reviews: a protocol for a methodological study. Syst Reviews. 2014;3(1):137.
27. Goldkuhle M, Narayan VM, Weigl A, Dahm P, Skoetz N. A systematic assessment of Cochrane reviews and systematic reviews published in high-impact medical journals related to cancer. BMJ Open. 2018;8(3):e020869.
28. Parmar MKB, Torri V, Stewart L. Extracting summary statistics to perform meta-analyses of the published literature for survival endpoints. Stat Med. 1998;17(24):2815–34.
29. Murad MH, Wang Z, Zhu Y, Saadi S, Chu H, Lin L. Methods for deriving risk difference (absolute risk reduction) from a meta-analysis. BMJ. 2023;381:e073141.
30. DerSimonian R, Laird N. Meta-analysis in clinical trials. Control Clin Trials. 1986;7(3):177–88.
31. Veroniki A, Jackson D, Viechtbauer W, Bender R, Knapp G, Kuss O et al. Recommendations for quantifying the uncertainty in the summary intervention effect and estimating the between-study heterogeneity variance in random-effects meta-analysis. Cochrane Database Syst Reviews. 2015;10(Suppl. 1):25–7.
32. Page MJ, Shamseer L, Altman DG, Tetzlaff J, Sampson M, Tricco AC, et al. Epidemiology and reporting characteristics of systematic reviews of biomedical research: a cross-sectional study. PLoS Med. 2016;13(5):e1002028.
33. Goldkuhle M, Hirsch C, Iannizzi C, Bora AM, Bender R, van Dalen EC, et al. Meta-epidemiological review identified variable reporting and handling of time-to-event analyses in publications of trials included in meta-analyses of systematic reviews. J Clin Epidemiol. 2023;159:174–89.
34. Goldkuhle M, Kreuzberger N, Bender R, Bora A, Burdett S, Hirsch C et al. Transparent reporting of meta-analyses of time-to-event outcomes based on aggregate data from randomized trials of interventions (META-TTE reporting guideline). 2023 22.08.2024 [cited 2024 22.08.2024]. https://osf.io/j5bmw
35. Saluja R, Cheng S, Delos Santos KA, Chan KKW. Estimating hazard ratios from published Kaplan-Meier survival curves: a methods validation study. Res Synthesis Methods. 2019;10(3):465–75.
36. Wei Y, Royston P, Tierney JF, Parmar MK. Meta-analysis of time-to-event outcomes from randomized trials using restricted mean survival time: application to individual participant data. Stat Med. 2015;34(21):2881–98.

Publisher's note

6.2. Characteristics, methods and reporting of trials included in meta-analyses of time-to-event outcomes *(Paper 2)*



Meta-epidemiological review identified variable reporting and handling of time-to-event analyses in publications of trials included in meta-analyses of systematic reviews (*Paper 2*)

? Characteristics, reporting and methods of time-to-event analyses in trial publications included in meta-analyses of time-to-event outcomes

Meta-epidemiological study (*systematic search* (*02/2017-08/2020*))

- 25 Cochrane reviews and 25 non-Cochrane reviews (Core Clinical Journals)
- All randomized trials in pairwise, hazard ratio (HR)-based meta-analyses
- Primary review outcomes and overall survival/all-cause mortality
- Duplicate extraction of review, trial and trial outcome data

235 trials included in meta-analyses → 315 individual trial time-to-event analyses

- Overall survival: 64% (201/315 trial analyses)
- Progression-free survival: 17% (52/315 trial analyses)

Trial methods and reporting

Outcomes	• Outcome definition reported	**61% (192/315 trial analyses)**
	• Censoring reasons reported	**41% (130/315 trial analyses)**
	• Follow-up start specified	**56% (175/315 trial analyses)**

	Available for trial analyses (n= 315)	**Used in review (n=315)**
• Available time-to-event data per trial	• Survival curves: 83% (263) • Log-rank p-values: 76% (240) • HR: 72% (226) • Time-point specific survival: 46% (145)	• HR: 5% (15) • P-values: 5% (15) • Other: 8% (25) • Not reported: 83% (260)
• Analyses types	• Intention-to-treat: 70% (220) • Per protocol: 8% (25) • Modified intention-to-treat: 5% (15) • As treated: 2% (7) • Not reported: 23% (73)	• Intention-to-treat: 69% (216) • Modified intention-to-treat: 5% (16) • Other: 3% (8) • Not reported: 24% (75)
• Adjusted analyses available	• Adjusted: 27% (86) • Stratified: 20% (62) • Unadjusted: 17% (54) • Not reported: 22% (70) • Not applicable (No HR): 28% (87)	• Unadjusted: 25% (80) • Stratified: 18% (56) • Adjusted: 13% (41) • Not reported: 44% (138)

Trial characteristics relevant for time-to-event meta-analysis (n=235)	**Handling**
Missing outcome data	**Reported: 40% (95)** (e.g., excluded from analysis: 18%, censored: 11%, …)
Proportional hazards	**Tested: 14% (21/145 applicable)** Non-proportional 1% (3) vs. proportional 1% (2) vs. not reported 7% (16)
Competing events	**Reported: 11% (7/66 applicable)** (e.g., cumulative incidence curve: 2%, …)
Treatment switching	**Reported: 1% (3/222 applicable)** (e.g., sensitivity analysis: 1%, …)
Informative censoring	**Reported: <1% (1)** (sensitivity analysis – results not shown)

- **Variable and partially deficient reporting of trials included in meta-analyses of time-to-event outcomes in systematic reviews**
- **Reporting deficiencies from trials continue at the review level**

Goldkuhle M et al. Meta-epidemiological review identified variable reporting and handling of time-to-event analyses in publications of trials included in meta-analyses of systematic reviews. Journal of Clinical Epidemiology. 2023;159:174-189.

Figure 8: Graphical abstract for paper 2 (162).

6.2.1. Publication status

This article was published in July 2023 in the Journal of Clinical Epidemiology.

The work shares of the individual participants and their involvement in this paper are detailed in appendix 11.2.1.

Results of this article were presented at the 27th Cochrane Colloquium in London, UK, in 2023 (172) and at the 25th Congress of the German Network for Evidence-based Medicine in Berlin 2024 (173) (chapter 9.4).

6.2.2. Synopsis (162, 172, 173)

The second meta-epidemiological study assessed, as previously highlighted, the characteristics, methods and reporting of trials included in meta-analyses of time-to-event outcomes in current systematic reviews (162, 172, 173).

The study was also performed according to an a-priori registered protocol (osf.io/5qxbd). It assessed trials that were included in 50 systematic reviews (25 Cochrane reviews and 25 non-Cochrane reviews from Core Clinical Journals), randomly selected from the reviews included in the meta-epidemiolocal study on review level (chapter 6.1).
Data was extracted for trials included in meta-analyses of either the primary review outcome, as designated by the review authors, or the first mentioned time-to-event outcome in the review publication. In addition, data was extracted for meta-analyses of overall survival or all-cause mortality, when assessed as time-to-event outcomes, as these are generally considered the most relevant review outcomes.
The data extraction at trial level and for each individual trial time-to-event outcome included information on trial characteristics (e.g., publication dates, outcomes), trial time-to-event outcomes (e.g., definition, composition and presentation), numeric data (e.g., number randomized, number analyzed, loss to follow-up and competing events, follow-up), general trial methods (e.g., types of analyses, adjustment for covariates), time-to-event outcome-specific methods (e.g., time-to-event analyses), available time-to-event data (e.g., type of estimators, Kaplan-Meier or other survival plots and their presentation), results of time-to-event analyses and other study characteristics of particular relevance to time-to-event analyses (e.g., censoring, competing events, treatment switching, proportional hazards) as well as their consideration in trials (e.g., in additional analyses, such as sensitivity analyses).
Data for systematic reviews that included these trials could be extracted from the review level data extraction (chapter 6.1). This allowed comparisons between the trial level and the review level data. Data for this sub-project were extracted in duplicate by two project participants and the statistical analysis was performed descriptively (171).

The meta-epidemiological study included a total of 235 individual trials with 315 individual time-to-event analyses that were included in relevant meta-analyses of the 50 selected systematic reviews.
The assessment suggests serious limitations in the reporting of time-to-event analyses in trial publications. The study showed, for example, that outcome definitions, explanations of censoring mechanisms and starting points of follow-up data were available for only 61%, 41% and 56% of the assessed trial outcomes. General trial analysis specific information, such as the types of analyses and adjustment for each time-to-event analysis, was more frequently described in trial publications, 77% and 78%, than in their associated reviews. Review authors reported the respective information for their included trial estimates in 75% and 50% of the

time. If reported, review authors most frequently included intention-to-treat analyses and unadjusted analyses.
The types of time-to-event summary data that were available in trial publications for individual analyses varied significantly. Most commonly, trialists reported Kaplan-Meier plots, log-rank P values and HRs. Looking at which of these trial data the review authors included in their meta-analyses showed that information on methods used to include individual trial estimates in reviews was overall rarely reported. If reported, review authors most frequently used directly reported HRs or P values.
Except for missing outcome data, trial characteristics with particular relevance for the reliability of time-to-event outcomes, such as informative censoring, treatment switching and proportional hazards, were addressed only sporadically in the trial publications, for example, in additional analyses. The reporting quality of trials and their including reviews appears similar in this respect.

Previous methodological research has indicated that time-to-event analyses are poorly reported (4-8). Focusing on trials included in meta-analyses and taking into account the review level, this meta-epidemiological study now points to specific complications for review authors and emphasizes that that poor reporting of time-to-event analyses extends over several levels. The results suggest that the reporting of time-to-event outcomes in trial publications and the handling of possible limitations by review authors require improvement, for example, with reporting standards and targeted guidance (see chapter 7.4).

6.2.3. Full-text publication

The supplementary material accompanying this publication is provided in appendix 11.2.2.

ELSEVIER

Journal of Clinical Epidemiology 159 (2023) 174–189

Journal of Clinical Epidemiology

ORIGINAL ARTICLE

Meta-epidemiological review identified variable reporting and handling of time-to-event analyses in publications of trials included in meta-analyses of systematic reviews

Marius Goldkuhle[a,*], Caroline Hirsch[a], Claire Iannizzi[a], Ana-Mihaela Bora[a], Ralf Bender[b], Elvira C. van Dalen[c], Lars G. Hemkens[d,e,f,g], Marialene Trivella[h,i], Ina Monsef[a], Nina Kreuzberger[a,1], Nicole Skoetz[a,1]

[a]*Evidence-Based Medicine, Department I of Internal Medicine, Faculty of Medicine and University Hospital Cologne, University of Cologne, Kerpener Str. 62, 50937 Cologne, Germany*

[b]*Department of Medical Biometry, Institute for Quality and Efficiency in Health Care, Im Mediapark 8, D-50670 Cologne, Germany*

[c]*Princess Máxima Center for Pediatric Oncology, Heidelberglaan 25, 3584CS Utrecht, The Netherlands*

[d]*Research Center for Clinical Neuroimmunology and Neuroscience Basel (RC2NB), University Hospital Basel and University of Basel, Basel, Switzerland*

[e]*Department of Clinical Research, University Hospital Basel, University of Basel, Basel, Switzerland*

[f]*Meta-Research Innovation Center at Stanford (METRICS), Stanford University, Stanford, CA, USA*

[g]*Meta-Research Innovation Center Berlin (METRIC-B), Berlin Institute of Health, Berlin, Germany*

[h]*Division of Cardiovascular Medicine, John Radcliffe Hospital, Oxford OX3 9DU, UK*

[i]*Department of Population Health, London School of Hygiene and Tropical Medicine, London, WC1E 7HT, UK*

Accepted 25 May 2023; Published online 30 May 2023

Abstract

Objectives: Previous findings indicate limited reporting of systematic reviews with meta-analyses of time-to-event (TTE) outcomes. We assessed corresponding available information in trial publications included in such meta-analyses.

Study Design and Setting: We extracted data from all randomized trials in pairwise, hazard ratio (HR)–based meta-analyses of primary outcomes and overall survival of 50 systematic reviews systematically identified from the Cochrane Database and Core Clinical Journals. Data on methods and characteristics relevant for TTE analysis of reviews, trials, and outcomes were extracted.

Results: Meta-analyses included 235 trials with 315 trial analyses. Most prominently assessed was overall survival (91%). Definitions (61%), censoring reasons (41%), and follow-up specifications (56%) for trial outcomes were often missing. Available TTE data per trial were most frequently survival curves (83%), log-rank *P* values (76%), and HRs (72%). When trial TTE data recalculation was reported, reviews mostly specified HRs or *P* values (each 5%). Reviews primarily included intention-to-treat analyses (64%) and analyses not adjusted for covariates (25%). Except for missing outcome data, TTE-relevant trial characteristics, for example, informative censoring, treatment switching, and proportional hazards, were sporadically addressed in trial publications. Reporting limitations in trial publications translate to the review level.

Conclusion: TTE (meta)-analyses, in trial and review publications, need clear reporting standards.

Keywords: Systematic review; Meta-analysis; Randomized trials; Time-to-event outcomes; Survival analysis; Reporting quality

1. Introduction

Researchers interested in effects of interventions on longer-term outcomes or outcomes that occur in all participants at some point often employ time-to-event (TTE)

[1] Contributed equally.

* Corresponding author. Evidence-based Medicine, Department I of Internal Medicine, University Hospital Cologne, Kerpener Str. 62, 50637 Cologne, Germany. Tel.: +49-221-478-62032; fax: +49-221-478-96654.

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

https://doi.org/10.1016/j.jclinepi.2023.05.023

What is new?

Key findings

- We identified variable and often insufficient reporting of time-to-event outcomes and associated methods in publications of randomized trials included in aggregate data meta-analyses of current systematic reviews.
- Limited reporting included critical information such as outcome definitions, methods, and trial characteristics relevant for assessing the certainty of time-to-event analyses, for example, informative censoring and proportional hazards. Available time-to-event data varied substantially between trial publications.
- Limitations in trial reporting translate to review publications as well.

What this adds to what is known?

- Previous methodological research suggested shortcomings in the reporting of time-to-event outcomes and analyses in study publications. Focusing on trials included in meta-analyses, we showed that these limitations have relevance for meta-analyses in current systematic reviews.

What are the implications and what should be changed?

- Trial authors should strictly adhere to available reporting guidelines for time-to-event analyses in randomized trial publications. Reporting standards for meta-analyses of time-to-event outcomes based on aggregate data are urgently needed.

outcomes [1,2]. TTE analyses measure the occurrence of an event, for example, death, disease progression, or wound healing, together with the time until its occurrence and, for individuals without an observed event (censored observation), accounts for their time under observation. Survival plots and probabilities estimated by using the method by Kaplan and Meier [3], hazard ratios (HRs) estimated by using the Cox model and various statistical tests, most prominently the log-rank test, constitute the most frequently used methods for TTE analyses [4,5]. Meta-analyses of TTE outcomes from aggregate trial data are commonly performed based on the HR, which, for individual trials, can be included directly or derived from various data sources in trial publications [4,6,7].

Because TTE analyses are complex, authors of evidence syntheses depend on rigorous reporting in trial publications to determine the credibility of their meta-analyses. Trial HRs are frequently estimated by using Cox models which assume at least approximate proportionality of the hazards of compared groups (proportional hazards) over the observation time [5,8−10]. Missing outcome data, competing events, and treatment switching impose challenges on interpretation of the results, especially when they lead to naive censoring of trial participants [1,11−15]. Finally, information on more general analytical trial characteristics is particularly relevant for TTE outcome meta-analyses and their interpretation. Unfortunately, previous studies have indicated that reporting of trials including TTE outcomes is often deficient [16−21].

We explored the characteristics, methodology, and handling of TTE analyses of trials included in meta-analyses of current systematic reviews.

2. Methods

We report our assessment in accordance with an adaption of the Preferred Reporting Items for Systematic Reviews and Meta-Analyses checklist to meta-epidemiological research [22]. The project is registered under: osf.io/5qxbd.

2.1. *Eligibility criteria*

We assessed publications of trials that were included in systematic reviews with a meta-analysis based on aggregate data from a minimum of two randomized controlled trials that evaluated a health-related TTE outcome by means of the HR. We did not impose limitations regarding intervention types, medical fields, or settings, but reviews should have been available as full-text articles published in English. Network meta-analyses, previous versions of updated reviews, and co-publications of Cochrane reviews (CR) were excluded.

2.2. *Identification and selection of reviews and trials*

The reviews were part of a separate study on the handling of TTE outcomes in systematic reviews (Goldkuhle et al. 2023, *submitted*). Briefly, we randomly selected 25 CR from a sample of CR until August 2020 and 25 non-Cochrane reviews (nCR) from a corresponding sample published during the same time (February 28, 2017 to August 18, 2020). nCR were identified in a systematic search performed by an experienced information specialist (I.M.) (Appendix A1; February 8, 2021) and limited to reviews published in Core Clinical Journals, as defined by the US National Library of Medicine, to ensure relevance of included reviews [23].

We assessed primary review outcomes or, if not applicable, the first TTE outcome reported in the abstract. If a review included overall survival/all-cause mortality in a TTE outcome meta-analysis, we included this analysis as well, because it is often considered the most relevant outcome of a study. For feasibility, we excluded analyses with more than 20 trials.

For the selected review outcomes, we identified all trial publications from which TTE data were included in applicable meta-analyses: Systematic reviews often cite multiple publications of individual included trials. In such cases, we prioritized publications and outcome data as reported by review authors. Otherwise, we selected trial publications including an HR and confidence interval that corresponded to a review's forest plot HR, or could be inverted accordingly. If no corresponding trial HR was reported or if it differed from the review reported trial HR in any of the referenced trial publications, and other TTE data that were reported could not be directly matched, we selected the publication that corresponded in follow-up duration and number of participants if possible. Where a corresponding trial publication reported multiple sources of TTE data and it was unclear which was selected by review authors, we noted this information.

Selection took place in duplicate and independently (N.K. and M.G.). Potential discrepancies were resolved by consulting a third author (N.S.).

2.3. Data extraction and statistical analysis

All extractions were performed in duplicate by two authors (N.K., C.I., C.H., A.B., and M.G.) with involvement of a third author (N.S.) in case of potential discrepancies. The data extraction sheet was developed and piloted a priori (Appendix A2). Data were analyzed descriptively by means of absolute and relative frequencies for categorical data and medians, means, and variability measures for count data. To illustrate how review authors approached items associated with those extracted on trial level, we present data extracted for reviews in the tables along with applicable trial level results.

3. Results

3.1. Search results for reviews and their included trials

A flow diagram (Appendix A3) illustrates our search.

The identified reviews included 235 trials in their primary and overall survival TTE outcome analyses, resulting in 315 individual trial analyses of TTE outcomes included in review meta-analyses. For outcomes of 18 trials, we did not extract data because either TTE data were not available in cited publications, or it was unclear which data were included in the review, or publications were not accessible, or data were received from a secondary source (Appendix A4).

3.2. Characteristics of included reviews

Appendix Table A5 presents the characteristics of included reviews in detail. Most reviews were published in 2019, addressed questions on neoplasms, and compared biologics/drugs to biologics/drugs. Reviews included a median of four studies (interquartile range [IQR] 2.25−5) and 1,521 participants (571−4,580.5) in TTE outcome meta-analyses. They compared a median of five outcomes (IQR 4−8), among them a median of two (IQR 2−2) of TTE outcomes.

3.3. Characteristics of trials included in review time-to-event outcome meta-analyses

Most trials were published between 2011 and 2015 (Table 1). TTE data in reviews were predominately available in first full-text publications of trials and from trials addressing superiority. When multiple publications were cited and primary publications defined by the review authors (e.g., by an asterisk in the list of references for individual trials in Cochrane reviews), we most often located applicable TTE data in these. Overall, the original publications of TTE data were completely clear for 89% (279/315 trial outcomes) of trial outcomes in that the trial HRs in reviews corresponded to those in trial publications, or the source was explicitly reported by review authors, or only single publications were referenced.

The median population randomized per trial was 266 (IQR 120−620). For 44% (141/315 trial outcomes) of all trial outcomes, the analyzed population differed from the randomized population. If reported, the analyzed population differed by a median of 2.3% (IQR 0.8%−7.5%) and up to 63.3% of the randomized population. Trials analyzed a median of 2 (IQR 2−3) TTE outcomes. Most prominent TTE outcome per trial was overall survival or all-cause mortality (91%; 214/235 trials). Few trials assessed safety data with TTE methods.

3.4. Characteristics of trial outcomes included in this assessment

Outcome definitions were provided for 61% (192/315 trial outcomes) assessed trial outcomes (Table 2). Death as a competing event was possible in 11% (35/315 trial outcomes) of trial outcomes. Planned reasons for censoring of study participants were reported for less than half of trial outcomes. Reasons were most frequently last known time points of individuals being event-free and end of follow-up. Loss to follow-up, alternative treatments, and competing events were less often reported. Finally, a follow-up starting point was given for 56% (175/315 trial outcomes) of trial outcomes, which was most frequently randomization.

3.5. Time-to-event methodological characteristics of the trials included in review time-to-event outcome meta-analyses

The most frequently available TTE results (Appendix Fig. A6) for individual trial outcomes were HRs or log

Table 1. Characteristics of trials included in the reviews time-to-event outcome meta-analyses

	Trial			Review		
Domain	Overall (*N* = 235)	Cochrane (*n* = 102)	Non-Cochrane (*n* = 133)	Overall (*N* = 50)	Cochrane (*n* = 25)	Non-Cochrane (*n* = 25)
Publication						
Publication year						
≤2000	9% (19)	18% (18)	1% (1)	24% (12)	44% (11)	0% (1)
2001–2005	14% (32)	15% (15)	13% (17)	34% (17)	36% (9)	32% (8)
2006–2010	20% (46)	25% (26)	15% (20)	48% (24)	60% (15)	36% (9)
2011–2015	31% (74)	25% (25)	37% (49)	64% (32)	52% (13)	76% (19)
2016–2020	27% (64)	18% (18)	35% (46)	54% (27)	36% (9)	72% (18)
Publication format						
First full publication/NOS	84% (197)	76% (78)	89% (119)	100% (50)	100% (25)	100% (25)
Updated analysis	9% (20)	11% (11)	7% (9)	22% (11)	20% (5)	24% (6)
Abstract	4% (10)	8% (8)	2% (2)	14% (7)	20% (5)	8% (2)
Other (e.g., final analysis, letter)	3% (8)	5% (5)	3% (3)	16% (8)	20% (5)	12% (3)
Trial design						
Superiority/NOS	87% (204)	83% (85)	89% (119)	96% (48)	92% (23)	100% (25)
Noninferiority	11% (27)	13% (13)	11% (14)	26% (13)	28% (7)	24% (6)
Equivalency	1% (3)	3% (3)	0% (0)	4% (2)	8% (2)	0% (0)
Other (i.e., equivalency, combined analysis)	2% (4)	4% (4)	0% (0)	6% (3)	12% (3)	0% (0)
Data availability						
Multiple references	31% (73)	61% (62)	8% (11)	60% (30)	76% (19)	44% (11)
Data in primary publication[a]						
Yes	29% (67)	55% (56)	8% (11)	58% (29)	88% (22)	28% (7)
No	3% (8)	8% (8)	0% (0)	14% (7)	28% (7)	0% (0)
No primary publication defined by review authors	7% (17)	4% (4)	10% (13)	22% (11)	8% (2)	36% (9)
Single publication referenced	63% (147)	33% (34)	85% (113)	76% (38)	56% (14)	96% (24)
Origin of TTE data clear[a,b]						
Review HR is trial HR	61% (143)	43% (44)	74% (99)	80% (40)	64% (16)	96% (24)
Reported by review authors	16% (38)	23% (23)	11% (15)	20% (10)	28% (7)	12% (3)
Single data source in cited publication(s)	15% (35)	21% (21)	11% (14)	28% (14)	28% (7)	28% (7)
HR recalculated but source not reported	12% (28)	18% (18)	8% (10)	40% (20)	56% (14)	24% (6)
Trial population						
Sample size of randomized population						
Median (IQR)	266 (120–620)	219 (108–605)	310 (149–627)	1,531 (499–3,318)	593 (358–1,692)	1,935 (1,473–3,766)
Mean (range)	663 (20–17,160)	602 (20–8,113)	707 (40–17,160)	2,946 (83–31,703)	2,216 (83–10,988)	3,676 (349–31,703)
Not reported	6% (13)	10% (10)	2% (3)	18% (9)	24% (6)	12% (3)
Proportion of randomized participants not in analysis (%)[c]						
Median (IQR)	2.3 (0.8–7.5)	3.7 (1–10.9)	1.7 (0.5–4.8)			
Mean (range)	9.1 (0–63.3)	7.4 (0–60.9)	10.7 (0–63.3)			
Unclear/Not reported	10% (31)	14% (18)	7% (13)	36% (18)	32% (8)	40% (10)
All randomized analyzed	55% (174)	44% (58)	63% (116)	94% (47)	92% (23)	96% (24)
Outcomes in trial publication						
Number of TTE event outcomes						

(*Continued*)

Table 1. Continued

	Trial			Review		
Domain	Overall (N = 235)	Cochrane (n = 102)	Non-Cochrane (n = 133)	Overall (N = 50)	Cochrane (n = 25)	Non-Cochrane (n = 25)
Median (IQR)	2 (2–3)	2 (2–3)	2 (2–3)			
Mean (range)	3.27 (1–49)	2.22 (1–6)	4.07 (1–49)			
Not reported/Unclear	1% (3)	2% (2)	1% (1)			
Assessed TTE outcomes						
ACM/OS	91% (214)	87% (89)	94% (125)			
Progression-free survival	37% (88)	16% (16)	54% (72)			
Disease-free survival	19% (44)	31% (32)	9% (12)			
Duration of response	8% (18)	0% (0)	14% (18)			
Time to progression	7% (16)	7% (7)	7% (9)			
Other[d]	349	82	267			
Safety data as TTE data	3% (8)	0% (0)	6% (8)	6% (3)	0% (0)	12% (3)

Abbreviations: ACM, all-cause mortality; HR, Hazard ratio; IQR, interquartile range; NOS, Not otherwise specified; OS, Overall survival; TTE, time-to-event.

[a] These data must be interpreted as "trials including at least one outcome fulfilling the respective item, for example, 29% of trials included at least one trial outcome for which data were available in the primary trial publication, as indicated by the review authors.

[b] Refers to whether the origin of time-to-event data for an extracted trial was completely clear and, if so, how. The origin was clear if the forest plot HR and confidence interval in a review publication for an individual trial outcome corresponded to an HR and confidence interval reported for that outcome in a respective trial publication, if the review authors explicitly reported the source of time-to-event data for that trial outcome [e.g., in case of data recalculation] or if only a single source of time-to-event data for a trial outcome was available in any trial publication cited in a review.

[c] These data are presented per trial outcome: N = 315 [CR: n = 131; nCR: n = 184].

[d] Other included, for example, cardiovascular death, event-free survival, relapse/recurrence-free survival, and myocardial infarction.

(HR)s, log-rank P values, and survival curves, in combination with time point–specific survival probabilities, median survival times, or both. Differences in available TTE data types existed between outcomes of overall survival/all-cause mortality, composite outcomes including death from any cause, and outcomes not including death from any cause (Appendix A7). Other data such as cumulative incidence rates and observed-expected events were given rarely. When reported, HRs were primarily calculated with Cox models and sporadically from log-rank results or, for example, Rank Preserving Structural Failure Time or Fine and Gray models (Table 3; Appendix A7).

The included reviews only scarcely reported the used sources of TTE data for an individual trial outcome; if done, most often it was recalculation from HRs or P values, with information such as events per trial arm.

For 79% (249/315 trial outcomes) of outcomes, trials provided Kaplan-Meier curves, occasionally with the censored individuals throughout follow-up and the individuals at risk over time. Sporadically reported were cumulative incidence curves and adjusted Kaplan-Meier curves. If assessable, we perceived censoring as balanced, regarding distribution over time and proportions, in 80% (96/120 trial outcomes) of applicable trial outcomes.

3.6. General methodological characteristics of the trials included in review time-to-event outcome meta-analyses

Analysis types available in trial publications for individual trial outcomes (Table 4; Appendix A8) were most often intention-to-treat (ITT) analyses alone and most analyses that were reported as ITT analyses were performed in the complete allocated population.

Trial outcome analyses that were included in meta-analyses of the reviews were mostly ITT analyses as referred to by the trial conductors (69%; 216/315 trial outcomes). Overall, more than half of trial outcome analyses that were included in meta-analyses were clearly performed in the complete allocated trial population, and in 88% (276/315) of analyses, participants were analyzed in their allocated arm.

If adjustment or stratification of trial outcome HRs was reported, the most frequently available combinations in trial publications for individual outcomes were a single HR that was adjusted for baseline characteristics (12%; 39/315 trial outcomes). An available HR was reported as unadjusted in 24% (54/228 trial outcomes) and as adjusted for 38% (86/228 trial outcomes) of trial outcomes. Yet, frequently the adjustment status of available HRs was not reported.

Trial outcome analyses that were included in meta-analyses were mostly unadjusted (45%; 80/177 trial outcomes) and 23% were adjusted (41/177 trial outcomes). For 44% (138/315 trial outcomes), the adjustment status could not be determined.

3.7. Trial results and results included in review time-to-event outcome meta-analyses

The relative effect of HRs from trials included in TTE outcome meta-analyses, as reported, for example, in forest

Table 2. Characteristics of time-to-event outcomes defined and analyzed in the included trials

	Trial outcome			Review			
Domain	**Overall (*N* = 315)**	**Cochrane (*n* = 131)**	**Non-Cochrane (*n* = 184)**	**Overall (*N* = 50)**	**Cochrane (*n* = 25)**	**Non-Cochrane (*n* = 25)**	**Handling in review**
Trial outcomes included in this assessment (Primary and overall survival/all-cause mortality review outcomes)							
ACM/OS	64% (201)	67% (88)	61% (113)	88% (44)	84% (21)	92% (23)	
Progression-free survival	17% (52)	8% (11)	22% (41)	20% (10)	12% (3)	28% (7)	
Disease-free survival	6% (20)	13% (17)	2% (3)	8% (4)	12% (3)	4% (1)	
Local control	3% (10)	4% (5)	3% (5)	4% (2)	4% (1)	4% (1)	
Stent failure	3% (8)	0% (0)	4% (8)	2% (1)	0% (0)	4% (1)	
Other[a]	6% (20)	7% (10)	5% (10)	24% (12)	32% (8)	16% (4)	
Primary trial outcome	42% (132)	37% (48)	46% (84)	76% (38)	76% (19)	76% (19)	
Outcome definition provided	61% (192)	59% (77)	63% (115)				Heterogenous definitions mentioned - 6% (3/50) in discussion - 2% (1/50) in results
Composite outcome							
Yes	26% (83)	19% (25)	32% (58)				
No	70% (221)	76% (99)	66% (122)				
Unclear	3% (11)	5% (7)	2% (4)				
Outcome composites defined	92% (76)	92% (23)	91% (53)				
Outcome composites consistent							
Yes	42% (132)	52% (68)	35% (64)	62% (31)	64% (16)	60% (15)	
No	5% (15)	5% (6)	5% (9)	16% (8)	16% (4)	16% (4)	
Unclear	1% (3)	2% (3)	0% (0)	6% (3)	12% (3)	0% (0)	
Not applicable	52% (165)	41% (54)	60% (111)	90% (45)	80% (20)	100% (25)	
Death as competing event possible							
Yes	11% (35)	11% (14)	11% (21)	22% (11)	20% (5)	24% (6)	
No	84% (264)	82% (108)	85% (156)	92% (46)	88% (22)	96% (24)	
Unclear	5% (16)	7% (9)	4% (7)	26% (13)	28% (7)	24% (6)	
Reasons for censoring provided							
Yes	41% (130)	33% (43)	47% (87)	74% (37)	68% (17)	80% (20)	
Unclear/Not reported	59% (185)	67% (88)	53% (97)	96% (48)	96% (24)	96% (24)	
Reasons for censoring							
Participant last known event-free	23% (72)	16% (21)	28% (51)	44% (22)	36% (9)	52% (13)	
End of follow-up	15% (47)	14% (18)	16% (29)	50% (25)	40% (10)	60% (15)	
Loss to follow-up	9% (28)	8% (11)	9% (17)	26% (13)	24% (6)	28% (7)	
Other[b]	6% (18)	2% (3)	8% (15)	28% (1)	12% (3)	40% (10)	
Unclear	1% (2)	0% (0)	1% (2)	2% (1)	0% (0)	4% (1)	
Follow-up start reported							Follow-up start included in any outcome definition - 38% (19/50) Randomization - 4% (2/50) Allocated treatment

(*Continued*)

Table 2. Continued

	Trial outcome			Review			
Domain	**Overall (N = 315)**	**Cochrane (n = 131)**	**Non-Cochrane (n = 184)**	**Overall (N = 50)**	**Cochrane (n = 25)**	**Non-Cochrane (n = 25)**	**Handling in review**
							- 2% (1/50) Enrollment
Yes	56% (175)	53% (70)	57% (105)	80% (40)	68% (17)	92% (23)	
No	34% (108)	35% (46)	34% (62)	82% (41)	72% (18)	92% (23)	
Unclear	0% (1)	1% (1)	0% (0)	2% (1)	4% (1)	0% (0)	
Not applicable	10% (31)	11% (14)	9% (17)	40% (20)	40% (10)	40% (10)	
Follow-up start							
Randomization	43% (135)	42% (55)	43% (80)	72% (36)	60% (15)	84% (21)	
Allocated treatment	7% (22)	5% (6)	9% (16)	14% (7)	4% (1)	24% (6)	
Other (e.g., enrollment, previous treatment)	6% (18)	8% (10)	5% (9)	24% (12)	24% (6)	24% (6)	
Not applicable	44% (139)	46% (60)	43% (79)	88% (44)	80% (20)	96% (24)	

Abbreviations: ACM, all-cause mortality; IQR, Interquartile range; MACE, Major adverse cardiac events; MI, Myocardial infarction; OS, overall survival; TIMI, Thrombolysis in myocardial infarction.

[a] Other included event-free survival, time to wound healing, event-free survival, major adverse cardiac events [MACE], thrombolysis in myocardial infarction [TIMI] major bleeding, ''composite of all-cause death, myocardial infarction or stroke'', time to recurrence, biochemical relapse-free survival, and time to death from prostate cancer.

[b] Other included alternative treatment, competing events, absence of postbaseline information, participant withdrawal or withdrawal of consent, and inadequate outcome assessment.

plots and including recalculated HRs, was predominately favoring the intervention that was indicated by the review authors (Appendix A9). As judged by 95% confidence intervals, 26% (81/315 trial outcomes) of trial analyses were statistically significantly favoring the review authors' defined intervention. Results differed between CR and nCR and between outcomes of overall survival/all-cause mortality, composite outcomes including death from any cause, and outcomes not including death from any cause (Appendix A9). HRs which were directly reported in trial publications by trial authors showed a similar distribution.

Where an HR was directly reported in a trial publication, an HR $<$ 1 most often indicated a decreased risk of the event in the intervention group (86%; 179/208 trial outcomes) and it was predominantly calculated based on the rate of events in each group (91%; 190/208 trial outcomes) in difference to the rate of participants not experiencing the event (absence of event).

HRs reported in the trial publications were directly applicable to trial HRs in meta-analyses for 51% (120/315 trial outcomes) of trial outcomes or had to be inverted in 7% (23/315 trial outcomes). In several cases, an available HR or its confidence interval differed from the HR in the meta-analysis, for example, reviews explicitly reported not to use a trial HR, or recalculated the confidence interval.

3.8. Specific trial characteristics with relevance for time-to-event analyses and interpretation

A measure of trial follow-up was available for 82% (191/235 trials) of trials (Table 5; Appendix A10). Respective measures were most frequently reported as single measure across trial outcomes and only seldomly reported specifically for an individual trial outcome. Follow-up was predominately reported as median follow-up across outcomes, and although seldomly reported, calculated as median survival including surviving/event-free individuals only.

Missing outcome data were reported per trial arm for 57% (134/235 trials) of trials. About a third reported no information at all. The remaining either reported information across arms or for individual outcomes. If reported, median missing outcome data per trial arm was most frequently less than 5% of the allocated population, in several cases, however, also substantially higher (Appendix A10). Outcome-specific missing outcome data were reported in few trials. Handling of missing outcome data consisted most frequently of entirely excluding or censoring respective individuals from the analysis. Regarding handling of potential informative censoring, one trial reported a sensitivity analysis but did not show any results.

If death was as a potential competing event for an assessed trial outcome, only few trials reported the number of these potential competing events per arm (proportions in Appendix A10). In response, trial authors presented survival time distributions as cumulative incidence curves in 21% (seven/33 trials) of applicable trials. In two of these trials, authors used Fine and Gray regression to calculate HRs as well.

About a third of trials reported information regarding receipt of the comparator treatment in the intervention group or vice versa (treatment switching). Rates were most frequently less than 10% of the allocated population; in some cases, however, they exceeded 20% and 50%

Table 3. Time-to-event specific methodological characteristics of trials included in the reviews time-to-event outcome meta-analyses

	Trial outcome				Review			
Domain	**Overall ($N = 315$)**	**ACM/OS ($n = 198$)**	**Combined, including ACM ($n = 77$)**	**Not including ACM ($n = 40$)**	**Overall ($N = 50$)**	**Cochrane ($n = 25$)**	**Non-Cochrane ($n = 25$)**	**Handling in review**
Time-to-event data available for trial outcomes in trial publications								
Time-to-event data								Five most frequent methods for TTE data - 62% (31/50) HR and CI - 42% (21/50) Set of methods (e.g., Tierney 2008 (4)) - 22% (11/50) Survival curves - 20% (10/50) log (HR) and standard error - 8% (4/50) HR and other information
Survival curves	83% (263)	85% (168)	90% (69)	65% (26)	92% (46)	84% (21)	100% (25)	
P value (log-rank)	76% (240)	75% (148)	87% (67)	63% (25)	94% (47)	92% (23)	96% (24)	
HR or log (HR)	72% (226)	68% (135)	95% (73)	45% (18)	90% (45)	80% (20)	100% (25)	
Time-point specific survival (per arm)	46% (145)	48% (95)	49% (38)	30% (12)	82% (41)	76% (19)	88% (22)	
Median survival (per arm)	40% (125)	39% (78)	51% (39)	20% (8)	58% (29)	56% (14)	60% (15)	
Type of test unclear or not reported	6% (20)	6% (12)	5% (4)	10% (4)	26% (13)	20% (5)	32% (8)	
Other[a]	10% (33)	11% (22)	4% (3)	18% (8)	46% (23)	60% (15)	32% (8)	
HR calculation								HR included in meta-analyses - 88% (44/50) HR/log (HR) NOS - 6% (3/50) Other (HR/log HR from Cox model and HR/log (HR) from Cox model, log-rank test and Kaplan-Meier curve)
Cox model	60% (188)	57% (113)	75% (58)	43% (17)	86% (43)	72% (18)	100% (25)	
Other[b]	3% (9)	3% (6)	1% (1)	5% (2)	14% (7)	16% (4)	12% (3)	
Unclear/Not reported	11% (36)	10% (20)	19% (15)	2% (1)	42% (21)	40% (10)	44% (11)	
No HR calculated	26% (82)	30% (59)	4% (3)	50% (20)	54% (27)	64% (16)	44% (11)	
Survival plots for trial outcomes in trial publications								
Survival plots								
Kaplan-Meier	79% (249)	81% (161)	88% (68)	50% (20)	92% (46)	84% (21)	100% (25)	
Other[c]	4% (14)	3% (6)	1% (1)	16% (7)	14% (7)	16% (4)	12% (3)	
No, no graphs were presented	17% (52)	16% (31)	10% (8)	33% (13)	60% (30)	64% (16)	56% (14)	
Number at risk reported								
Yes	58% (184)	55% (108)	78% (60)	40% (16)	88% (44)	76% (19)	100% (25)	

(*Continued*)

Table 3. Continued

	Trial outcome				Review			
Domain	**Overall (N = 315)**	**ACM/OS (n = 198)**	**Combined, including ACM (n = 77)**	**Not including ACM (n = 40)**	**Overall (N = 50)**	**Cochrane (n = 25)**	**Non-Cochrane (n = 25)**	**Handling in review**
No	27% (86)	33% (65)	13% (10)	25% (11)	58% (29)	64% (16)	52% (13)	
Not applicable	14% (45)	13% (25)	9% (7)	33% (13)	56% (28)	56% (14)	56% (14)	
Censoring reported								Handling of nonadministrative censoring - 2% (1/50) Mentioned as bias criterion
Marked on plot	38% (119)	37% (74)	49% (38)	18% (7)	68% (34)	64% (16)	72% (18)	
On plot and with individuals at risk	3% (11)	3% (5)	8% (6)	0% (0)	14% (7)	12% (3)	16% (4)	
No	43% (136)	45% (90)	34% (26)	50% (20)	80% (40)	80% (20)	80% (20)	
Not applicable	16% (49)	15% (29)	9% (7)	33% (13)	62% (31)	68% (17)	56% (14)	
Censoring balanced								
Yes	30% (96)	31% (61)	40% (31)	10% (4)	66% (33)	64% (16)	68% (17)	
No	8% (24)	6% (12)	14% (11)	3% (1)	28% (14)	20% (5)	36% (9)	
Unclear	3% (9)	3% (5)	3% (2)	5% (2)	14% (7)	4% (1)	24% (6)	
Not applicable	59% (186)	61% (121)	43% (33)	80% (32)	88% (44)	92% (23)	84% (21)	
Data recalculation from trials reported in reviews for an individual trial outcome								
Data recalculation								
HR and other information (e.g., events)	5% (15)	4% (8)	1% (1)	15% (6)	4% (2)	4% (1)	4% (1)	
P value and other information (e.g., events)	5% (15)	6% (12)	3% (2)	3% (1)	8% (4)	12% (3)	4% (1)	
Other[d]	8% (25)	8% (16)	8% (6)	7% (3)	20% (10)	28% (7)	12% (3)	
Not reported	83% (260)	82% (162)	88% (68)	75% (30)	86% (43)	76% (19)	96% (24)	

Abbreviations: AAR, Absolute risk reduction; ACM, All-cause mortality; CI, Confidence interval; HR, Hazard ratio; O-E, Observed–expected; OS, overall survival; NOS, Not otherwise specified; RMST, Restricted mean survival time; RPSFT, Rank Preserving Structural Failure Time.

[a] Other includes median cumulative incidence [per arm], mean and standard deviation per arm, O-E events [log-rank] or hazard rates, or Wilcoxon-Gehan test.

[b] Other includes HR calculated from log rank tests, HR from Cox and RPSFT models, HR from Cox and time-dependent Cox models, Cox Markov model, and Cox and Fine and Gray models.

[c] Other includes cumulative incidence curves, adjusted Kaplan-Meier curves and unclear type of curves.

[d] Other includes HR and confidence intervals, individual participant data [recalculated or from publication], survival curves, and time-point specific survival times.

(Appendix A10). Six trials reported treatment switching as trial protocol specified, otherwise as anticipated, for example, sample size calculations, or explicitly excluded the option. If treatment switching was reported, the most prominent reason was related to the course of disease, for example, disease progression. Additional analyses to deal with treatment switching were reported only in single trials.

According to trial reporting, proportionality of hazards for outcomes analyzed as HRs was assessed by statistical tests, for example, log-log or Schoenfeld residuals, or by visual inspection of survival plots in 11% (19/166 trials) and 1% (two/166 trials) of trials. In only five trials, results of these assessments were reported, thrice as nonproportional and twice as reasonably proportional.

4. Discussion

4.1. Principal findings

The origin of included TTE data was determinable through our investigation in almost all reviews, but only rarely due to explicit reporting by review authors. Overall survival was the most commonly used TTE outcome in

Table 4. General methodological characteristics of trials included in the reviews time-to-event outcome meta-analyses

	Trial outcome				Review			
Domain	**Overall (N = 315)**	**ACM/OS (n = 198)**	**Combined, including ACM (n = 77)**	**Not including ACM (n = 40)**	**Overall (N = 50)**	**Cochrane (n = 25)**	**Non-Cochrane (n = 25)**	**Handling in review**
Trial outcome analyses available in trial publications								Eligible analyses reported - 42% (21/50) ITT Included analyses reported - 20% (10/50) ITT - 6% (3/50) ‘‘Analysis not reported in trials’’ - 18% (9/50) ‘‘Not reported for all trials’’
Available analyses								
ITT	70% (220)	68% (135)	79% (61)	60% (24)	96% (48)	96% (24)	96% (24)	
Per protocol	8% (25)	8% (16)	8% (6)	8% (3)	40% (20)	48% (12)	32% (8)	
mITT	5% (15)	5% (9)	8% (6)	3% (1)	20% (10)	8% (2)	32% (8)	
As treated	2% (7)	2% (3)	3% (2)	5% (2)	8% (4)	0% (0)	16% (4)	
Unclear/Not reported	23% (73)	25% (50)	13% (10)	33% (13)	60% (30)	72% (18)	48% (12)	
Trial outcome analyses included in review meta-analyses								
Included analysis								
ITT	69% (216)	67% (133)	79% (61)	55% (22)	96% (48)	96% (24)	96% (24)	
mITT	5% (16)	5% (9)	6% (5)	5% (2)	18% (9)	8% (2)	28% (7)	
Other (e.g., per protocol, as treated)	3% (8)	2% (5)	1% (1)	5% (2)	30% (15)	28% (7)	32% (8)	
Unclear/Not reported	25% (78)	27% (53)	13% (10)	34% (15)	62% (31)	72% (18)	52% (13)	
Analysis in complete population								
Yes	55% (174)	56% (110)	55% (42)	55% (22)	96% (48)	96% (24)	96% (24)	
No	32% (100)	31% (62)	31% (24)	35% (14)	70% (35)	80% (20)	60% (15)	
Unclear/Not reported	9% (27)	9% (17)	9% (7)	7% (3)	14% (7)	28% (7)	40% (10)	
Not applicable (e.g., subgroups only)	4% (14)	5% (9)	5% (4)	3% (1)	14% (7)	8% (2)	20% (5)	
Analysis in allocated arm								
Yes	88% (276)	88% (175)	91% (70)	78% (31)	98% (49)	96% (24)	100% (25)	
No	0% (1)	0% (0)	0% (0)	3% (1)	2% (1)	0% (0)	4% (1)	
Unclear/Not reported	10% (38)	11% (23)	8% (7)	15% (8)	44% (22)	60% (15)	28% (7)	
Adjusted, unadjusted, stratified analyses available in trial publications								Eligible covariate adjustments - 4% (2/50) Adjusted only

(*Continued*)

Table 4. Continued

	Trial outcome				Review			
Domain	Overall (*N* = 315)	ACM/OS (*n* = 198)	Combined, including ACM (*n* = 77)	Not including ACM (*n* = 40)	Overall (*N* = 50)	Cochrane (*n* = 25)	Non-Cochrane (*n* = 25)	Handling in review
								- 4% (2/50) Hierarchical (adjusted before unadjusted) - 4% (2/50) Unadjusted only - 2% (1/50) Both - 2% (1/50) Hierarchical (unadjusted before adjusted) - 6% (3/50) Unclear Stratified HRs eligible: 2% (1/50) Handling differently adjusted HRs - 2% (1/50) Only adjusted included, others likely excluded - 2% (1/50) Unadjusted recalculated - 8% (4/50) Unclear Included adjustment mentioned - 2% (1/50) In results
Adjusted	27% (86)	25% (50)	36% (28)	20% (8)	62% (31)	56% (14)	68% (17)	
Stratified	20% (62)	18% (36)	29% (22)	10% (4)	36% (18)	24% (6)	48% (12)	
Unadjusted	17% (54)	17% (34)	17% (13)	18% (7)	54% (27)	48% (12)	60% (15)	
Unclear/Not reported	22% (70)	21% (42)	34% (26)	5% (2)	54% (27)	40% (10)	68% (17)	
Not applicable (No HR reported)	28% (87)	31% (62)	5% (4)	53% (21)	54% (27)	64% (16)	44% (11)	
Adjusted, unadjusted, stratified analyses included in review meta-analyses								
Covariate adjustment of included analysis								
Unadjusted	25% (80)	28% (56)	16% (12)	30% (12)	60% (30)	68% (17)	52% (13)	
Stratified	18% (56)	16% (31)	27% (21)	10% (4)	34% (17)	20% (5)	48% (12)	
Adjusted	13% (41)	12% (24)	16% (12)	13% (5)	44% (22)	28% (7)	60% (15)	
Unclear/Not reported	44% (138)	44% (87)	42% (32)	48% (19)	76% (38)	72% (18)	80% (20)	

Abbreviations: ACM, all-cause mortality; HR, hazard ratio; ITT, intention-to-treat; mITT, modified intention-to-treat; OS, overall survival.

trials and reviews. Only around half of trials provided definitions for their assessed outcomes and few gave reasons for censoring. Available TTE summary data for individual trial outcomes consisted most frequently of a combination of HRs, log rank *P* values, survival curves, and either median or time point–specific survival times. Yet, data used for recalculation of summary data in reviews were only seldomly reported for individual trial outcomes.

Analyses included in reviews most frequently used individuals in their allocated trial arms, but only little more than half were clearly performed in the complete allocated trial populations. Trial effect measures included in reviews were mostly unadjusted for covariates, and information on adjustment of available HRs was often not reported in trial publications.

Numerical missing outcome data were available per trial arm for little more than half of trials and only rarely for individual outcomes. Trial conductors often handled it by excluding or censoring affected participants. For informative censoring, one trial indicated sensitivity analyses.

Table 5. Handling of specific trial characteristics with relevance for time-to-event outcomes in the trials included in the reviews time-to-event outcome meta-analyses

Domain	Trial ($N = 235$)	Review ($N = 50$)	Handling in review
Follow-up in trials			
Follow-up measure available			Foreseen follow-up time reported - 8% (4/50) Longest follow-up - 6% (3/50) Minimum duration of follow-up required - 4% (2/50) Maximum duration of follow-up specified
Follow-up reported across outcomes	79% (185)	96% (48)	
Follow-up reported for outcomes	3% (6)	10% (5)	
No follow-up measure reported	19% (44)	44% (22)	
Available follow-up measures			Handling varying follow-up reported - 10% (5/50) Sensitivity analyses (e.g., shorter/longer follow-up - 12% (6/50) Other (e.g., metaregression, study exclusion, risk of bias) - 2% (1/50) Unclear
Median	66% (154)	92% (46)	
Minimum	25% (59)	56% (28)	
Maximum	23% (53)	54% (27)	
IQR/lower and upper range of IQR	18% (43)	56% (28)	
Other, for example, mean, fixed time-point, standard deviation	12% (29)	28% (14)	
Follow-up calculation			Varying follow-up mentioned - 24% (12/50) In discussion - 8% (4/50) In results, - 6% (3/50) In results and in discussion
Median, surviving patients only	8% (19)	26% (13)	
Median, all patients	5% (11)	16% (8)	
Other[a]	5% (12)	18% (9)	
Unclear/Not reported	62% (146)	86% (43)	
Not applicable	20% (47)	46% (23)	
Reported missing outcome data in trials			
Reported per arm	57% (134)	98% (49)	
Reported per outcome			Handling missing data reported - 68% (34/50) Mentioned as risk of bias criterion in methods - 40% (20/50) Contact with authors - 8% (4/50) Sensitivity analyses (according to rate missing) - 4% (2/50) Single imputation
Yes	4% (9)	12% (6)	
Complete/no loss at trial level	11% (26)	28% (14)	
Complete/no loss at outcome level	3% (8)	12% (6)	
No	84% (198)	100% (50)	
Handling			Missing data mentioned - 56% (28/50) In results - 8% (4/50) In results and discussion
Excluded from analysis	18% (42)	42% (21)	
Censored	11% (26)	34% (17)	
Complete/no loss at trial level	11% (25)	28% (14)	

(*Continued*)

Table 5. Continued

Domain	Trial (N = 235)	Review (N = 50)	Handling in review
Single or multiple imputation	1% (2)	4% (2)	
Unclear/Not reported	59% (139)	92% (46)	
No missing data	3% (8)	12% (6)	
Censoring in trials			
Handling			
Sensitivity analysis (results not shown)	0% (1)	2% (1)	
Death as competing event in trials			
Handling reported			
Yes[b]	3% (7)	10% (5)	Handling of deaths as competing events not reported or discussed. No outcomes with death as competing event assessed: 66% (33/50) of reviews
No	25% (59)	54% (27)	
Not applicable	86% (202)	92% (46)	
Treatment switching in trials			
Prespecified			
Reported as not planned or allowed	4% (10)	10% (5)	Handling treatment switching reported - 2% (1/50) Mentioned as risk of bias criterion in methods - 2% (1/50) Presence reported for each trial - 2% (1/50) Sensitivity analysis (e.g., according to rate)
Reported as anticipated, for example, protocol, sample size	3% (8)	12% (6)	
Unclear/Not reported	93% (218)	94% (47)	
Not applicable	0% (1)	2% (1)	
Switching reasons			Treatment switching mentioned - 6% (3/50) In results - 4% (2/50) In discussion
Course of disease (e.g., disease progression)	12% (29)	32% (16)	
Participant (e.g., choose to switch)	9% (20)	20% (10)	
Other[c]	11% (26)	28% (14)	
Not reported	13% (30)	38% (19)	
Not applicable	64% (151)	88% (44)	
Handling reported			
Yes[d]	1% (3)	2% (1)	
No	93% (219)	98% (49)	
Not applicable	8% (19)	18% (9)	
Proportional hazards			
Assumption tested			Proportional hazards assessment not reported
Test (e.g., log-log, Schoenfeld residuals)	8% (19)	32% (16)	
Visual inspection of curves	1% (2)	4% (2)	
No	52% (124)	88% (44)	
Not applicable (e.g., no HR)	29% (69)	52% (26)	
Test results			Handling nonproportional hazards not reported

(*Continued*)

Table 5. Continued

Domain	Trial (N = 235)	Review (N = 50)	Handling in review
Nonproportional	1% (3)	6% (3)	
Reasonably proportional	1% (2)	4% (2)	
Not reported	6% (16)	26% (13)	
Not applicable	92% (216)	100% (50)	

Abbreviations: CI, confidence interval; IQR, Interquartile range; LTFU, Loss to follow-up; RPSFT, Rank preserving structural failure time.

[a] Other included, for example, the reverse Kaplan-Meier method, median follow-up excluding censored individuals and mean follow-up.

[b] Handling of included cumulative incidence curves alone or together with a Fine and Gray model.

[c] Other included, for example, administrative [e.g., interim analysis], precondition [e.g., allergies], intervention related [e.g., adverse events], and investigator/physician [e.g., physicians decision].

[d] Handling included, for example, rank preserving structural failure time models and sensitivity analyses, either treating crossovers as outcome events or excluding crossovers.

Treatment switching was reported for about a third of trials, with in some cases considerably high rates and most often due to the participants' disease. Proportional hazards were assessed in 10% of trials, but results of such assessments were even more seldomly reported.

4.2. Comparison to review level handling of time-to-event data

Like their included trials, definitions of TTE outcomes of interest were provided only in half of reviews. Relevant follow-up information was infrequently defined. Although reviews included predominately ITT analyses, eligible and included analysis types as well as details on adjustment of estimates were often not reported. Review methods to obtain TTE data varied substantially, most present were direct inclusion of the HR and complete sets of recalculation methods, and were only seldomly reported for an individual outcome. In reviews respectively, trial characteristics relevant to TTE analysis (e.g., variable follow-up, informative censoring, competing events, treatment switching, and proportional hazards) were sporadically included in additional assessments (e.g., sensitivity analyses or certainty assessments) and scarcely mentioned in review texts.

4.3. Strengths and limitations

We ensured robustness of our extraction results through a priori developed forms and duplicate performance of relevant steps. Nevertheless, we must acknowledge potential limitations: first, we used random sampling to generate a representative but manageable set of reviews. Second, we aimed to extend our exploration to reviews often considered the methodological gold standard and based our sample on a fixed number of CR. We ensured relevance of the included reviews through selecting nCR published in Core Clinical Journals. Third, the limited number of included systematic reviews led to imbalances between characteristics of CR and nCR. These appear, however, typical for comparisons of both and a comparison was not our primary intent [24,25]. Fourth, we imposed a restriction to primary and all-cause death including review outcomes which often constitute a subgroup of outcomes that is reported with greater rigor. Fifth, for feasibility, we limited our assessment to comparisons of 20 trials. But, because the total number of excluded reviews was small (2/74 [3%] CR and 22/401 [5%] nCR during full-text screening), we assume minimal impact.

4.4. Relation to other work

Previous studies support our findings of deficient reporting of TTE analysis-relevant information in trial publications, including the start and end points of observations, censoring and follow-up information, assumptions, such as proportional hazards in Cox models, and details on statistical modeling as well as numbers of events and censored observations [16−20,26]. Batson et al. [19] discuss implications of limited trial-level reporting to meta-analysis and particularly promote openness to alternative approaches when assumptions underlying the Cox model HR are in question. Our assessment focusses on trials that are included in TTE outcome meta-analyses and confirms their findings. In addition, we show that insufficient trial reporting is also transferred to review publications.

Kahale et al. [23] assessed the handling of missing outcome data in systematic reviews and trials included in meta-analyses of dichotomous data and found that the approach to missing outcome data was explained only in little more than a third of their assessed trials. Determining missing outcome data handling for TTE trial outcomes constitutes a particular hardship. Available reporting does not permit the distinction between loss to follow-up censoring and censoring for administrative causes (e.g., end of follow-up) so that trial participants with potential missing outcome data can be excluded without visibly reducing the analysis sample. We focused our extraction on explicitly reported handling of missing outcome data and assume that in many of the "not reported" cases, lost individuals were naively censored. Kahale et al. found that a minor proportion of their assessed reviews consistently approached missing outcome data in included trials in their analyses, which agrees with our findings.

4.5. Explanations, implications, and further research

Limited reporting in trial publications imposes complications for all who rely on reported information to evaluate the credibility of TTE outcome effects from trials, for example, for meta-analyses. With the recently published Consolidated Standards of Reporting Trials extension to trial outcomes, in addition to general Consolidated Standards of Reporting Trials guidance, some of the reporting issues we identified might improve, for example, appropriate outcome definitions and details on statistical methods, handling of missing outcome data, and specification of the analysis population [27]. Overall, trial authors should adhere to available reporting guidelines and suggestions, both for general outcome reporting as well as for TTE outcome-specific information, for example, for survival curves [17,18,27—29].

In response to current reporting limitations on trial level, review authors are encouraged to rigorously follow available guidance and to explicitly report deficiencies in trial publications they encounter [4,6,7]. Still, additional guidance and further research on the optimal translation of TTE-related trial issues to meta-analyses of aggregate data are needed.

5. Conclusion

The poor reporting of TTE outcomes and associated methods in trial publications limits not only the usefulness of these trials but also that of the systematic reviews and meta-analyses relying on them.

Funding

Funded by the German Research Foundation (DFG) (project SK 146/2-1; https://gepris.dfg.de/gepris/projekt/437030575).

CRediT authorship contribution statement

M.G. contributed to conceptualization, data curation, formal analysis, investigation, methodology, project administration, writing—original draft, and writing—review and editing; C.H. contributed to formal analysis, investigation, and writing—review and editing; C.I. contributed to formal analysis, investigation, and writing—review and editing; A.M.B. contributed to formal analysis, investigation, and writing—review and editing; R.B. contributed to conceptualization and writing—review and editing; E.v.D. contributed to conceptualization and writing—review and editing; L.G.H. contributed to conceptualization and writing—review and editing; I.M. contributed to systematic search; M.T. contributed to conceptualization and writing—review and editing; N.K. contributed to conceptualization, data curation, formal analysis, investigation, methodology, and writing—review and editing; N.S. contributed to conceptualization, methodology, supervision, and writing—review and editing.

Declaration of competing interest

One included non-Cochrane review was co-authored by a project participant (L.G.H.), who did not appraise data or resolve conflicts for these reviews. All other authors declare no relevant conflicts of interest.

Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.jclinepi.2023.05.023.

References

[1] Leung K-M, Elashoff RM, Afifi AA. Censoring issues in survival analysis. Annu Rev Publ Health 1997;18(1):83—104.
[2] Lagakos SW. General right censoring and its impact on the analysis of survival data. Biometrics 1979;35:139—56.
[3] Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. J Am Stat Assoc 1958;53:457—81.
[4] 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.
[5] Kleinbaum DG, Klein M. Survival analysis. 3 ed. New York, NY: Springer-Verlag; 2012.
[6] 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.
[7] Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, et al. Cochrane handbook for systematic reviews of interventions: Cochrane. 2020. Available at www.training.cochrane.org/handbook. Accessed March 29, 2023.
[8] Hernán MA. The hazards of hazard ratios. Epidemiology 2010;21(1):13—5.
[9] Stensrud MJ, Hernán MA. Why test for proportional hazards? JAMA 2020;323:1401—2.
[10] Rulli E, Ghilotti F, Biagioli E, Porcu L, Marabese M, D'Incalci M, et al. Assessment of proportional hazard assumption in aggregate data: a systematic review on statistical methodology in clinical trials using time-to-event endpoint. Br J Cancer 2018;119:1456—63.
[11] Austin PC, Fine JP. Accounting for competing risks in randomized controlled trials: a review and recommendations for improvement. Stat Med 2017;36:1203—9.
[12] Schumacher M, Ohneberg K, Beyersmann J. Competing risk bias was common in a prominent medical journal. J Clin Epidemiol 2016;80:135—6.
[13] Sullivan TR, Latimer NR, Gray J, Sorich MJ, Salter AB, Karnon J. Adjusting for treatment switching in oncology trials: a systematic review and recommendations for reporting. Value Health 2020;23(3):388—96.
[14] Ishak KJ, Proskorovsky I, Korytowsky B, Sandin R, Faivre S, Valle J. Methods for adjusting for bias due to crossover in oncology trials. Pharmacoeconomics 2014;32:533—46.
[15] Goldkuhle M, Bender R, Akl EA, van Dalen EC, Nevitt S, Mustafa RA, et al. GRADE Guidelines: 29. Rating the certainty in time-to-event outcomes-Study limitations due to censoring of participants with missing data in intervention studies. J Clin Epidemiol 2021;129:126—37.

[16] Zhu X, Zhou X, Zhang Y, Sun X, Liu H, Zhang Y. Reporting and methodological quality of survival analysis in articles published in Chinese oncology journals. Medicine 2017;96(50):e9204.

[17] Abraira 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, Stepniewska KA. Review of survival analyses published in cancer journals. Br J Cancer 1995;72:511−8.

[19] 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.

[20] 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.

[21] Salika T, Turner RM, Fisher D, Tierney JF, White IR. Implications of analysing time-to-event outcomes as binary in meta-analysis: empirical evidence from the Cochrane Database of Systematic Reviews. BMC Med Res Methodol 2022;22:73.

[22] Murad MH, Wang Z. Guidelines for reporting meta-epidemiological methodology research. BMJ Evidence-based Medicine 2017;22(4):139−42.

[23] Kahale LA, Khamis AM, Diab B, Chang Y, Lopes LC, Agarwal A, et al. Meta-analyses proved inconsistent in how missing data were handled across their included primary trials: a methodological survey. Clin Epidemiol 2020;12:527−35.

[24] Page MJ, Shamseer L, Altman DG, Tetzlaff J, Sampson M, Tricco AC, et al. Epidemiology and reporting characteristics of systematic reviews of biomedical research: a cross-sectional study. PLoS Med 2016;13(5):e1002028.

[25] Goldkuhle M, Narayan VM, Weigl A, Dahm P, Skoetz N. A systematic assessment of Cochrane reviews and systematic reviews published in high-impact medical journals related to cancer. BMJ Open 2018;8(3):e020869.

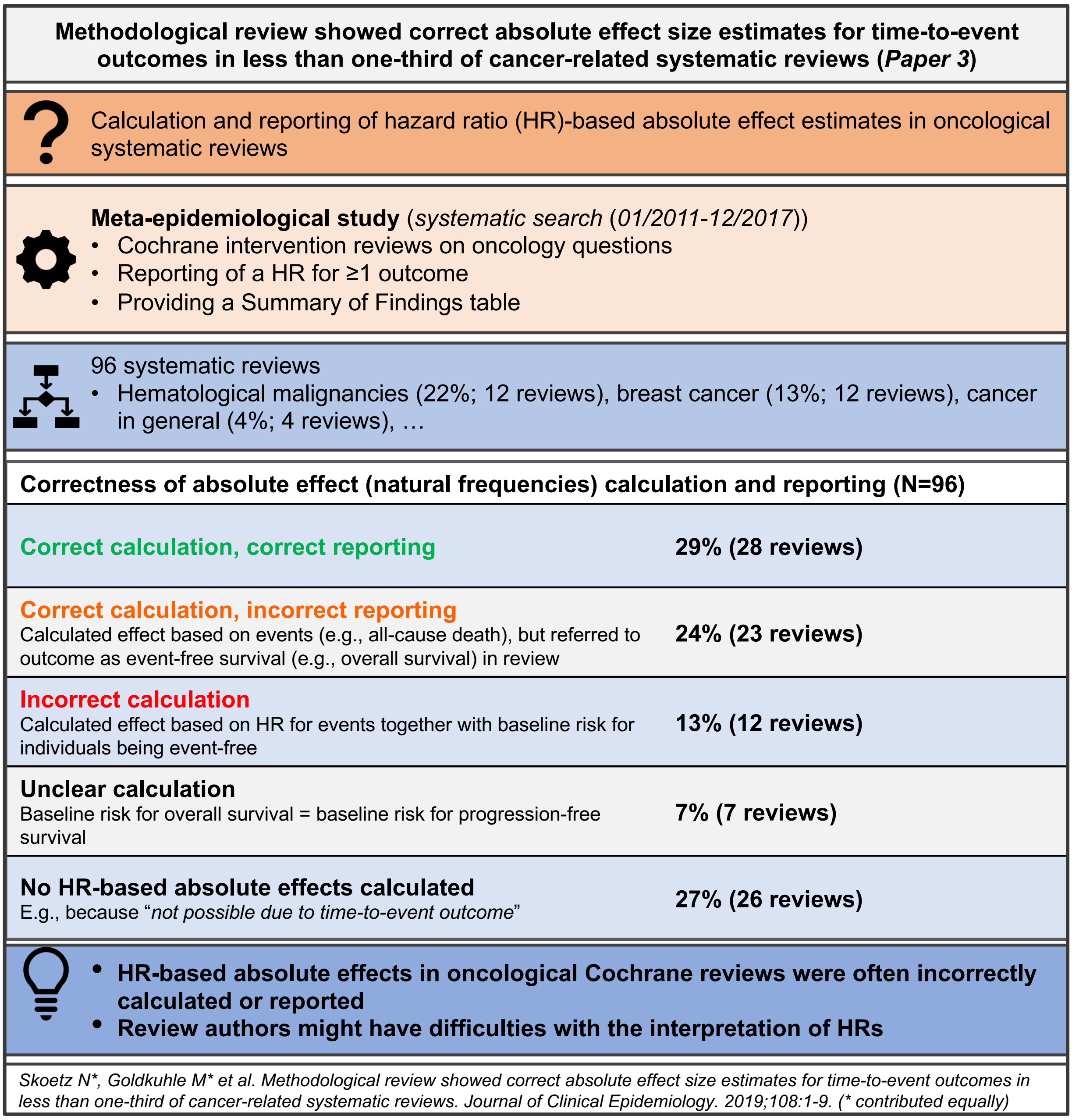
[26] Vervölgyi E, Kromp M, Skipka G, Bender R, Kaiser T. Reporting of loss to follow-up information in randomised controlled trials with time-to-event outcomes: a literature survey. BMC Med Res Methodol 2011;11:130.

[27] Butcher NJ, Monsour A, Mew EJ, Chan A-W, Moher D, Mayo-Wilson E, et al. Guidelines for reporting outcomes in trial reports: the CONSORT-outcomes 2022 extension. JAMA 2022;328:2252−64.

[28] Morris TP, Jarvis CI, Cragg W, Phillips PPJ, Choodari-Oskooei B, Sydes MR. Proposals on Kaplan-Meier plots in medical research and a survey of stakeholder views: KMunicate. BMJ Open 2019;9(9):e030215.

[29] Schulz KF, Altman DG, Moher D, CONSORT Group. CONSORT 2010 Statement: updated guidelines for reporting parallel group randomised trials. BMC Med 2010;8(1):18.

6.3. Presentation of results of meta-analyses of time-to-event outcomes in form of absolute effects in systematic reviews *(Paper 3)*



Methodological review showed correct absolute effect size estimates for time-to-event outcomes in less than one-third of cancer-related systematic reviews (*Paper 3*)

?

Calculation and reporting of hazard ratio (HR)-based absolute effect estimates in oncological systematic reviews

Meta-epidemiological study (*systematic search* (*01/2011-12/2017*))

- Cochrane intervention reviews on oncology questions
- Reporting of a HR for ≥1 outcome
- Providing a Summary of Findings table

96 systematic reviews

- Hematological malignancies (22%; 12 reviews), breast cancer (13%; 12 reviews), cancer in general (4%; 4 reviews), …

Correctness of absolute effect (natural frequencies) calculation and reporting (N=96)	
Correct calculation, correct reporting	**29% (28 reviews)**
Correct calculation, incorrect reporting Calculated effect based on events (e.g., all-cause death), but referred to outcome as event-free survival (e.g., overall survival) in review	**24% (23 reviews)**
Incorrect calculation Calculated effect based on HR for events together with baseline risk for individuals being event-free	**13% (12 reviews)**
Unclear calculation Baseline risk for overall survival = baseline risk for progression-free survival	**7% (7 reviews)**
No HR-based absolute effects calculated E.g., because "*not possible due to time-to-event outcome*"	**27% (26 reviews)**

- **HR-based absolute effects in oncological Cochrane reviews were often incorrectly calculated or reported**
- **Review authors might have difficulties with the interpretation of HRs**

Skoetz N, Goldkuhle M* et al. Methodological review showed correct absolute effect size estimates for time-to-event outcomes in less than one-third of cancer-related systematic reviews. Journal of Clinical Epidemiology. 2019;108:1-9. (* contributed equally)*

Figure 9: Graphical abstract for paper 3 (11).

6.3.1. Publication status

This article was published in April 2019 in the Journal of Clinical Epidemiology.

The work shares of the individual participants and their involvement in this paper are detailed in appendix 11.3.1.

This paper was awarded with Cochrane's Bill Silverman prize 2020 and is part of the consortium of articles awarded with the German Network for Evidence-based Medicine's David Sackett prize 2021 (chapter 9.3).

6.3.2. Synopsis (11)

As previously outlined, the third meta-epidemiological study that is part of the consortium of papers in this dissertation focused on the calculation of absolute effect estimates for the results of meta-analyses of time-to-event outcomes (11, 174).

The assessment was performed according to an a-priori registered protocol. It included a sample of Cochrane intervention reviews addressing oncological research questions and reporting a GRADE Summary of Findings (see chapter 3.3.4) with a pooled effect estimate for at least one time-to-event outcome. This sample was selected because oncological systematic reviews frequently address time-to-event outcomes and because Summary of Findings tables are mandatory for Cochrane reviews. Such tables require the presentation of absolute effect measures, commonly either natural frequencies or a risk difference, calculated based on the HR and a baseline/ population risk estimate.
A systematic search for the respective Cochrane reviews was conducted on the Cochrane Database of Systematic Reviews. The extracted data included, besides general review characteristics, particularly the calculation and reporting of absolute effect estimates.

Eligible systematic reviews must have presented absolute effect measures for the time-to-event outcomes overall survival and disease/progression-free survival for at least one review question. The correctness of the reported absolute effects was then assessed by comparing the two: overall survival is, by definition, included in the composite outcome disease/progression-free survival, which measures the time to occurrence/progression of disease or all-cause death. A comparison of the baseline risk estimates that review authors chose for the calculation of their absolute effects allows to determine whether these estimates represent the risk of an event in the population (overall survival rate < progression/ disease-free survival rate) or its absence (overall survival rate > progression/ disease-free survival rate). As previously explained (chapter 3.1.3 and 3.4.4), HRs are commonly calculated based on the event rates in the compared trial arms. Usually, a HR < 1 represents a lower hazard rate of the event of interest in the experimental arm. If a baseline risk that represents the absence of event (e.g., survival) is multiplied with a HR calculated based on events (e.g., mortality) to derive the absolute risk of the event in the experimental arm, this number will reverse the absolute effect and will falsely indicate that an inferior comparator group is beneficial. This phenomenon was used to determine if the review authors applied the correct calculation for their absolute effects in the selected reviews. Furthermore, it was assessed how the authors reported their outcomes (either as events, e.g., mortality, as absence of events, e.g., survival, or both, and whether they provided any reasoning for their decision), whether they reported additional information on the selection and source of applied baseline risks and whether the HR indeed was calculated based on rate of events.

From 483 potentially eligible Cochrane reviews, 96 reviews corresponded to the specified inclusion criteria and most of them addressed hematological malignancies or breast cancer interventions.
In the seldom cases in which the review authors specified the origin of their baseline risk in footnotes along the Summary of Findings tables, this information was not sufficient to determine whether the baseline risks were applicable to the rate of events or absence of event in the population.
Overall, in 29% of the review authors correctly calculated their absolute effects for time-to-event outcomes and a HR for events was multiplied with a baseline risk applicable to the risk of events or reverse. In 24% of reviews, absolute effects were correct, yet the reporting of

review authors was inconsistent when they shuffled between reporting their outcomes as events and the absence thereof without any reasoning.
Incorrect absolute effects were found in 13% of reviews, which occurred because authors falsely multiplied HRs for events and the baseline/ population rate of its absence or reverse. Finally, calculation was indeterminable in 7% of reviews, because the baseline rates were the same in both groups. In 27% of reviews no absolute effects for time-to-event outcomes were calculated, with the explicit reasoning that calculation was not performed because the outcomes were time-to-event outcomes.

Overall, this study highlighted serious flaws in the calculation of absolute effects of time-to-event outcomes in a considerable number of Cochrane systematic reviews, which are widely considered as the gold standard. The results imply that the interpretation of the HR is not straightforward, at least for some review authors. This clearly demonstrated that additional guidance on the calculation of absolute effects for time-to-event outcomes is needed, potentially also including alternative estimates to the commonly used (see paper 5).
A central, direct implication of the project was the adaptation of the GRADEpro Guideline Development Tool (GDT) software's (www.gradepro.org/) underlying statistical package. Previously, like several other meta-analysis programs, the software provided false absolute effect estimates when baseline risks for absence of events (e.g., survival) and HRs were used for calculation (175).

6.3.3. Full-text publication

The supplementary material accompanying this publication is provided in appendix 11.3.2.

Journal of Clinical Epidemiology 108 (2019) 1−9

Journal of Clinical Epidemiology

ORIGINAL ARTICLE

Methodological review showed correct absolute effect size estimates for time-to-event outcomes in less than one-third of cancer-related systematic reviews

Nicole Skoetz[a,*,1], Marius Goldkuhle[a,1], Aaron Weigl[a], Kerry Dwan[b], Valérie Labonté[c], Philipp Dahm[d], Joerg J. Meerpohl[e], Benjamin Djulbegovic[f], Elvira C. van Dalen[g]

[a]*Department I of Internal Medicine, University Hospital of Cologne, Kerpener Str. 62, Cologne 50937, Germany*

[b]*Editorial and Methods Department, Cochrane, St Alban's House, 57-59 Haymarket, London SW1Y 4QX, UK*

[c]*Institute for Evidence in Medicine (for Cochrane Germany Foundation), Medical Center, Faculty of Medicine, University of Freiburg, Breisacher Str. 153, Freiburg 79110, Germany*

[d]*Department of Urology, Minneapolis Veterans Administration Medical Center and University of Minnesota, Minneapolis VA Health Care System, Urology Section 112D, One Veterans Drive, Minneapolis, MN 55417, USA*

[e]*Institute for Evidence in Medicine (for Cochrane Germany Foundation), Medical Center, Faculty of Medicine, University of Freiburg, Breisacher Str. 153, Freiburg 79110, Germany*

[f]*Department of Supportive Medicine and Department of Hematology, City of Hope, 1500 Duarte Rd, Duarte, CA 91010, USA*

[g]*Department of Pediatric Oncology, Emma Children's Hospital, Amsterdam UMC, University of Amsterdam, PO Box 22660, Amsterdam 1100 DD, the Netherlands*

Accepted 5 December 2018; Published online 10 December 2018

Abstract

Objectives: To evaluate in how many cancer-related Cochrane reviews hazard ratio (HR)-based absolute effects in summary of findings (SoF) tables have been correctly calculated and reported.

Study Design and Setting: We identified all Cochrane cancer intervention reviews that reported an HR for at least one outcome and provided a SoF table, published between January 2011 and December 2017 in the Cochrane Database of Systematic Reviews.

Results: In 28 reviews (29%) of 96 included Cochrane reviews, absolute effects in the SoF tables were calculated in a correct manner. In 23 reviews (24%), absolute effects had been correctly calculated, but there was no explanation given why authors calculated event-free survival (e.g., overall survival) throughout the review but reported number of events in SoF tables (e.g., death). Twelve reviews (13%) provided incorrect absolute effects. For seven reviews (7%), it was unclear if absolute effects were correctly calculated. In 26 (27%) reviews, no absolute effects based on the given HR were calculated.

Conclusions: In less than one-third of cancer-related Cochrane reviews, absolute effect size estimates were correctly calculated and reported. There is a need for guidance on how to calculate and report absolute effect estimates based on HR data.

Keywords: Time-to-event; Hazard ratio; Absolute effects; Summary of findings; Methodological; Review

Conflict of interest statement: All authors have completed the ICMJE uniform disclosure form at www.icmje.org/coi_disclosure.pdf and declare: authors received travel funds from Cochrane only to present findings at one international meeting and one international conference for the submitted work; no financial relationships with any organizations that might have an interest in the submitted work in the previous 3 years; no other relationships or activities that could appear to have influenced the submitted work." NS, AW and MG are part of Cochrane Haematological Malignancies, EvD part of Cochrane Childhood Cancer, VL and JJM part of Cochrane Germany, and KD staff member of the Cochrane Editorial Unit.

Funding: Partial support from Cochrane was received to present data during the Cochrane Midyear Meeting 2017 in Geneva, Switzerland, and during the Global Evidence Summit 2017 in Cape Town, South Africa.

Authors' Contributions: N.S. initiated the project, designed the data collection tools and analysis, selected studies, extracted and analysed data, and drafted and revised the article. M.G. was involved in the design of the data collection tools and analysis, selected studies, extracted and analysed data, prepared tables for presentation, and revised the article. EvD designed the data collection tools and analysis, selected studies, extracted and analysed data, and drafted and revised the article. A.W. selected studies, extracted and analysed data, and revised the article. P.D., B.G., K.D., J.J.M., and V.L. extracted data and revised the article. All authors approved the final version of the article. All authors had full access to all of the data (including statistical reports and tables) in the study and can take responsibility for the integrity of the data and the accuracy of the data analysis.

[1] Contributed equally.

* Corresponding author. Tel.: +49221 478 96651; fax: +49221 478 96654.

E-mail address: nicole.skoetz@uk-koeln.de (N. Skoetz).

https://doi.org/10.1016/j.jclinepi.2018.12.006
0895-4356/© 2018 Elsevier Inc. All rights reserved.

What is new?

Key findings:

- We identified errors in the presentation of absolute effect measures in summary of findings (SoF) tables of cancer-related Cochrane reviews and describe common pitfalls to avoid.
- The errors in calculation of hazard ratios can be minimized if the review authors first assess direction of effect measure (event or nonevent) and then accordingly calculate the respective corresponding absolute effects. For example, the event that is typically measured is mortality (death), but the outcome reported is often overall survival (1- mortality), which goes into opposite direction.

What this adds to what is known?

- The appropriate presentation of absolute effect size estimates based on hazard ratios in SoF tables has not been evaluated previously.

What is the implication, what should change now?

- There is an urgent need for additional training materials and guidance for authors on how to calculate and present absolute effects based on time-to-event data.

1. Introduction

Absolute effect estimates are more understandable to patients, clinicians, and other users of systematic reviews than relative effect measures and are the recommended effect measure to communicate risks [1]. They reflect the clinical importance of outcomes and can ground exaggerated outcome perceptions of clinicians and patients, which may occur if solely relative effects are reported [2−4]. Absolute effects provide important supplementary information that considers risk-specific control event rates over a given time period. Absolute effect estimates are a routine part of the user-friendly format of ’summary of findings" (SoF) tables or evidence profiles [5]. Reviews published by Cochrane, which is widely known for establishing methodological standards for conducting and reporting high quality systematic reviews, regularly include such SoF tables. SoF tables are prepared according to the GRADE guidance papers and can be calculated using software products such as GRADEpro GDT (gradepro.org) or MAGICapp (app.magicapp.org) [5].

In many fields of health care, in particular oncology, analyses that assess the time to a given event for one or several groups of patients are commonly used. For patients with cancer, one of the most relevant outcomes is overall survival (OS). It describes the survival time of patients until death for any reason which occurs within a certain period of follow-up. In addition, another outcome like progression-free survival (PFS), that is the survival time without detectable worsening of disease (progress, relapse, death) over a considered time-period, is often assessed. This outcome measure provides complimentary information for OS. Both outcomes are so called time-to-event outcomes, as they involve the assessment of both whether a particular event occurs, and also when it occurs [6]. To compare time-to-event outcomes of two groups of patients, hazard ratios (HR) with corresponding confidence intervals that provide relative effect size estimates are used.

The calculation of absolute effects based on HR is error prone because both beneficial (event-free survival) and adverse effects (events) can easily be confused and because calculation of HR is based on difficult to interpret exponential functions. As there is currently no written guidance on how to calculate absolute effects based on HR, and how to best present these in SoF tables, it might be especially difficult for review authors to do this properly, as well as for journal editors and peer reviewers to identify mistakes. Another potential challenge arises around the consistent definition of time-to-event outcomes across all parts of the review including the abstract, results section, and the SoF table. Because of that absolute effects based on reported time-to-event outcomes are difficult to calculate, present, and interpret. Although often the event is measured (e.g., death), the event-free survival (e.g., OS) is reported throughout individual studies and the corresponding systematic review, review authors must be aware to calculate the respective absolute effect (for the event or for event-free survival). Until a recent update (September 2018), the GRADEpro GDT software allowed calculation of absolute effects based on HR only for outcomes and baseline risks corresponding to events (such as mortality) but not for event-free survival (like overall survival).

In this methodological review, we evaluated in how many current cancer-related Cochrane reviews absolute effects based on HR in SoF tables have been correctly calculated and reported.

2. Materials and methods

We report our methodological review according to the Preferred Reporting Items for Systematic Reviews and Meta-analyses (PRISMA) [7]. The project was conducted according to an a priori developed protocol. As this is a methodological review, it was not eligible for a registration in the International Prospective Register of Systematic Reviews (PROSPERO). The protocol can be accessed on request from the review authors.

2.1. Eligibility criteria

Cancer-related Cochrane intervention reviews and overviews of reviews were eligible for inclusion if they provided an SoF table in which the effect size for at least one time-to-event outcome based on a pooled HR was included. Reviews in which an HR for an outcome was given from a single study only were also eligible. Reviews that reported HR from several studies for the same outcome but did not pool the respective HR were excluded. This was to ensure that only one single HR per outcome in the review was reported, which could be used to calculate one corresponding absolute effect estimate. The results for this time-to-event outcome must have been mentioned in at least one of the following sections: abstract, methods, or results. Therapeutic, preventive, or prophylactic intervention reviews were eligible. Reviews not meeting all these criteria were excluded. We excluded reviews in which effects were presented in risk ratios or odds ratios (ORs) only. We used the original English version of each Cochrane review for data extraction and assessment.

2.2. *Study identification and selection*

We systematically identified all Cochrane intervention reviews that examined questions in the context of oncology, irrespective of type of cancer, stage of disease, type of intervention, outcomes assessed, or study design of included studies. This was done by using the function: "Browse by topic" and by choosing the following options: "Cancer" and "Stage: Review" in the Cochrane Database of Systematic Reviews. This function is based on tags, which are manually applied by the operators of the Cochrane Library.

The current version of SoF tables was first described in 2010, and the first GRADE guidelines were published in 2011 [5,8]. We did not expect any Cochrane Reviews to include SoF tables before 2011. Therefore, we restricted the included Cochrane Reviews to a 6-year period between January 2011 and December 2017. In case a review was published more than once during this time period, for example, as primary publication and as an update, we included only the most recent publication. Three authors (N.S., A.W., and M.G.) independently screened titles, abstracts, and full texts identified in the Cochrane Database of Systematic Reviews for eligibility according to the inclusion criteria. Review screening was carried out in one step because it is necessary to view a review full text to assess the availability of SoF tables and time-to-event outcomes. If any disagreement regarding the inclusion of reviews occurred, the authors tried to resolve it by discussion or involved another author (EvD) until consensus was achieved.

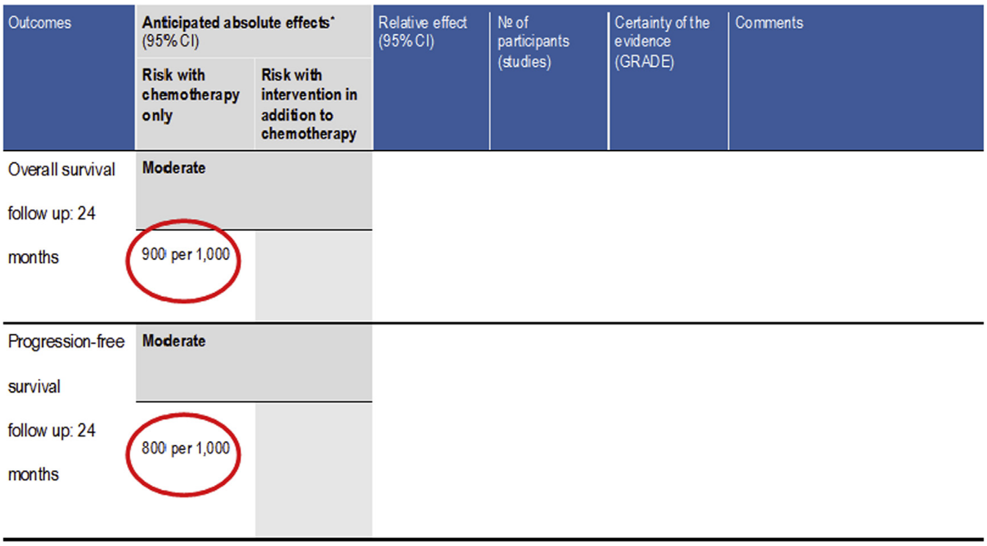
2.3. *Data extraction*

All included reviews were randomly allocated to eight members of the research team (N.S., M.G., P.D., A.W., V.L., J.J.M., K.D., EvD) to be extracted independently in duplicate. In case one of these individuals was involved in the publication of a specific Cochrane review (e.g., as an author or member of the editorial group), this particular review was reassigned to other, nonconflicted members of the research team. We used a dedicated pilot-tested extraction form. Any discrepancies during data extraction were resolved through discussion or if necessary with involvement of a third author.

To classify the baseline characteristics of the included SRs, we extracted information on the cancer type (e.g., breast, lung, colorectal), but also "cancer in general" and "mixed" (multiple diseases, but not cancer in general) and year of publication. To examine how absolute effects were calculated, we extracted data for the first two time-to-event outcomes, which were reported in a SoF table and the description for these outcomes with corresponding HR and their 95% confidence intervals as reported in abstract, methods section, and/or results section. For Cochrane reviews in which a (pooled) HR was given, we assumed that this effect measure and its associated confidence intervals had been correctly calculated. We extracted the first two HR outcomes because in cancer reviews these are commonly OS and PFS. Overall survival as an outcome measures includes only the single event "death", whereas PFS includes the events "death," "progression," and "relapse". Comparing the reported baseline risks for these two outcomes allowed to determine whether the baseline risks applied to events or event-free survival (as described in the next paragraphs). In addition, we extracted the description of the same outcomes as reported in the SoF table and the absolute effects as well as information regarding assumption of the underlying baseline risk. If several SoF tables were included, we used data from the first SoF table that listed an eligible time-to-event outcome.

For each outcome, we interpreted the meaning of an HR < 1, that is, whether this favored the control or intervention arm, based on the choice of the event as documented in the methods section of the review. If absolute effects were reported in the SoF table and review authors provided information on how the control group risk had been determined, we extracted this information. In case review authors did not provide information on how they determined the control group risk, we assessed whether they used the number of people with the event (e.g., people being dead at a specific time point) or the number of people event-free (e.g., people being alive at a specific time point) to calculate absolute effects for the intervention group. If at least two HR outcomes, like OS and PFS, were reported in the review, we compared the absolute numbers in the estimated control group risk for both outcomes, as shown in Figs. 1—3. If the absolute number for the outcome OS (based on the event people being dead) was lower than for the outcome PFS (based on the event people with progressive disease), we assumed authors had used number of people being event-free to calculate absolute effects for the intervention arm (see Fig. 1).

If the number was lower for OS than for PFS (see Fig. 2), we assumed authors had used number of people with the event to estimate the control group risk and calculate numbers for the intervention arm.



Outcomes	Anticipated absolute effects* (95% CI)		Relative effect (95% CI)	№ of participants (studies)	Certainty of the evidence (GRADE)	Comments
	Risk with chemotherapy only	Risk with intervention in addition to chemotherapy				
Overall survival follow up: 24 months	Moderate 900 per 1,000					
Progression-free survival follow up: 24 months	Moderate 800 per 1,000					

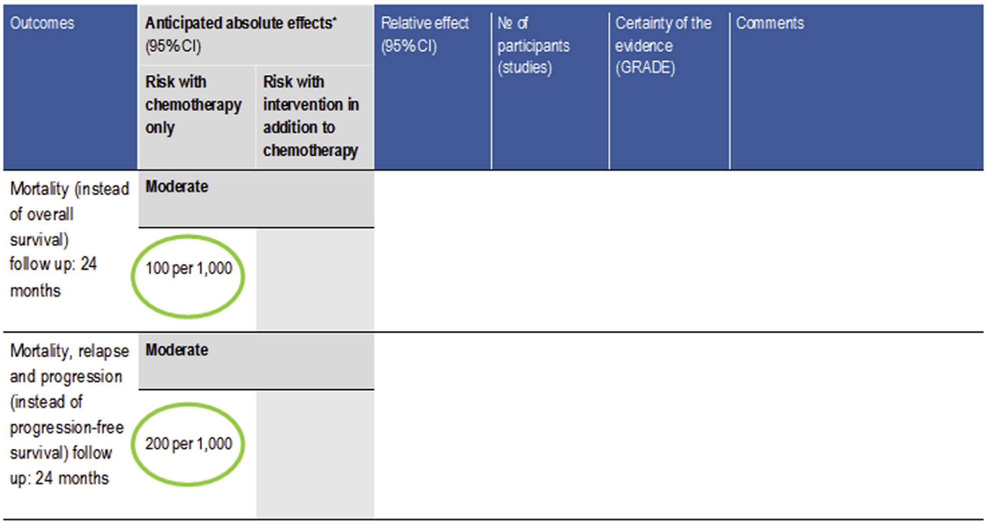
Fig. 1. Extract from an exemplar SoF table. Numbers for the estimated control group risks are marked in red (for the outcomes overall survival and the combined outcome progression-free survival). (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

If the absolute numbers in both estimated control groups were identical (see Fig. 3), it was impossible to judge whether authors used absolute numbers of people with event or people being event-free to determine control group risk for their calculations. One would expect a higher overall number of people who survived compared to the number of people who survived without progression (PFS), as OS is based on deaths only, but PFS is the sum of people being dead, with relapse, or progression.

In case authors reported event-free survival like OS and PFS throughout the review but used number of events (i.e., people being dead or with progress) to calculate absolute effects, we extracted information on how authors commented on this in the SoF table like "Instead of OS, mortality is reported in this SoF table, for technical reasons".

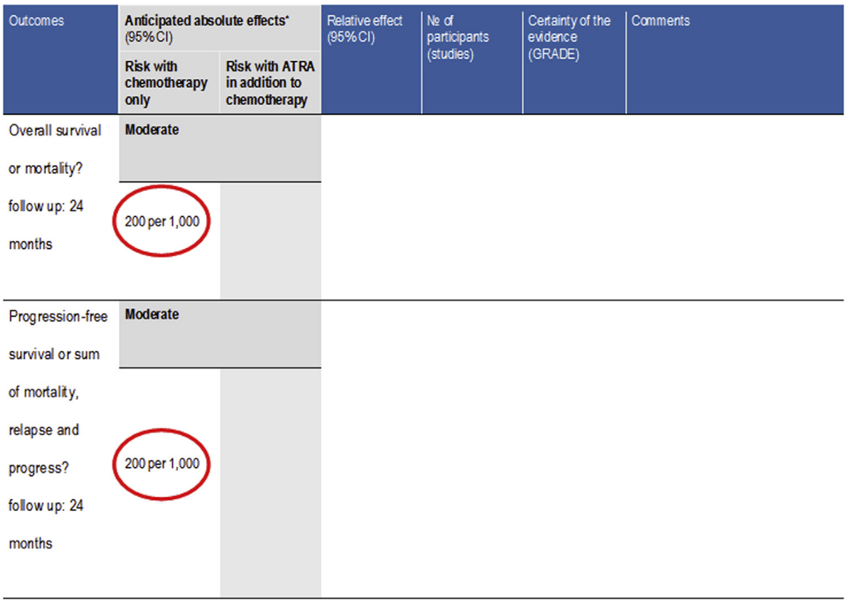
2.4. *Recalculation of absolute effect size estimates*

To check whether review authors calculated the absolute effects from the HR outcome correctly, we recalculated absolute effects based on methods described by Tierney et al. [9]. This article recommends using



Outcomes	Anticipated absolute effects* (95% CI)		Relative effect (95% CI)	№ of participants (studies)	Certainty of the evidence (GRADE)	Comments
	Risk with chemotherapy only	Risk with intervention in addition to chemotherapy				
Mortality (instead of overall survival) follow up: 24 months	Moderate 100 per 1,000					
Mortality, relapse and progression (instead of progression-free survival) follow up: 24 months	Moderate 200 per 1,000					

Fig. 2. Extract from an exemplar SoF table. Numbers for the estimated control group risks are marked in green (for the outcomes mortality and the outcome sum of mortality, relapse, and progress). (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)



Outcomes	Anticipated absolute effects* (95% CI)		Relative effect (95% CI)	№ of participants (studies)	Certainty of the evidence (GRADE)	Comments
	Risk with chemotherapy only	Risk with ATRA in addition to chemotherapy				
Overall survival or mortality? follow up: 24 months	Moderate 200 per 1,000					
Progression-free survival or sum of mortality, relapse and progress? follow up: 24 months	Moderate 200 per 1,000					

Fig. 3. Extract from an exemplar SoF table. Numbers for the estimated control group risks are marked in red (for the outcomes overall survival and progression-free survival). (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

event-free survival data to interpret HR using the following formula:

$$\text{Corresponding intervention risk, per 1000} = (\exp[\ln(\text{proportion of patients event} - \text{free}) \times HR]) \times 1000, \text{ per 1000}$$

These calculations were made based on the HR and baseline risk estimates, which were reported in the Cochrane review.

3. Results

3.1. Search results

As shown in the study flowchart (see Fig. 4), our search led to 483 Cochrane reviews, which were determined to be cancer-related. Study selection showed that 210 of these included an SoF table and of these, 96 reported for at least one time-to-event outcome an HR in the SoF table. The references of the included Cochrane reviews are listed in the online Appendix.

3.2. General characteristics of included reviews

The vast majority of studies, 52 (54%), was published in 2017 or 2016, see also Appendix Table 1. Only one review, without an update publication, has been published in 2011. The largest number of reviews evaluated interventions for hematological malignancies (21 reviews, 22%), 12 reviews (13%) assessed interventions for patients with breast cancer, and four reviews (4%) evaluated "cancer in general". No overviews of reviews were identified.

Five reviews (5%) reported only one outcome with an HR. Ten reviews (10%) mentioned in a comment or footnote, how they determined the baseline risk in the control group, but none of this information was useful to evaluate whether review authors used the number of people with an event or the number of people being event-free to calculate absolute effect size estimates. This was because authors did not report whether the patients were alive or dead at the respective time point. Also, transparent reporting of information where baseline risk data is derived from was often missing.

3.3. Presentation of absolute effect estimates in included reviews

Table 1 summarizes the presentation of absolute effect estimates in the included reviews.

3.3.1. Absolute effects correctly calculated with consistent labeling of outcomes throughout the review

Twenty-eight reviews (29%) correctly calculated absolute effects and labeled the time-to-event outcomes in a

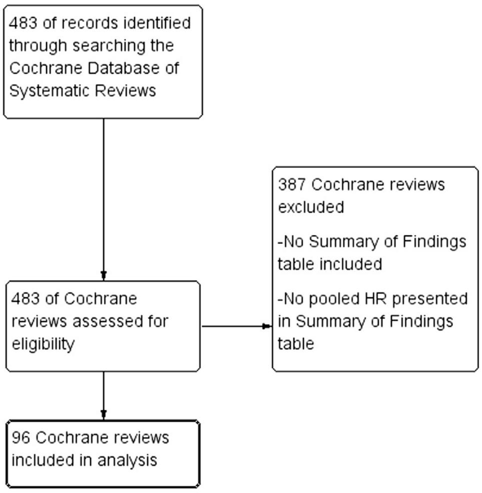

483 of records identified through searching the Cochrane Database of Systematic Reviews
387 Cochrane reviews excluded
-No Summary of Findings table included
-No pooled HR presented in Summary of Findings table
483 of Cochrane reviews assessed for eligibility
96 Cochrane reviews included in analysis

Fig. 4. PRISMA flow diagram of Cochrane reviews.

consistent manner throughout the review, that is, making a clear distinction between people being event-free (e.g., people alive at a specific time point) and people with an event (e.g., people dead at specific time point). Accordingly, time-to-event outcomes were labeled consistently throughout abstract, methods, results section, and SoF table of the respective reviews (see an example in Fig. 5).

3.3.2. Absolute effects correctly calculated but inconsistent presentation of outcomes in SoF table and other parts of the review

Twenty-three reviews (24%) correctly calculated the absolute effects. However, there was inconsistency in how labels for the time-to-event outcomes were used throughout the review. In the SoF table, events (e.g., number of deaths) were used to calculate the absolute effect, whereas in other parts of the review, event-free survival (e.g., OS) was reported, without any explanation in the comment section as shown in Fig. 5, why the name of the outcome changed within the review.

3.3.3. Incorrect calculation of absolute effects

Twelve reviews (13%) provided incorrect absolute effect estimates. The underlying reason was that instead of correctly entering the number of people with the event, the review authors entered the number of people without an event into calculation software, then applying the HR. This led to incorrect results with less people instead of more being alive in the favored arm (see Fig. 6). The review authors reported these incorrect results in the SoF table only; none of the review authors reported these incorrect numbers in the abstract, plain language summary, results, or discussion section.

3.3.4. Unclear results

In seven reviews (7%), it was unclear how review authors determined the control group risk and whether direction of results was correct. This was the case when the control group risk for both outcomes of interest (e.g., OS and PFS) was identical, as shown in Fig. 3.

3.3.5. No absolute effects calculated

Twenty-six reviews (27%) did not calculate an absolute effect. However, five of these reviews reported mean survival ranges for both the control and intervention arm or weighted mean survival with 95% confidence intervals in the SoF table but without any explanation how these had been calculated. Thus, their provenance remained unclear, and it could not be judged whether they were correct or incorrect.

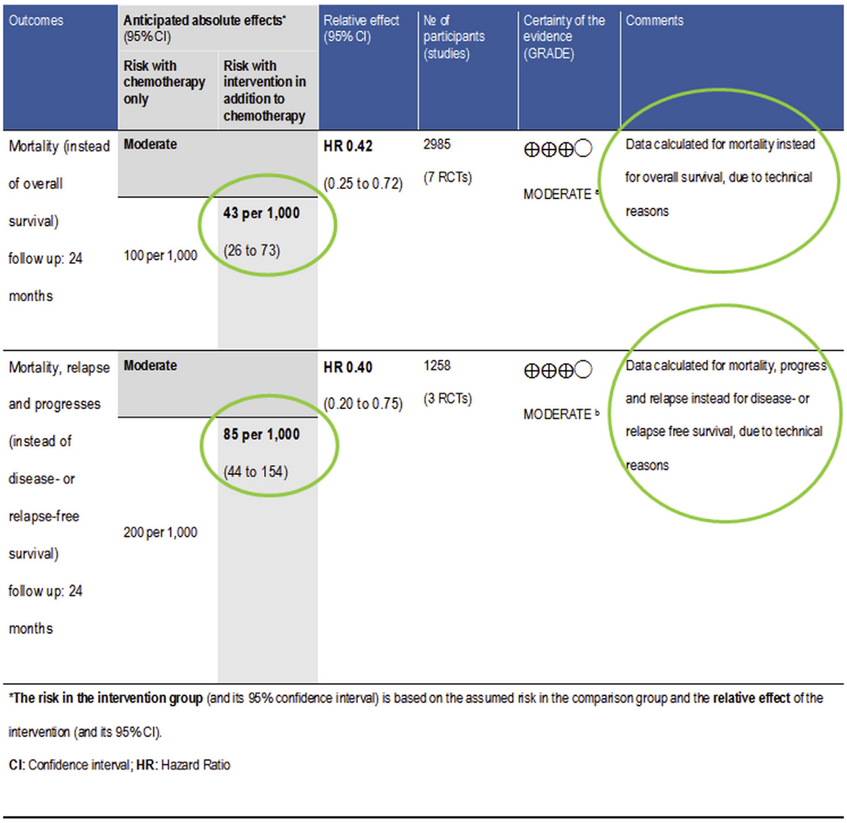
4. Discussion

4.1. Principal findings

This methodological review shows that absolute effects based on time-to-event outcomes are calculated correctly and presented in a readily interpretable way in less than one in three Cochrane reviews related to cancer (out of 96 reviews). In about every fourth review, the absolute effects were correctly calculated but the respective outcomes were labeled inconsistently and potentially misleading without any comment why authors calculated event-free survival (e.g., OS) throughout the review but reported number of events (e.g., death) in SoF tables. Twelve percent provided incorrect absolute effects in the SoF tables, because inappropriate data were entered into the calculation software. As the review authors did not report the results of the incorrect calculations in the abstract, results section, or

Table 1. Presentation of absolute effect estimates in the included Cochrane reviews

Calculation of absolute effects and labeling of outcomes	Cochrane reviews (*N* = 96)	Figure
Absolute effects correctly calculated	28 + 23 (53%)	5
Consistent labeling of outcomes throughout the review	28 (29%)	
Inconsistent labeling of outcomes throughout the review	23 (24%)	
Absolute effects incorrectly calculated	12 (13%)	6
Unclear results	7 (7%)	3
No absolute effects calculated	26 (27%)	



Outcomes	Anticipated absolute effects* (95% CI)		Relative effect (95% CI)	№ of participants (studies)	Certainty of the evidence (GRADE)	Comments
	Risk with chemotherapy only	Risk with intervention in addition to chemotherapy				
Mortality (instead of overall survival) follow up: 24 months	**Moderate** 100 per 1,000	**43 per 1,000** (26 to 73)	**HR 0.42** (0.25 to 0.72)	2985 (7 RCTs)	⊕⊕⊕○ MODERATE [a]	Data calculated for mortality instead for overall survival, due to technical reasons
Mortality, relapse and progresses (instead of disease- or relapse-free survival) follow up: 24 months	**Moderate** 200 per 1,000	**85 per 1,000** (44 to 154)	**HR 0.40** (0.20 to 0.75)	1258 (3 RCTs)	⊕⊕⊕○ MODERATE [b]	Data calculated for mortality, progress and relapse instead for disease- or relapse free survival, due to technical reasons

*__The risk in the intervention group__ (and its 95% confidence interval) is based on the assumed risk in the comparison group and the __relative effect__ of the intervention (and its 95% CI).

CI: Confidence interval; **HR**: Hazard Ratio

Fig. 5. Extract from an exemplar SoF table. Numbers for the correctly calculated absolute effects and comments are marked in green, assuming the HR < 1 favors the intervention group. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

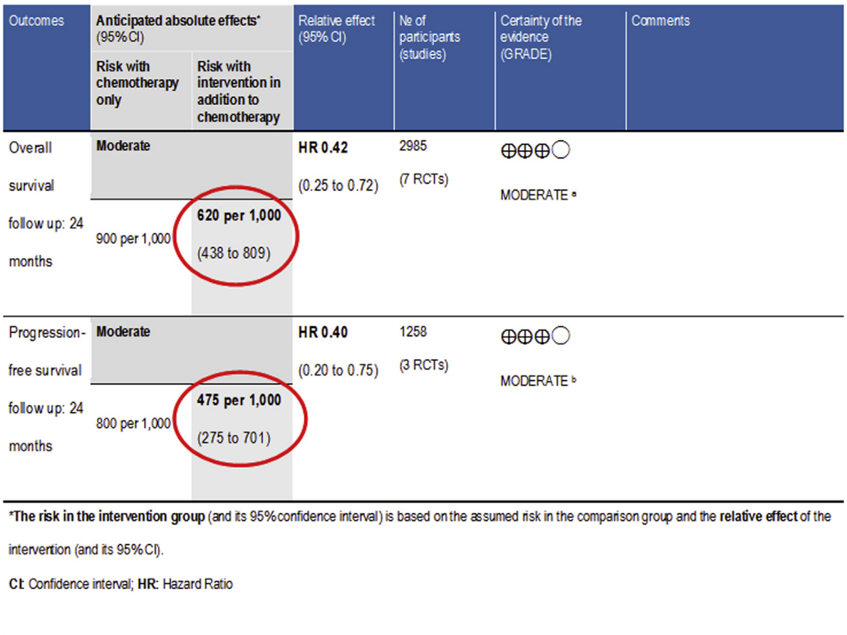
conclusion, this error is not transferred to other sections of the respective reviews. For 7%, it was completely unclear if absolute effects were correctly calculated, because it was unclear whether control group risk numbers described people with event or people without event. In the remaining 27%, no absolute effect had been calculated at all.

4.2. Strengths and limitations of this methodological review

We performed this methodological review based on an a priori developed protocol. The strength to the validity of our findings is based on a comprehensive search, a screening process done in duplicate, a piloted data extraction form, and data abstraction in duplicate. The reliability of this work is ensured through adherence to the review methods proposed by Cochrane and reporting in accordance with the PRISMA standards. The small sample size might have influenced our results; however, at the date of the search, no more cancer-related Cochrane reviews including time-to-event outcomes described as HR in a SoF table were available. Another limitation is that we evaluated only cancer-related reviews; therefore, our findings are directly applicable to cancer reviews only. Cochrane reviews evaluating patients with other types of disease and time-to-event outcomes should also be assessed, as these reviews might report other outcomes like time to hospitalization, time to discharge, or time to recovery, which could result in different findings. Although the focus of this review was on Cochrane reviews only, which has recently mandated the inclusion of absolute effects and SoFs, we expect similar issues to affect non-Cochrane reviews. We evaluated 10 high-impact cancer journals publishing systematic reviews within the same time period as mentioned above but could not identify any review out of 177 reporting absolute effects for HR outcomes or presenting SoF tables (unpublished data).

4.3. Strengths and limitations in relation to other methodological reviews

To date, there has not been other research directed to the reporting of absolute effect size estimates specifically



Outcomes	Anticipated absolute effects* (95% CI)		Relative effect (95% CI)	№ of participants (studies)	Certainty of the evidence (GRADE)	Comments
	Risk with chemotherapy only	Risk with intervention in addition to chemotherapy				
Overall survival follow up: 24 months	Moderate 900 per 1,000	620 per 1,000 (438 to 809)	HR 0.42 (0.25 to 0.72)	2985 (7 RCTs)	⊕⊕⊕○ MODERATE a	
Progression-free survival follow up: 24 months	Moderate 800 per 1,000	475 per 1,000 (275 to 701)	HR 0.40 (0.20 to 0.75)	1258 (3 RCTs)	⊕⊕⊕○ MODERATE b	

*The risk in the intervention group (and its 95% confidence interval) is based on the assumed risk in the comparison group and the relative effect of the intervention (and its 95% CI).

CI: Confidence interval; HR: Hazard Ratio

Fig. 6. Extract from an exemplar SoF table. Numbers for the incorrectly calculated absolute effect, assuming that HR < 1 favors the intervention group. The incorrect numbers are marked in red. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

for time-to-event outcomes in SoF tables. With regard to its specific focus and relevance to all investigators using time-to-event outcomes in SoF tables, our review is unique.

Prior methodological work assessing the frequency of reporting of absolute effects any outcome type systematic review revealed that they are rarely reported. Alonso-Coello et al. assessed 98 Cochrane and 104 non-Cochrane systematic reviews, which were published in 2010 and not limited to a certain field of disease. Overall only 36.1% of these reviews presented absolute effect estimates for the most patient-important outcome. In 32 systematic reviews that included SoF tables, all authors calculated absolute effects. Twenty-nine of these reviews were Cochrane reviews and three reviews were non-Cochrane reviews. The relative effect measures on which the absolute effects were based were primarily risk ratios or ORs; a smaller amount of only 6.4% were calculated from HRs. If absolute effects were reported, the source of the baseline risk applied for calculation was often not given [10]. This is in accordance with our findings showing sparse reporting of where baseline risk data are derived from. Agarwal et al. performed a methodological review to assess the frequency of reporting of absolute effects in the abstracts of systematic reviews. They included 96 Cochrane and 94 non-Cochrane reviews published in 2010 and revealed that absolute effects were reported in the abstracts of 22.5% of the respective systematic reviews. Again, the relative effect estimates corresponding to the calculated absolute effects where predominantly relative risks and ORs, only a small number (5.8%) were HR [2].

Our review demonstrated incorrect calculation and reporting of absolute effects based on HR in the SoF tables of cancer-related Cochrane reviews. Prior work suggested flaws in the calculation of absolute effects of clinical studies. A review evaluating 734 randomized controlled trials, published in high impact general medical journals of which 373 investigated time-to-event outcomes, found that only half of randomized controlled trials reporting number-needed-to-treat or number-needed-to harm from such outcomes used appropriate calculation methods [11]. Prior studies have also addressed the challenge of calculating numbers-needed-to-treat for time-to-event data in the setting of competing risks and reviewed the potential issue of varying follow-up times [12,13].

4.4. Meaning of this methodological review: explanations, implications, and further research

Our methodological assessment shed light on the problems review authors face when they try to calculate absolute effects based on time-to-event data in SoF tables. There is currently no written guidance on calculating HR-based absolute effects and how best to present them in SoF tables. Therefore, it may be particularly difficult for reviewers to do this properly and also for journal editors and peer reviewers to identify mistakes. This work demonstrates the need for additional training and guidance of review authors working with time-to-event data. A workshop addressing this issue was held at an international conference already and further workshops are planned. Also, additional training materials in the form of written materials, online modules, and webinars for review authors and other GRADE users on how to calculate absolute effects are currently being developed and will soon be disseminated. Moreover, the GRADEpro GDT software has been adapted in September 2018 to provide review authors the choice as to whether to enter the number of people with events or number of people without events for the control group risk. This will allow better consistency of use of outcomes throughout the review. In addition, there appears need for additional oversight of review authors to identify incorrect and misleading information before publication.

A further key aspect to consider is how review authors should estimate the absolute risk in the control group for time-to-event outcomes: should it be based on one study or on all included studies or on data from representative observational studies and which time point should be used? Confidence intervals of calculated absolute effects do not incorporate uncertainty in the assumed control risks and are not considered by the calculation according to Tierney et al. [6]. This is of special concern if we look at long-term survival with a low or moderate mortality and a corresponding high number of censored patients (i.e., a low number of patients under risk and a high censoring rate). These aspects will be considered in another article.

5. Conclusion

Based on our systematic review, in less than one in three cancer-related Cochrane reviews that included at least one time-to-event outcome, absolute effect size estimates were correctly calculated and appropriately reported. This was due to missing comments and/or entering incorrect numbers into the GRADEpro GDT software. There is an urgent need for additional training materials and guidance for review authors, editors, and peer-reviewers on how to calculate and present absolute effects based on HR data.

Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.jclinepi.2018.12.006.

References

[1] Gigerenzer G, Gaissmaier W, Kurz-Milcke E, Schwartz LM, Woloshin S. Helping doctors and patients make sense of health statistics. Psychol Sci Public Interest 2007;8:53—96.

[2] Agarwal A, Johnston BC, Vernooij RWM, Carrasco-Labra A, Brignardello-Petersen R, Neumann I, et al. Authors seldom report the most patient-important outcomes and absolute effect measures in systematic review abstracts. J Clin Epidemiol 2017;81:3—12.

[3] Malenka DJ, Baron JA, Johansen S, Wahrenberger JW, Ross JM. The framing effect of relative and absolute risk. J Gen Intern Med 1993;8:543—8.

[4] Naylor C, Chen E, Strauss B. Measured enthusiasm: does the method of reporting trial results alter perceptions of therapeutic effectiveness? Ann Intern Med 1992;117:916—21.

[5] Guyatt G, Oxman AD, Akl EA, Kunz R, Vist G, Brozek J, et al. GRADE guidelines: 1. Introduction-GRADE evidence profiles and summary of findings tables. J Clin Epidemiol 2011;64:383—94.

[6] van Dalen EC, Tierney JF, Kremer LCM. Tips and tricks for understanding and using SR results. No. 7: time-to-event data. Evidence Based Child Health A Cochrane Rev J 2007;2:1089—90.

[7] Moher D, Liberati A, Tetzlaff J, Altman DG. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. BMJ 2009;339:b2535.

[8] Rosenbaum SE, Glenton C, Oxman AD. Summary-of-findings tables in Cochrane reviews improved understanding and rapid retrieval of key information. J Clin Epidemiol 2010;63:620—6.

[9] 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.

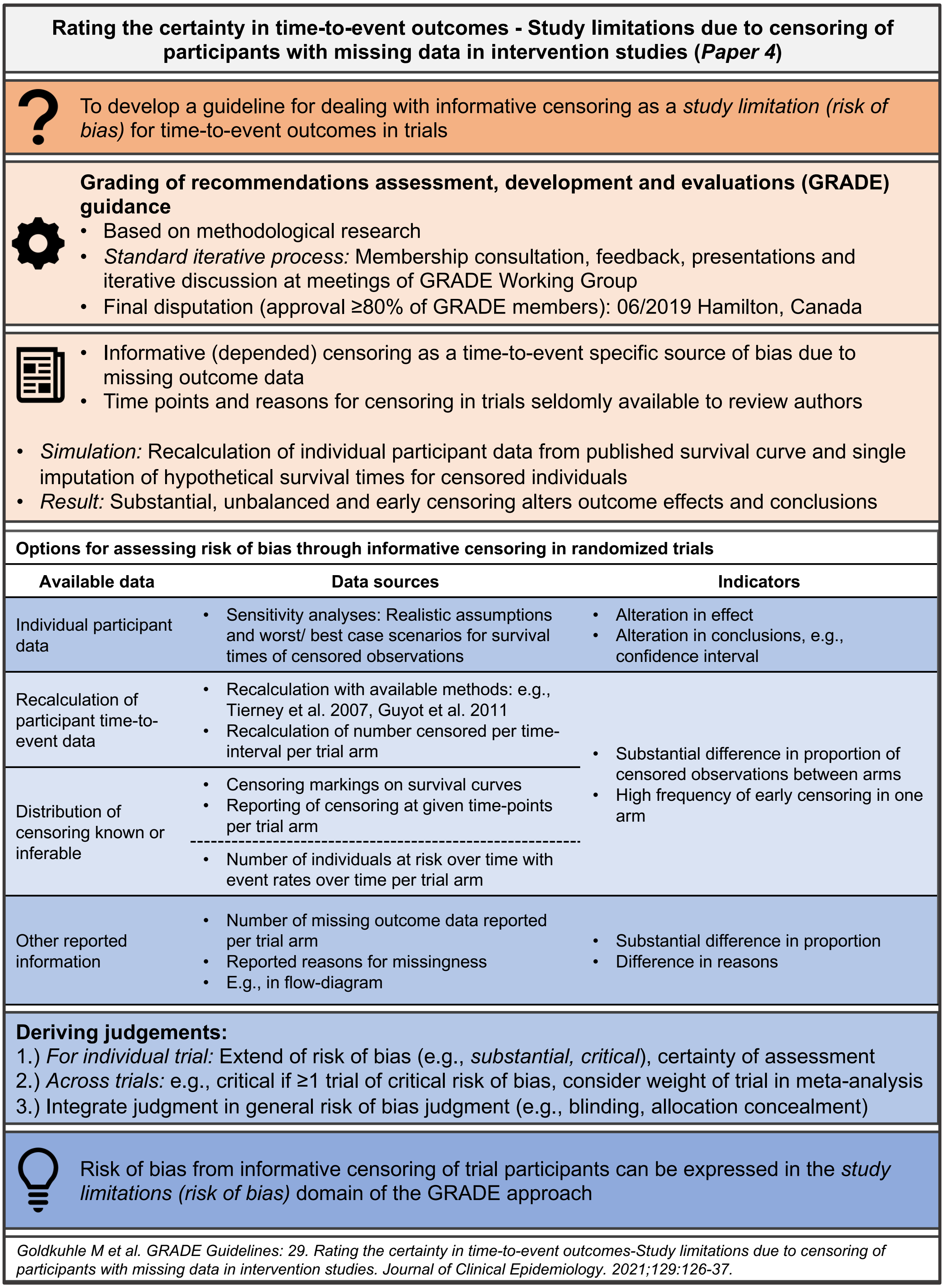
[10] Alonso-Coello P, Carrasco-Labra A, Brignardello-Petersen R, Neumann I, Akl EA, Vernooij RWM, et al. Systematic reviews experience major limitations in reporting absolute effects. J Clin Epidemiol 2016;72:16—26.

[11] Hildebrandt M, Vervölgyi E, Bender R. Calculation of NNTs in RCTs with time-to-event outcomes: a literature review. BMC Med Res Methodol 2009;9:21.

[12] Bender R, Kromp M, Kiefer C, Sturtz S. Absolute risks rather than incidence rates should be used to estimate the number needed to treat from time-to-event data. J Clin Epidemiol 2013;66:1038—44.

[13] Gouskova NA, Kundu S, Imrey PB, Fine JP. Number needed to treat for time-to-event data with competing risks. Stat Med 2014;33:181—92.

6.4. Guidance for handling informative censoring as a study-limitation in evidence syntheses *(Paper 4)*



Rating the certainty in time-to-event outcomes - Study limitations due to censoring of participants with missing data in intervention studies (*Paper 4*)

? To develop a guideline for dealing with informative censoring as a *study limitation (risk of bias)* for time-to-event outcomes in trials

Grading of recommendations assessment, development and evaluations (GRADE) guidance

- Based on methodological research
- *Standard iterative process:* Membership consultation, feedback, presentations and iterative discussion at meetings of GRADE Working Group
- Final disputation (approval ≥80% of GRADE members): 06/2019 Hamilton, Canada

- Informative (depended) censoring as a time-to-event specific source of bias due to missing outcome data
- Time points and reasons for censoring in trials seldomly available to review authors

- *Simulation:* Recalculation of individual participant data from published survival curve and single imputation of hypothetical survival times for censored individuals
- *Result:* Substantial, unbalanced and early censoring alters outcome effects and conclusions

Options for assessing risk of bias through informative censoring in randomized trials

Available data	Data sources	Indicators
Individual participant data	• Sensitivity analyses: Realistic assumptions and worst/ best case scenarios for survival times of censored observations	• Alteration in effect • Alteration in conclusions, e.g., confidence interval
Recalculation of participant time-to-event data	• Recalculation with available methods: e.g., Tierney et al. 2007, Guyot et al. 2011 • Recalculation of number censored per time-interval per trial arm	• Substantial difference in proportion of censored observations between arms • High frequency of early censoring in one arm
Distribution of censoring known or inferable	• Censoring markings on survival curves • Reporting of censoring at given time-points per trial arm • Number of individuals at risk over time with event rates over time per trial arm	
Other reported information	• Number of missing outcome data reported per trial arm • Reported reasons for missingness • E.g., in flow-diagram	• Substantial difference in proportion • Difference in reasons

Deriving judgements:
1.) *For individual trial:* Extend of risk of bias (e.g., *substantial, critical*), certainty of assessment
2.) *Across trials:* e.g., critical if ≥1 trial of critical risk of bias, consider weight of trial in meta-analysis
3.) Integrate judgment in general risk of bias judgment (e.g., blinding, allocation concealment)

Risk of bias from informative censoring of trial participants can be expressed in the *study limitations (risk of bias)* domain of the GRADE approach

Goldkuhle M et al. GRADE Guidelines: 29. Rating the certainty in time-to-event outcomes-Study limitations due to censoring of participants with missing data in intervention studies. Journal of Clinical Epidemiology. 2021;129:126-37.

Figure 10: Graphical abstract for paper 4 (2).

6.4.1. Publication status

This article was published in January 2021 in the Journal of Clinical Epidemiology.

The work shares of the individual participants and their involvement in this paper are detailed in appendix 11.4.1.

This paper is part of the consortium of articles awarded with the German Network for Evidence-based Medicine's David Sackett prize 2021 (chapter 9.3).

6.4.2. Synopsis (2)

Informative censoring, elaborated in chapter 3.2.1, constitutes a time-to-event specific source of bias that is often neglected in evidence syntheses, as emphasized in the previously reported meta-epidemiological studies. Moreover, no formalized guidelines on how to identify a potential risk of bias in primary trial publications and how deal with the issue in evidence syntheses were available.

The fourth article contributing to this dissertation is a guidance article systematically developed within and according to the standards of the renowned GRADE Working Group (2). It complements previous GRADE guidance that focusses on the assessment and consideration of bias through missing outcome data for binary and continuous outcomes in systematic reviews and evidence-based guidelines (126).

The systematic methods required to develop GRADE guidelines are formalized elsewhere (see also chapter 3.3.4) (142, 143). In brief, the guideline was developed through iterative discussions in face-to-face and digital meetings of internationally recognized experts. These included methodologists, biostatisticians and clinicians who have already made important contributions to the field. As basis, the development process as well as the guidance article require a conceptual and theoretical elaboration of the problem, evidence to back the issue and practical examples that show how the proposed principles are to be applied.

In addition, within this particular article, a straightforward modeling approach was used to demonstrate how loss to follow-up censoring can affect the estimates of trial outcomes: after recalculation of individual participant data from a suitable survival curve and confirmation of correct recalculation with help of the approach proposed by Guyot et al. 2012 (145), realistic event times were imputed for individuals with censored observations to simulate informative censoring. The resulting estimate indicated a loss of statistical significance of the original estimate and increased uncertainty due to informative censoring.

After its final presentation and disputation, the article was accepted for submission for publication by about 100 attending GRADE Working Group members at a GRADE meeting in Hamilton, Canada.

Besides a comprehensive theoretical background on the sources and potential impact of the issue, the guidance includes concrete suggestions for the identification potential informative censoring from reported trial information.

Depending on the data from trial publications available to evidence synthesis authors, several sources are suggested: most straightforward for a judgement is the direct reporting of details on censoring with reasons per trial arm. If information on missing outcome data per trial arm is reported in sufficient detail, this data may also be helpful in some cases. Furthermore, if Kaplan-Meier plots are presented, censoring markings on or below curves (as in figure 1), the number of individuals at risk or information generated through recalculation of summary data can be applied for inference on the distribution of censoring in the compared trial groups.

Particular attention for informative censoring is required, for example, if censoring patterns between groups are not explainable by relative effects, if censoring occurs early in one group and/ or is apparently differently distributed across the follow-up time. In figure 1 of chapter 3.1.1, an example that is also used in the article, the number of censored individuals reported below the curve is substantially higher in the neratinib group as compared to the placebo group across the entire follow-up period – while the number of individuals at risk is consequentially lower. The plot and the reported HR indicate, however, that treatment with neratinib considerably improves the outcome invasive disease-free survival. Why the number of individuals at risk is lower in the preferred treatment group throughout the entire observation period would require further investigation and may pose a risk of bias (2, 12).
Finally, the article describes how to derive and formalize judgments on the degree of concern regarding a risk of bias raised through potential informative censoring in a particular trial, how to translate this concern to a body of evidence of multiple trials and how to incorporate such summary judgements into an overall GRADE certainty of evidence rating.

In line with the overall goals of this dissertation project, this GRADE guidance informs authors of evidence syntheses about an often-neglected methodological issue of time-to-event analysis. It is intended to improve the communication of the results of meta-analyses of time-to-event outcomes and their certainty in the future.

6.4.3. Full-text publication

The supplementary material accompanying this publication is provided in appendix 11.4.2.

Journal of Clinical Epidemiology 129 (2021) 126−137

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

[a]*Department 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*

[b]*Department of Medical Biometry, Institute for Quality and Efficiency in Health Care, Im Mediapark 8, D-50670 Cologne, Germany*

[c]*Department of Internal Medicine, American University of Beirut, P.O.Box 11-0236, Lebanon, Canada*

[d]*Department of Health Research Methods, Evidence, and Impact, McMaster University, 1280 Main St. W., Hamilton, Ontario, L8S 4K1, Canada*

[e]*Princess Máxima Center for Pediatric Oncology, Heidelberglaan 25, 3584CS, Utrecht, The Netherlands*

[f]*Department of Biostatistics, University of Liverpool, Block F, Waterhouse Building, 1-5 Brownlow Street, Liverpool, L69 3GL, UK*

[g]*Department of Medicine, University of Kansas Health System, 3901 Rainbow Blvd, MS3002, Kansas City, KS, 66160, USA*

[h]*Department of Medicine, McMaster University, 1280 Main St. W., Hamilton, Ontario, L8S 4K1, Canada*

[i]*McMaster GRADE Centre & Michael G DeGroote Cochrane Canada Centre, McMaster University, Hamilton, Ontario, Canada*

[j]*Centre for Statistics in Medicine, University of Oxford, Botnar Research Centre, Windmill Rd, Oxford, OX3 7LD, UK*

[k]*City of Hope, 1500 Duarte Rd, Duarte, CA, 91010, USA*

[l]*Unit 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).

https://doi.org/10.1016/j.jclinepi.2020.09.017

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.

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 administrational 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-trea 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

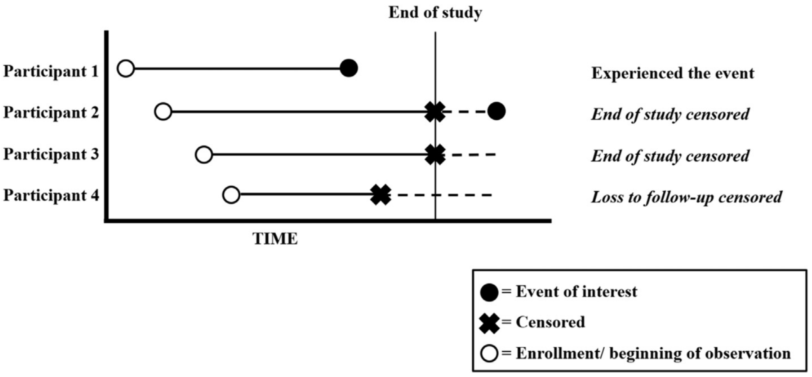


Fig. 1. Types of censoring: For participant 1, the occurrence of the outcome event is observable. Participants 2 and 3 are censored because of the administrational closure of the study. The variation in their duration of follow-up and the differing censoring time points result from the staggered recruiting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

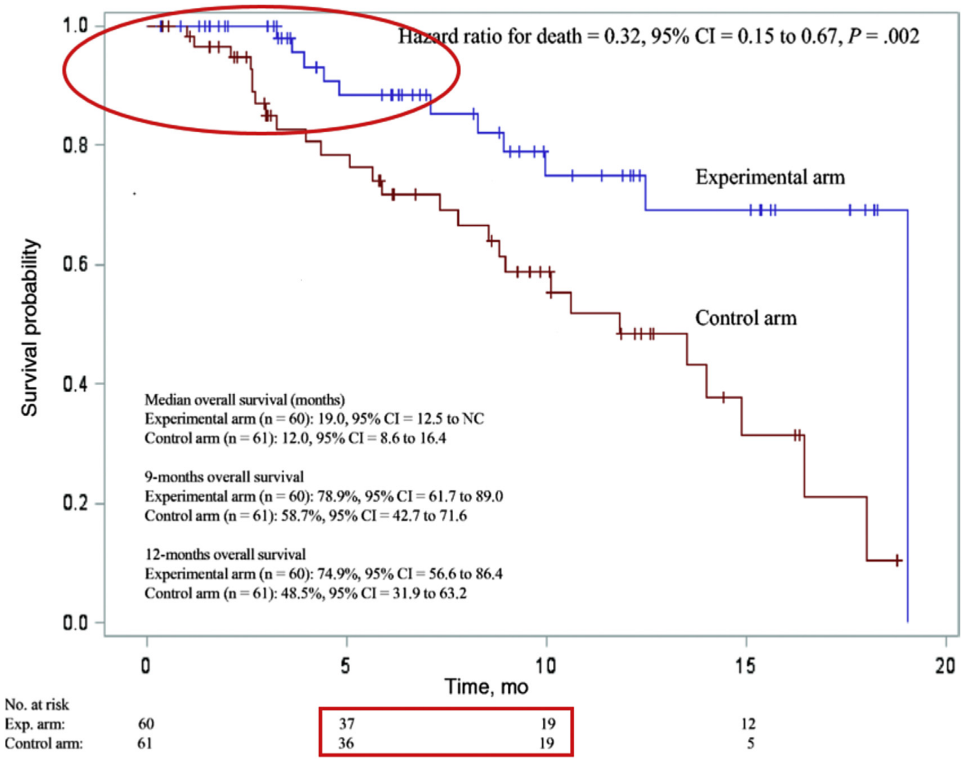
To demonstrate the impact of early censoring on the results, we considered a hypothetical scenario in which all participants who were censored before 7 months of follow-up experience the event 1 month after censoring, that is these data are no longer censored but are counted as events. This assumption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



No. at risk				
Exp. arm:	60	37	19	12
Control arm:	61	36	19	5

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 reconsent 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.

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 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.

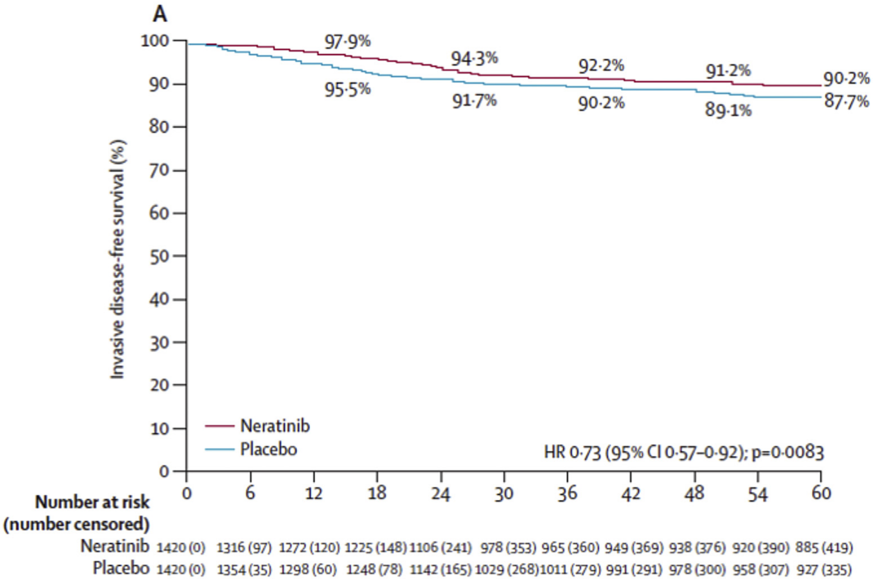
6.4. *Rating the risk of bias resulting from censoring of participants because of missing follow-up data and deriving an overall judgment for an individual study*

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



Number at risk (number censored)	0	6	12	18	24	30	36	42	48	54	60
Neratinib	1420 (0)	1316 (97)	1272 (120)	1225 (148)	1106 (241)	978 (353)	965 (360)	949 (369)	938 (376)	920 (390)	885 (419)
Placebo	1420 (0)	1354 (35)	1298 (60)	1248 (78)	1142 (165)	1029 (268)	1011 (279)	991 (291)	978 (300)	958 (307)	927 (335)

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. **Marialene 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] Abraira 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, Stepniewska 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 trials2019. 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:l4898.

[25] Cochrane Community. Glossary: The Cochrane Colloaboration. 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.

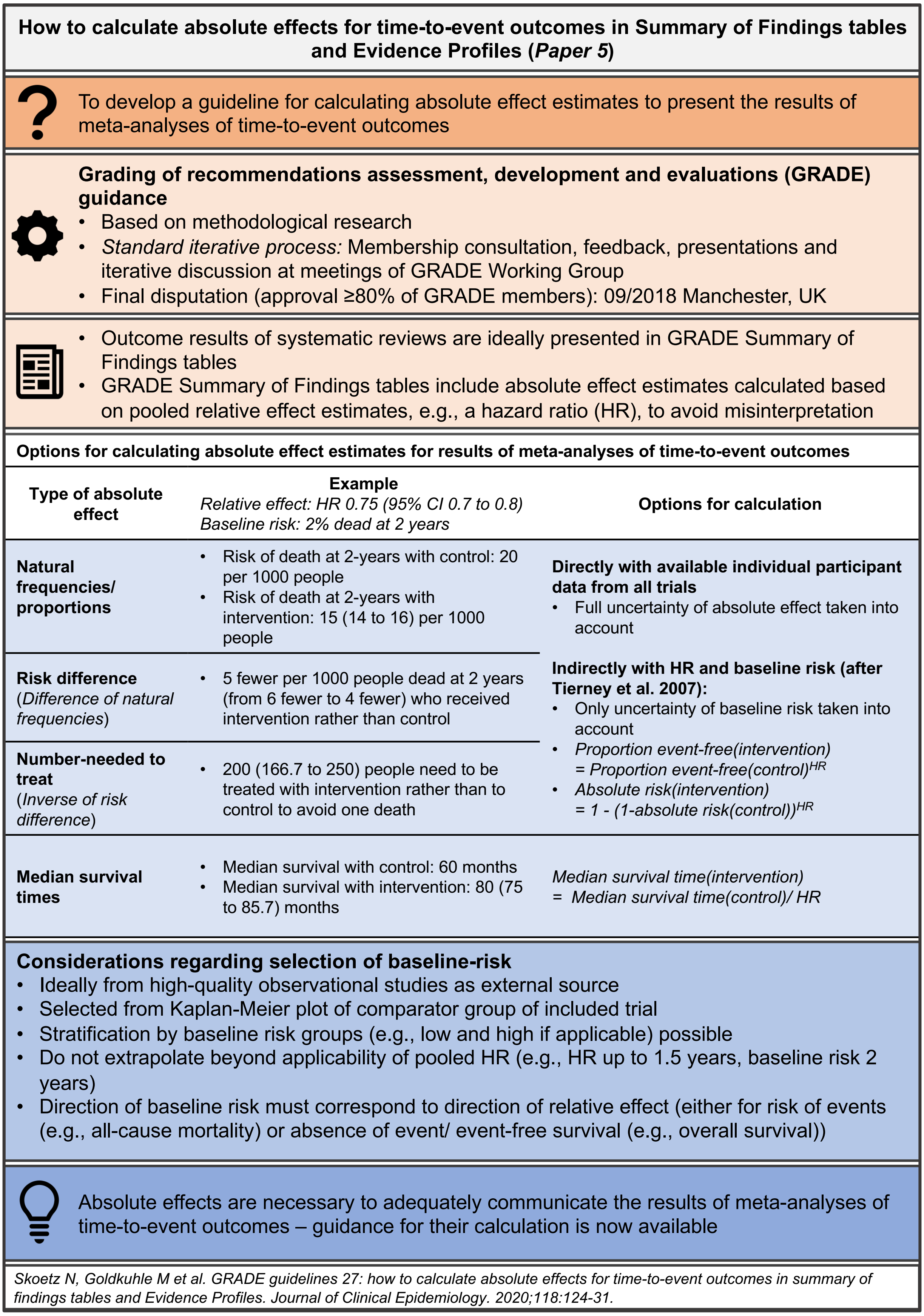
[51] Cole SR, Hernán MA. Constructing inverse probability weights for marginal structural models. Am J Epidemiol 2008;168:656−64.

[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.

6.5. Guidance for presenting the results of meta-analyses of time-to-event outcomes in form of absolute effects *(Paper 5)*



How to calculate absolute effects for time-to-event outcomes in Summary of Findings tables and Evidence Profiles (*Paper 5*)

To develop a guideline for calculating absolute effect estimates to present the results of meta-analyses of time-to-event outcomes

Grading of recommendations assessment, development and evaluations (GRADE) guidance

- Based on methodological research
- *Standard iterative process:* Membership consultation, feedback, presentations and iterative discussion at meetings of GRADE Working Group
- Final disputation (approval ≥80% of GRADE members): 09/2018 Manchester, UK

- Outcome results of systematic reviews are ideally presented in GRADE Summary of Findings tables
- GRADE Summary of Findings tables include absolute effect estimates calculated based on pooled relative effect estimates, e.g., a hazard ratio (HR), to avoid misinterpretation

Options for calculating absolute effect estimates for results of meta-analyses of time-to-event outcomes

Type of absolute effect	Example *Relative effect: HR 0.75 (95% CI 0.7 to 0.8)* *Baseline risk: 2% dead at 2 years*	Options for calculation
Natural frequencies/ proportions	• Risk of death at 2-years with control: 20 per 1000 people • Risk of death at 2-years with intervention: 15 (14 to 16) per 1000 people	**Directly with available individual participant data from all trials** • Full uncertainty of absolute effect taken into account
Risk difference (*Difference of natural frequencies*)	• 5 fewer per 1000 people dead at 2 years (from 6 fewer to 4 fewer) who received intervention rather than control	**Indirectly with HR and baseline risk (after Tierney et al. 2007):** • Only uncertainty of baseline risk taken into account • *Proportion event-free(intervention) = Proportion event-free(control)*^{HR} • *Absolute risk(intervention) = 1 - (1-absolute risk(control))*^{HR}
Number-needed to treat (*Inverse of risk difference*)	• 200 (166.7 to 250) people need to be treated with intervention rather than to control to avoid one death	
Median survival times	• Median survival with control: 60 months • Median survival with intervention: 80 (75 to 85.7) months	*Median survival time(intervention) = Median survival time(control)/ HR*

Considerations regarding selection of baseline-risk

- Ideally from high-quality observational studies as external source
- Selected from Kaplan-Meier plot of comparator group of included trial
- Stratification by baseline risk groups (e.g., low and high if applicable) possible
- Do not extrapolate beyond applicability of pooled HR (e.g., HR up to 1.5 years, baseline risk 2 years)
- Direction of baseline risk must correspond to direction of relative effect (either for risk of events (e.g., all-cause mortality) or absence of event/ event-free survival (e.g., overall survival))

Absolute effects are necessary to adequately communicate the results of meta-analyses of time-to-event outcomes – guidance for their calculation is now available

Skoetz N, Goldkuhle M et al. GRADE guidelines 27: how to calculate absolute effects for time-to-event outcomes in summary of findings tables and Evidence Profiles. Journal of Clinical Epidemiology. 2020;118:124-31.

Figure 11: Graphical abstract for paper 5 (10).

6.5.1. Publication status

This article was published in February 2020 in the Journal of Clinical Epidemiology.

The work shares of the individual participants and their involvement in this paper are detailed in appendix 11.5.1.

This paper is part of the consortium of articles awarded with the German Network for Evidence-based Medicine's David Sackett prize 2021 (chapter 9.3).

6.5.2. Synopsis (10)

The fifth article included in this dissertation is a GRADE guideline for the calculation of absolute effect estimates for time-to-event outcomes in meta-analyses (10). Presentation of meta-analysis results as absolute effects is critical for the comprehensible communication of the magnitude of the pooled relative effect estimates. As previously shown theoretically and empirically (chapter 3.3), the calculation of absolute effect estimates from pooled HRs is associated with certain complexities (11). A GRADE guideline tackling these difficulties therefore constitutes a valuable resource in order to improve the correct communication of evidence synthesis results of time-to-event outcomes.

Like the GRADE guideline on informative censoring (chapter 6.4), this guideline was systematically developed following the previously elaborated process underlying all GRADE guideline articles (10, 142, 143). Similar to the previously elaborated GRADE guideline article, the guidance was developed under continuous discussion with international experts in small- and large group meetings and involved the generation of conceptual and empirical evidence, which is provided through the meta-epidemiological study that addresses the calculation of absolute effect estimates in cancer-related Cochrane reviews (chapter 6.3) (11). The final guidance article was discussed and approved by about 60 attending members of the GRADE Working Group at a GRADE meeting in Manchester, UK.

The article provides an in-depth elaboration of the role of absolute effect estimates to communicate the results of evidence syntheses. It presents important considerations for the interpretation and presentation of the outputs of meta-analyses of time-to-event outcomes, including the correct interpretation of the HR.
The guidance gives detailed explanations for several different methods to calculate absolute effect estimates for meta-analyses of time-to-event outcomes: direct calculation methods for natural frequencies and the risk difference, which require individual participant data, and indirect calculation methods, which use aggregate time-to-event data from the included trials or secondary evidence. The article also highlights the central considerations underlying these methods. These include, for example, rules for choosing adequate sources and time points to for baseline risks from internal control groups as well as explanations for selecting external sources of evidence.
Potential shortcomings of the presented methods are also discussed. For instance, that most indirect methods cannot take into account the uncertainty of the baseline risk without more advanced statistical techniques. In consequence, absolute effect estimates are presented only with the uncertainty of the relative effects so that, for example, the calculated CIs represent the statistical uncertainty of the pooled HR alone.
Finally, the guidance suggests and elaborates several options for alternative absolute effects for time-to-event outcomes, such as the number-needed-to-treat at a given time point, median

survival times or their difference between treatment groups, which may be more comprehensible to certain groups of decision makers.

This guideline is intended improve the calculation and presentation of absolute effect estimates for time-to-event outcomes in evidence syntheses and thus contribute to the overreaching goal of this dissertation, namely to improve meta-analyses of time-to-event outcomes in evidence syntheses and the decisions that are based on their results.

6.5.3. Full-text publication

The supplementary material accompanying this publication is provided in appendix 11.5.2.

ELSEVIER

Journal of Clinical Epidemiology 118 (2020) 124−131

Journal of Clinical Epidemiology

SERIES

GRADE guidelines 27: how to calculate absolute effects for time-to-event outcomes in summary of findings tables and Evidence Profiles

Nicole Skoetz[a,*], Marius Goldkuhle[a], Elvira C. van Dalen[b], Elie A. Akl[c], Marialena Trivella[d], Reem A. Mustafa[e], Artur Nowak[f], Philipp Dahm[g], Holger Schünemann [h], Ralf Bender[i], GRADE Working Group

[a]Department I of Internal Medicine, University of Cologne, Center for Integrated Oncology Aachen Bonn Cologne Duesseldorf, Kerpener Street 62, 50937 Cologne, Germany
[b]Princess Máxima Center for Pediatric Oncology, Heidelberglaan 25, 3584CS Utrecht, The Netherlands
[c]Department of Internal Medicine, American University of Beirut, Lebanon
[d]Centre for Statistics in Medicine, University of Oxford, Oxford, UK
[e]Department of Medicine, University of Kansas Health System, 3901 Rainbow Blvd, MS3002, Kansas City, KS 66160, USA
[f]Evidence Prime Inc, 175 Longwood Road South, Suite 305, Hamilton, Ontario L8P 0A1, Canada
[g]Minneapolis VA Health Care System, Urology Section 112D, One Veterans Drive, Minneapolis, MN 55417, USA
[h]Department of Health Research Methods, Evidence, and Impact, McMaster University, 1280 Main Street W, Hamilton, Ontario L8S 4K1, Canada
[i]Department of Medical Biometry, Institute for Quality and Efficiency in Health Care, Im Mediapark 8, D-50670 Cologne, Germany

Accepted 10 October 2019; Published online 9 November 2019

Abstract

Objectives: To provide GRADE guidance on how to prepare Summary of Findings tables and Evidence Profiles for time-to-event outcomes with a focus on the calculation of the corresponding absolute effect estimates.

Study Design and Setting: This guidance was justified by a research project identifying frequent errors and limitations in the presentation of time-to-event outcomes in the Summary of Findings tables. We developed this guidance through an iterative process that included membership consultation, feedback, presentation, and discussion at meetings of the GRADE Working Group.

Results: Review authors need to carefully consider the definition of the outcome of interest; although often the event is used as label for the outcome of interest (e.g., death or mortality), the event-free survival (e.g., overall survival) is reported throughout individual studies. Review authors should calculate the absolute effect correctly, either for the event or absence of the event. We also provide examples on how to calculate the absolute effects for events and the absence of events for various baseline or control group risks and time points.

Conclusions: This article aids in the development of Summary of Findings tables and Evidence Profiles, including time-to-event outcomes, and addresses the most common scenarios when calculating absolute effects in order to provide an accurate interpretation. © 2019 Elsevier Inc. All rights reserved.

Keywords: GRADE guidance; Time-to-event outcomes; Hazard ratio; Absolute effects; Summary of findings table; Evidence Profile

1. Introduction

The GRADE approach provides a systematic and transparent framework for rating the certainty of evidence and moving from the evidence to a recommendation or decision. Therefore, the GRADE guidelines are highly relevant for systematic review authors, health technology assessment, and clinical practice guidelines developers [1]. The assessment of the certainty of the evidence is presented in GRADE Summary of Findings tables or GRADE Evidence Profiles, together with absolute effect estimates for relative effects [2,3]. A recent methodological systematic review showed that review authors might have difficulties calculating absolute effects for time-to-event outcomes [4].

Analyses that assess the time to a given event for one or several groups of patients are used in clinical studies in some fields, in particular, oncology. These time-to-event analyses are valuable, particularly when the event of interest can occur at any point over an extended period of time and the time till event occurrence carries important value. A distinct feature of time-to-event analytic techniques is to

* Corresponding author. Tel.: +49221 478 96651; fax: +49221 478 96654.

E-mail address: nicole.skoetz@uk-koeln.de (N. Skoetz).

https://doi.org/10.1016/j.jclinepi.2019.10.015

What is new?

Key findings

- The GRADE Working Group describes the preferred approach for presenting absolute effects for time-to-event outcomes in Summary of Findings tables or Evidence Profiles and provides guidance on how to avoid common pitfalls.
- Systematic review authors should be cautious whether as event mortality (e.g., people being dead) or survival (e.g., people who are alive) was used in the considered publications.
- In most cases, the absolute effect will be estimated indirectly from the hazard ratio and an adequate baseline risk. If the estimation uncertainty of the baseline risk is a relevant source of the total estimation uncertainty, it should be taken into account in the estimation of the absolute effect.

What is the implication and what should change now?

- Systematic review authors and guideline developers are advised to use the herein presented approaches to derive and present absolute effects of time-to-event outcomes, in order to support clinical decision-making and healthcare recommendations whenever they use the GRADE approach.

incorporate the censored information, which refers to information from study participants who did not experience the event of interest during the follow-up period. To compare the effects of different interventions/management strategies on time-to-event outcomes between two groups, hazard ratios (HRs) with corresponding confidence intervals derived from Cox regression models are routinely calculated as the relative effect measure.

Although their use is not limited to the field of oncology, the resulting Kaplan Meier curves, also referred to as survival curves are closely associated with oncology. For patients with cancer, one of the most relevant outcomes is overall survival (OS). Progression-free survival (PFS), disease-free survival (DFS), and event-free survival are also often assessed outcomes as they provide complementary information to OS. In addition, time-to-event analyses can describe outcomes other than survival, such as time to hospital admission, time to passage of a ureteral stone, or time to the occurrence of specified adverse events. These examples are time-to-event outcomes, as they involve the assessment of both whether a particular event occurs, and also when it occurred [5].

Absolute effect estimates (i.e., risk difference, the number needed to treat) provide important supplementary information to relative effect estimates by considering the control event rate over a given time period. As they take into account the underlying baseline risk for the event of interest in the study groups, absolute effect estimates are less vulnerable to exaggerated effect interpretation than relative effect estimates and allow a more appropriate assessment of the clinical relevance of effects [6]. Especially the absolute difference of events in both arms for one outcome at specific time points is essential for decision making and are a routine part of GRADE Summary of Findings tables and GRADE Evidence Profiles. They are automatically calculated by GRADE's official app GRADEpro GDT (gradepro.org) [2,3]. The formula for calculation of the absolute effects varies depending on whether the relative effect estimate is a risk ratio or hazard ratio [7].

As mentioned above, a recent methodological systematic review showed that less than 30% of oncological Cochrane Reviews calculated absolute effects for time-to-event outcomes correctly and presented results in an easily interpretable way [4]. The main source of error is the confusion around whether the study authors describe the proportion of participants with a given event (e.g., death from any cause) or the proportion of participants who are event-free (e.g., overall survival). Furthermore, interpretation of results in the Summary of Findings tables was hampered by the lack of explanation of which baseline risk (BLR) review authors used to estimate the absolute effect or by entering incorrect numbers like the number of events instead of numbers of patients being event-free into the GRADEpro GDT software.

Given the above-described confusion and the lack of written GRADE guidance available on how to calculate absolute effects for time-to-event outcomes based on HR and how to avoid common pitfalls, the members of the GRADE time-to-event Working Group developed this guidance incorporating feedback from methodologists and stakeholders. The findings from a methodological review evaluating the presentation of absolute effects from time-to-event data in 97 cancer-related Cochrane reviews was presented first at the GRADE meeting in 2017, in Rome, Italy. This meeting was followed by two small group discussions during the GRADE biannual meetings in Cape Town, South Africa, 2017, and Bogota, Colombia, 2018, and one large group discussion in Manchester, UK, 2018, involving more than 80 international experts, where the paper was formally approved. To prepare the presentations and incorporate the feedback from the meetings, the group of authors met in a 60-minute online conference and came to a consensus regarding this GRADE guidance.

2. Direct calculation of the absolute effect

It should be noted that when individual participant data are available and if the risk difference is an appropriate measure of the meta-analysis, the absolute effect of an intervention in individual trials can and should be estimated directly based upon the individual participant data and not indirectly via the estimates of the hazard ratio and the baseline risk. Therefore, in this case, not only is the estimation uncertainty of the hazard ratio taken into account, but the full uncertainty of the absolute effect estimate is also automatically taken into account.

In studies of time-to-event data, there is usually a staggered entry of patients into the study leading to varying follow-up times and censored observations. Sometimes, studies have a recruitment period with a staggered entry (say, for 1 year) and a fixed follow-up period for all patients, say for 2 years. In this case, you have complete observations for a period of 2 years. In the case of a single study with individual participant data and complete observation for all patients at least for a minimum time period, a specific time point with complete observations should be chosen, and the corresponding 2 × 2 table should be prepared. The usual methods for binary data can be applied to yield appropriate point and interval estimates for the risk difference [8].

In the case of staggered entry of the patients over the whole study duration, no adequate time period with complete observation may be available. In the case of a single study with individual participant data and incomplete observation, reviewers should apply methods for survival data. Methods based upon Kaplan-Meier curves [7,9] and the Cox regression model [10−12] are available to perform point and interval estimates for the risk difference at different time-points.

Sometimes it might be useful to choose the risk difference as effect measure for the meta-analysis (e.g., in the case of rare events and empty cells). In this case, a pooled risk difference by means of the usual meta-analytic methods represents an adequate measure of the absolute effect [13]. In all other cases, the estimation of the absolute effect should be performed indirectly from the pooled HR, and adequate estimation of the baseline risk.

3. Indirect calculation of the absolute effect

3.1. Assumptions for this guidance paper

For calculations of absolute effects from a pooled hazard ratio (HR), we assume that the latter is correctly calculated and applicable in the considered situation. Besides unadjusted HRs, the HRs adjusted for prognostic factors can also be used if the adjustment is performed adequately for the considered research question.

3.2. Assumptions to estimate baseline risks

The baseline risk used to calculate absolute effect size estimates should be appropriate for the individuals and their characteristics to which it is intended to be applied to. Typically, the calculation of absolute effects in systematic reviews is based on the baseline risk from included trials. However, trials could enroll individuals with a higher than average baseline risk as a way to increase their statistical power, or they could include patients with a lower than average baseline risk, as patients with comorbidities might have been excluded.

3.3. Use of the baseline risk from an external source

Large, representative observational studies at low risk of bias or systematic reviews of those studies may provide adequate baseline event rates. This approach has been previously reported for binary outcomes using appropriate observational studies, with defined prognostic markers for different risk groups. If an appropriate estimate for the baseline risk with 95% confidence interval (CI) is available from an external source, for example, from an observational study or registry, it is possible to estimate the absolute effect by taking the uncertainties of the HR and the baseline risk estimates into account (see section 3.5). For representing multiple risk groups in the population, studies with different baseline risks could be grouped accordingly (i.e., into risk groups like high, moderate, and low). For each risk group, the baseline risk estimates of representative studies could then be used to calculate the corresponding absolute risks in the intervention arm. It must be noted that systematic review authors should not extrapolate the HR beyond the follow-up period that it represents. For example, if the (pooled) HR is calculated for a follow-up period of 1.5 years, a baseline risk at 1 year from an eligible observational study may be suitable to estimate a corresponding absolute risk at 1 year. Whereas, the same HR should not be extrapolated and applied together with a baseline risk for 5 years to estimate an absolute risk at 5 years. This is because we have evidence that HR is constant only within the period of 1.5 years. After this period, the HR could potentially increase or decrease.

3.4. Use of the baseline risk from the control groups of the included studies

If no suitable observational data are available to estimate the baseline risk, data from Kaplan-Meier survival curves from the control groups of the trials included in the systematic review may be used to estimate the baseline risk. An option here is to select the curve from a trial representative for the control group of interest that is estimated to be at low risk of bias. It is as well an option to choose the curves from multiple trials representing different baseline risk groups (e.g., high, moderate, and low). Again, as mentioned for observational studies, trials with different baseline risks could be grouped, and effect estimates of representative trials for each risk group could be used to calculate the absolute effect for the intervention arm.

Oftentimes, toward the end of the reported observation time, only a small number of patients may still be at risk, with most patients having either experienced the event of interest or being censored. Therefore, the review authors should ideally choose a time point from the middle of the observation time of the respective Kaplan-Meier survival curve rather than at the end. This recommendation requires, however, that a sufficient number of events has happened up to the chosen time. In case there is a high degree of follow-up after this time point, meaning that none or only a few individuals are censored for a later eligible time point, it is possible to choose a later time point, where a larger number of events may have occurred. The chosen time point should be consistent across the different risk groups and clearly reported. Here again, it is important to point out that the HR should not be extrapolated and combined with a baseline risk estimate for a time-period that it does not represent (see section 3.3).

Sometimes, trials included in a meta-analysis, report on HRs only without presenting survival curves and survival rates at specific time points. In this case, no adequate estimates for the control group risk can be extracted from these trials and the observational data should be used to estimate the control group risk.

3.5. *Uncertainty of the baseline risk estimate*

Comparable to guidance for the calculation of absolute effects for binary data [14], only the uncertainty of the pooled HR is taken into account when grading the certainty of the body of evidence, not the uncertainty of the time point from the Kaplan-Meier survival curve and the corresponding baseline risk. The calculation of absolute effects is, therefore, conditional, based on the assumption that a given baseline risk is true.

As noted above, the baseline risk to estimate the absolute effect comes ideally from appropriate large, representative observational studies at low risk of bias. If this study is large, the standard error of the baseline risk estimate may be quite small, so that this uncertainty is negligible. In this case, the methods described in the next section can be used to estimate the absolute effect by using the baseline risk from the observational study.

However, in settings in which it appears important to take the uncertainty of the baseline risk estimate into account, which could be when the uncertainty of the baseline risk is a relevant source of the total uncertainty [15], a general method called Propagating Imprecision (PropImp) can be used to estimate the absolute effects [16]. Preconditions are that the baseline risk estimate comes from a source that is independent of the meta-analysis and that adequate point and confidence intervals are available for the baseline risk and the pooled HR. The computationally intensive PropImp approach is described in detail elsewhere, and an MS Excel sheet can be made available to facilitate implementation [16].

If large, representative observational studies at low risk of bias are not available, trials included in the meta-analysis may then provide the estimates of baseline risks [14]. In this case, the uncertainties of the baseline risk and the relative effect are correlated. Thus, only complex methods, including resampling, are available to take the uncertainty of the baseline risk into account to construct a valid confidence interval for the absolute effect [17]. However, it is not always necessary to take the uncertainty of the baseline risk estimate into account. If the standard error of the baseline risk estimate is small and the standard error of the pooled HR is the main source of the total uncertainty, the uncertainty of the baseline risk estimate is negligible. Under these circumstances, we can also take up the conditional view. Especially if we calculate the absolute effect for different risk groups, it makes sense to present the various absolute effects conditional on the corresponding assumed baseline risks. In this case, it is sufficient to take only the uncertainty of the pooled HR into account.

3.6. *Transparent reporting*

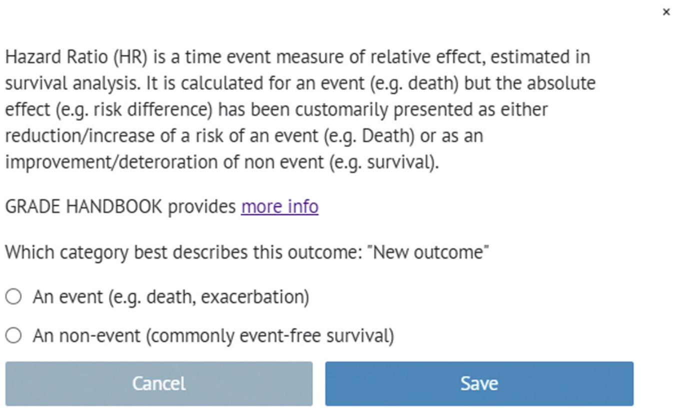
As suggested by Santesso et al. [18], transparent reporting of where baseline risk data come from is very important. It should be clearly described in the explanatory footnotes where the baseline risk comes from and which specific time point has been chosen. The time-to-event outcome and the corresponding absolute effects in the Summary of Findings table or Evidence Profile should be labeled in a consistent manner throughout the review (e.g., in the abstract, methods, and results section). The reviewers need to make a clear distinction between people who are event-free (e.g., people alive at a specific time point) and people with an event (e.g., people dead at a specific time point). If both, events and absence of events are reported in different sections of the review, a clear explanation is needed to avoid confusing the reader.

The calculated absolute effects should be reported in the Summary of Findings table and in addition at least in the abstract [19], as absolute effect estimates are more understandable to patients, clinicians, and other users of evidence syntheses than relative effect measures and are the recommended effect measure to communicate risks [20].

The specific time point of the baseline risk, which was used to calculate an absolute effect size estimate, should be provided rather than time ranges. Sometimes, review authors use the total number of events observed across several included trials with different follow-up durations to inform the baseline risk to estimate the corresponding absolute effect size estimate. This is not helpful to users since clinical decision-making is based on effect size estimates at a certain time point (e.g., 5 years or 60 months), and absolute effect size estimates will vary greatly depending on the time-point chosen.

4. Estimating and presenting absolute effects

First, we suggest to clearly define what is meant by event (e.g., people being dead) or by event-free survival (e.g.,



×

Hazard Ratio (HR) is a time event measure of relative effect, estimated in survival analysis. It is calculated for an event (e.g. death) but the absolute effect (e.g. risk difference) has been customarily presented as either reduction/increase of a risk of an event (e.g. Death) or as an improvement/deteroration of non event (e.g. survival).

GRADE HANDBOOK provides more info

Which category best describes this outcome: "New outcome"

○ An event (e.g. death, exacerbation)

○ An non-event (commonly event-free survival)

Fig. 1. Options to determine the definition and category of the event of interest (event [cumulative incidence] or non-event [survival]) of a time-to-event outcome in the GRADEpro GDT software, which is used to create Summary of Findings tables and Evidence Profiles.

people who are alive) and to estimate the desired proportion by labeling clearly whether this is the proportion of patients with event or patients being event-free (please see Fig. 1).

4.1. *Calculations of absolute effects for event-free survival (e.g., overall survival, progression-free survival)*

Calculation of absolute effects is based on methods as described by Tierney et al. [5] under the assumption of proportional hazards. Let p_i, $i = 0{,}1$, be the proportion of event-free patients up to a given time point in the control ($i = 0$) and intervention group ($i = 1$), respectively, and HR the hazard ratio for the comparison of the hazard between the intervention and the control group (intervention/control). Then the proportion of event-free patients in the intervention group can be calculated as:

$$p_1 = exp(ln(p_0) \times \mathrm{HR}) = p_0^{\mathrm{HR}}.$$

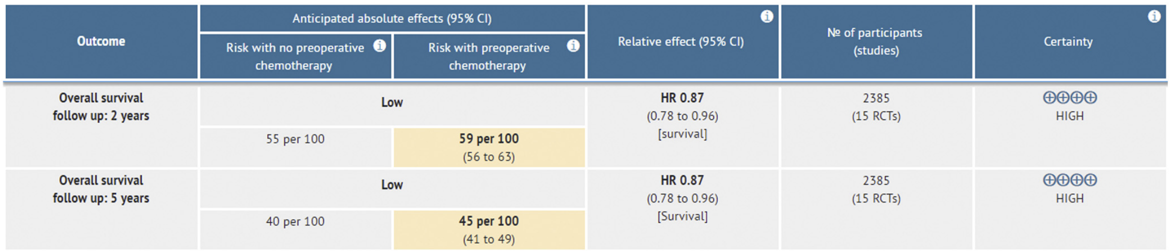
As an example, a pooled HR of 0.42 (95% CI 0.25 to 0.72) is used, indicating a lower risk of death over time in the intervention group. Estimating a proportion of patients with event-free survival in the control group at the time point 2 years of 0.9 we obtain:

$$p_1 = exp(ln(0.9) \times 0.42) = .0.9^{0.42} = 0.957.$$

This means that 96 of 100 people with this disease will be alive with the experimental intervention at 2 years. Then, the upper and lower confidence limits for the corresponding intervention risk are obtained by replacing HR by their upper and lower confidence limits, respectively (e.g., replacing 0.42 with 0.25, then with 0.72, in the example above), according to the substitution method of Daly (please see Fig. 2) [22].

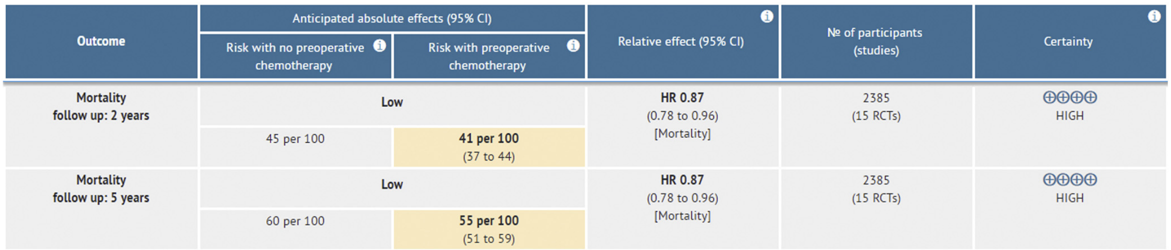
4.2. *Calculation of absolute effects for events (e.g., mortality)*

For obtaining absolute effects for time-to-event outcomes reported as events, such as mortality, a similar formula can be used. Let r_i, $i = 0{,}1$, be the proportion of patients with event up to a given time point in the control ($i = 0$) and intervention group ($i = 1$), respectively (i.e.,



Outcome	Anticipated absolute effects (95% CI)		Relative effect (95% CI)	№ of participants (studies)	Certainty
	Risk with no preoperative chemotherapy	Risk with preoperative chemotherapy			
Overall survival follow up: 2 years	**Low**		**HR 0.87** (0.78 to 0.96) [survival]	2385 (15 RCTs)	⊕⊕⊕⊕ HIGH
	55 per 100	**59 per 100** (56 to 63)			
Overall survival follow up: 5 years	**Low**		**HR 0.87** (0.78 to 0.96) [Survival]	2385 (15 RCTs)	⊕⊕⊕⊕ HIGH
	40 per 100	**45 per 100** (41 to 49)			

Fig. 2. Example: Calculations for event-free survival (overall survival) at two time points, based on an example in lung cancer patients [21]. In this example, an HR < 1 favors the intervention group, so more people will be alive in the intervention arm compared to the control arm. Please note that the term ''risk'' in the column headings misleadingly addresses the ''risk'' of surviving.



Outcome	Anticipated absolute effects (95% CI)		Relative effect (95% CI)	№ of participants (studies)	Certainty
	Risk with no preoperative chemotherapy	Risk with preoperative chemotherapy			
Mortality follow up: 2 years	**Low**		**HR 0.87** (0.78 to 0.96) [Mortality]	2385 (15 RCTs)	⊕⊕⊕⊕ HIGH
	45 per 100	**41 per 100** (37 to 44)			
Mortality follow up: 5 years	**Low**		**HR 0.87** (0.78 to 0.96) [Mortality]	2385 (15 RCTs)	⊕⊕⊕⊕ HIGH
	60 per 100	**55 per 100** (51 to 59)			

Fig. 3. Example: Calculations for events (mortality) at two time points, based on an example in lung cancer patients [21]. In this example, an HR < 1 favors the intervention group, so fewer people will be dead in the intervention arm compared to the control arm.

r_0 is the baseline risk), then risk of an event in the intervention group can be calculated by

$$r_1 = 1 - exp(ln(1-r_0) \times HR) = 1 - (1-r_0)^{HR}.$$

Fig. 3 gives an example of the presentation in GRADEpro.

4.3. Graphical presentation

For supporting the interpretation of systematic review results, the GRADEpro software provides the opportunity to present review findings graphically in an interactive summary of findings table [23]. A feature of this format allows visualizing a corresponding absolute effect for the comparison of an intervention arm to a control arm for each outcome. In six steps, the absolute number of events for a specific time point in the control group (the baseline risk), the estimated number of events in the intervention group, the risk difference and the associated statistical uncertainty are presented in an easily comprehensible way (please see Fig. 4 for an example).

4.4. Calculation of numbers needed to treat based on events or event-free survival

Numbers needed to treat with confidence intervals can also be calculated as the inverse of the risk differences between intervention and control arm [24].

Risk difference: control group risk—intervention group risk (95% CI [control group risk—upper CI; control group risk—lower CI])

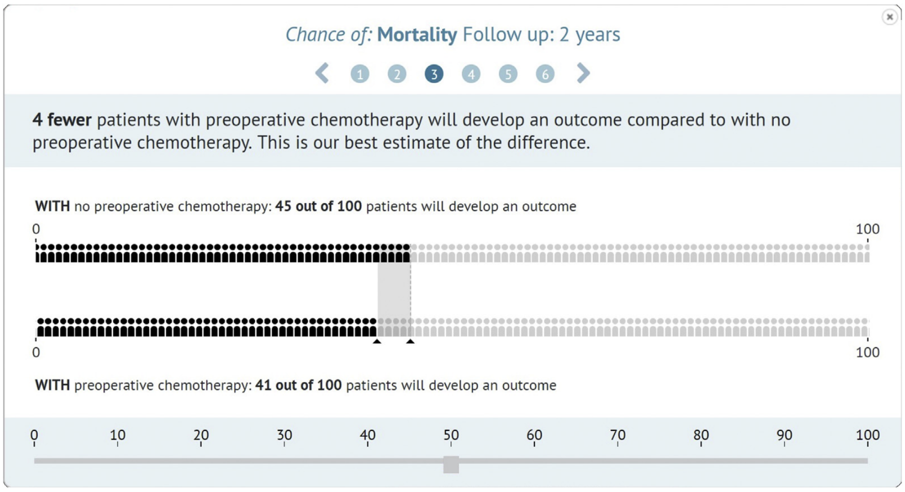
Example from above for events (mortality at 2 years)

Risk difference: 45/100 (control group)—41/100 (95% CI 37 to 44) (intervention group) = 4/100 (95% CI 1/100 to 8/100)

1/Risk difference = 25 (95% CI 12.5 to 100)

Meaning that 25 (13 to 100) people need to be treated to avoid one death at 2 years.

Similar to the afore outlined calculations of absolute effects utilizing the HR and corresponding baseline risk, the number needed to treat is strongly depending on the size of the chosen baseline risk [25]. Therefore, here, we propose to present the numbers needed to treat and the corresponding upper and lower confidence intervals



Chance of: **Mortality** Follow up: 2 years

4 fewer patients with preoperative chemotherapy will develop an outcome compared to with no preoperative chemotherapy. This is our best estimate of the difference.

WITH no preoperative chemotherapy: **45 out of 100** patients will develop an outcome

WITH preoperative chemotherapy: **41 out of 100** patients will develop an outcome

Fig. 4. Example: Graphical presentation of the absolute number of events in the control and the intervention arm at 2 yr in the interactive Summary of Findings table, based on an example in lung cancer patients [21].

across a range of baseline risks to represent different risk groups.

4.5. Calculation of median survival time

Calculation of median (event-free) survival time while applying the HR is one of the options presented in the paper by Tierney et al. [5] for individual trials. This option might be of great interest to patients, physicians, and stakeholders for clinical decision-making, but user testing is needed. One necessary condition is that the median survival time has been reached in the control group, meaning that for overall survival, 50% of the patients at risk already died. For obtaining the median survival time in the intervention group (MST_1) from the median survival time in the control group (MST_0) and the pooled HR, the following formula can be used (the calculation is based upon the assumption that MST_0 is fixed):

$$MST_1 = MST_0/HR$$

As an example, we consider the pooled hazard ratio of HR = 0.42 (95% CI 0.25 to 0.72). In this case, $HR < 1$ is defined as favoring the intervention arm. Assuming a median survival time in the control group of 80 months, we obtain:

$$MST_1 = 80\,\text{months}/0.42 = 190.5\,\text{months}$$

Again, only the uncertainty of the HR is taken into account, not that of the median survival time. Upper and lower confidence limits for the corresponding intervention risk are obtained by replacing HR by their upper and lower confidence limits, respectively (e.g., replacing 0.42 with 0.25, then with 0.72, in the example above).

The difference of the median survival times between the intervention and the control group can be calculated by $MST_1 - MST_0 = 190.5$ months -80 months $=$ 110.5 months.

5. Summary

Absolute effect estimates, especially absolute risk differences, provide essential information to guide clinical decision-making and the formulation of healthcare recommendations.

For time-to-event outcomes, the GRADE approach focuses on absolute effect estimates that are calculable from a hazard ratio and an applicable baseline risk as these will most frequently be available to the systematic review and guideline authors. Thus, GRADE focusses on risk differences and, on occasion, the number needed to treat or median survival times. We here present several approaches that are suitable to calculate the corresponding estimates and accord to the available data. In situations were sufficient data (e.g., IPD) or complete information for all study participants for a fixed follow-up duration is available, we advise review authors to use direct estimation methods, which are outlined in this document. As these data are often not available, we also guide review or guideline authors on how to calculate absolute effects indirectly.

When calculating absolute effect estimates, review authors must consider the direction of the effect (which intervention is favored with an $HR < 1$?) and whether the cumulative incidence of the event or event-free survival is reported, as given by the definition of the outcome. Authors, as well as users of systematic reviews, should be aware of potential mistakes in the calculation of absolute effects and should include the direction of the relative effect into their judgment.

6. Further considerations and unresolved issues

The GRADEpro GDT software has been adapted to provide systematic review authors and guideline developers the opportunity to choose from the number of people with a given event or without an event at a specific time point when presenting absolute effect size estimates. This allows consistency of reported outcomes throughout the review and lets authors and guideline developers choose the format that seems most suitable to questions at hand.

In this guidance paper, we focused only on the correct calculation of absolute effects and interpretation of the direction of effect—event vs. event-free survival. There are a number of unresolved issues related to meta-analyses of time-to-event outcomes and grading the certainty of the evidence body. The GRADE Working Group is aiming to address the following issues in subsequent guidance:

- Time-to-event outcomes have features that typically incorporate observations based on censoring [26]. Further challenging aspects are to assess the certainty of the evidence for censoring mechanisms that are not independent of the outcome leading to a potential risk of bias.
- Treatment-switching is nowadays common in cancer trials, which also might introduce bias in time-to-event analyses. Assessment of this bias is particularly difficult as the time points of switching are usually not given. How to grade the certainty of the evidence in case of treatment switching will be elucidated in another guidance paper
- In competing risk settings, sometimes Kaplan-Meier survival analyses are performed, which might overestimate a potential effect [27], which will also be the focus of another paper.
- In cases where the proportional hazards assumption is invalid, alternative effect measures to the HR, such as the difference of the restricted mean survival time (RSMT) between the groups, can be used [28].

CRediT authorship contribution statement

Nicole Skoetz: Writing - original draft, Methodology, Writing - review & editing, Project administration, Conceptualization, Supervision. **Marius Goldkuhle:** Writing - original draft, Writing - review & editing, Methodology, Project administration, Conceptualization. **Elvira C. van Dalen:** Writing - original draft, Writing - review & editing, Methodology, Conceptualization. **Elie A. Akl:** Writing - original draft, Writing - review & editing, Methodology. **Marialena Trivella:** Writing - original draft, Writing - review & editing, Methodology. **Reem A. Mustafa:** Writing - original draft, Writing - review & editing, Methodology. **Artur Nowak:** Writing - original draft, Writing - review & editing, Software. **Philipp Dahm:** Writing - original draft, Writing - review & editing, Methodology. **Holger Schünemann:** Writing - original draft, Methodology, Writing - review & editing. **Ralf Bender:** Writing - original draft, Writing - review & editing, Methodology, Project administration, Conceptualization, Supervision.

Acknowledgments

Article history: Slides presented at GRADE meetings in Rome (2017), Cape Town (2017), Bogota (2018) and Manchester (2018).

References

[1] Guyatt GH, Oxman AD, Schunemann HJ, Tugwell P, Knottnerus A. GRADE guidelines: a new series of articles in the Journal of Clinical Epidemiology. J Clin Epidemiol 2011;64:380−2.

[2] Schünemann HJ, Vist GE, Glasziou P, Akl E, Skoetz N, Guyatt GH. Chapter 14: completing summary of findings tables and grading the certainty of evidence. In: Higgins JPT, Chandler J, Cumston M, Li T, Page MJ, Welch V, editors. Cochrane handbook for systematic reviews of interventions version 6 (updated January 29, 2019). Chichester (UK): The Cochrane Collaboration; 2019:2019. Available at https://training.cochrane.org/handbooks. Accessed September 19, 2019.

[3] Guyatt G, Oxman AD, Akl EA, Kunz R, Vist G, Brozek J, et al. GRADE guidelines: 1. Introduction-GRADE evidence profiles and summary of findings tables. J Clin Epidemiol 2011;64:383−94.

[4] Skoetz N, Goldkuhle M, Weigl A, Dwan K, Lebonte V, Dahm P, et al. Methodological review showed correct absolute effect size estimates for time-to-event outcomes in less than one-third of cancer-related systematic reviews. J Clin Epidemiol 2019;108:1−9.

[5] 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.

[6] Carling CLL, Kristoffersen DT, Montori VM, Herrin J, Schünemann HJ, Treweek S, et al. The effect of alternative summary statistics for communicating risk reduction on decisions about taking statins: a randomized trial. PLoS Med 2009;6(8):e1000134.

[7] Altman DG, Andersen PK. Calculating the number needed to treat for trials where the outcome is time to an event. BMJ 1999;319(7223):1492−5.

[8] Newcombe RG. Interval estimation for the difference between independent proportions: comparison of eleven methods. Stat Med 1998;17:873−90.

[9] Bender R, Kromp M, Kiefer C, Sturtz S. Absolute risks rather than incidence rates should be used to estimate the number needed to treat from time-to-event data. J Clin Epidemiol 2013;66:1038−44.

[10] Austin PC. Absolute risk reductions and numbers needed to treat can be obtained from adjusted survival models for time-to-event outcomes. J Clin Epidemiol 2010;63:46−55.

[11] Laubender RP, Bender R. A note on calculating asymptotic confidence intervals for the adjusted risk difference and number needed to treat in the Cox regression model. Stat Med 2014;33:798−810.

[12] Laubender RP, Bender R. Estimating adjusted risk difference (RD) and number needed to treat (NNT) measures in the Cox regression model. Stat Med 2010;29:851−9.

[13] Deeks J, Higgins JPT, Altman D, on behalf of the Cochrane Statistical Methods Group. Chapter 9: analysing data and undertaking meta-analyses. In: Higgins JPT, Churchill R, Chandler J, Cumpston MS, editors. Cochrane handbook for systematic reviews of interventions version 5.2.0 (updated June 2017). Cochrane: Cochrane; 2017:2017.

[14] Guyatt GH, Oxman AD, Santesso N, Helfand M, Vist G, Kunz R, et al. GRADE guidelines: 12. Preparing summary of findings tables-binary outcomes. J Clin Epidemiol 2013;66:158−72.

[15] Spencer FA, Iorio A, You J, Murad MH, Schunemann HJ, Vandvik PO, et al. Uncertainties in baseline risk estimates and confidence in treatment effects. BMJ 2012;345:e7401.

[16] Newcombe RG. Propagating imprecision: combining confidence intervals from independent sources. Commun Stat - Theor Methods 2011;40(17):3154−80.

[17] Newcombe RG, Bender R. Implementing GRADE: calculating the risk difference from the baseline risk and the relative risk. Evid Based Med 2014;19(1):6−8.

[18] Santesso N, Carrasco-Labra A, Langendam M, Brignardello-Petersen R, Mustafa RA, Heus P, et al. Improving GRADE evidence tables part 3: detailed guidance for explanatory footnotes supports creating and understanding GRADE certainty in the evidence judgments. J Clin Epidemiol 2016;74:28−39.

[19] Agarwal A, Johnston BC, Vernooij RW, Carrasco-Labra A, Brignardello-Petersen R, Neumann I, et al. Authors seldom report the most patient-important outcomes and absolute effect measures in systematic review abstracts. J Clin Epidemiol 2017;81:3−12.

[20] Gigerenzer G, Gaissmaier W, Kurz-Milcke E, Schwartz LM, Woloshin S. Helping Doctors and patients make sense of health statistics. Psychol Sci Public Interest 2007;8(2):53−96.

[21] Preoperative chemotherapy for non-small-cell lung cancer: a systematic review and meta-analysis of individual participant data. Lancet 2014;383(9928):1561−71.

[22] Daly LE. Confidence limits made easy: interval estimation using a substitution method. Am J Epidemiol 1998;147:783−90.

[23] Schunemann HJ, Mustafa R, Brozek J, Santesso N, Alonso-Coello P, Guyatt G, et al. GRADE Guidelines: 16. GRADE evidence to decision frameworks for tests in clinical practice and public health. J Clin Epidemiol 2016;76:89−98.

[24] Veroniki AA, Bender R, Glasziou P, Straus SE, Tricco AC. The number needed to treat in pairwise and network meta-analysis and its graphical representation. J Clin Epidemiol 2019;111:11−22.

[25] Smeeth L, Haines A, Ebrahim S. Numbers needed to treat derived from meta-analyses–sometimes informative, usually misleading. BMJ 1999;318(7197):1548−51.

[26] Kleinbaum DG, Klein M. Survival analysis. A self-learning text. New York: Springer; 2012. Available at http://www.springer.com/de/book/9781441966452. Accessed September 19, 2019.

[27] Lacny S, Wilson T, Clement F, Roberts DJ, Faris P, Ghali WA, et al. Kaplan-Meier survival analysis overestimates cumulative incidence of health-related events in competing risk settings: a meta-analysis. J Clin Epidemiol 2018;93:25−35.

[28] Royston P, Parmar MKB. Restricted mean survival time: an alternative to the hazard ratio for the design and analysis of randomized trials with a time-to-event outcome. BMC Med Res Methodol 2013;13:152.

7. Discussion

7.1. Short summary

Included in this dissertation are three meta-epidemiological studies that systematically assessed the characteristics, methods and reporting of time-to-event outcomes in current evidence syntheses.

A meta-epidemiological study of 217 meta-analyses of time-to-event outcomes performed in 100 recent systematic reviews revealed great variability in the characteristics and methods between reviews and significant limitations in their reporting (167). For instance, outcome definitions were reported for less than half of the assessed review outcomes. Only limited information on general review methods, including information on the relevant types of analyses and their adjustment for covariates, were available. Review authors obtained time-to-event data from trial publications using diverse methods, with additional variability in the complexity and the number of approaches per review.

Particularly problematic appeared the consideration of trial characteristics that affect the reliability of time-to-event analyses, for example, different follow-up durations amongst trials, informative censoring, competing events, treatment switching and proportional hazards. Seldomly were these characteristics included in additional analyses, such as sensitivity analyses, or mentioned and discussed by the review authors.

A second meta-epidemiological study systematically examined the characteristics, methods and reporting of 235 trials with 315 eligible trial time-to-event analyses included in meta-analyses of time-to-event outcomes of a sample of the previously considered reviews (162). As for their including reviews, the meta-epidemiological study shows that trial time-to-event analyses are conducted with great variability in methods. They are associated with significant reporting limitations, such as for outcome definitions, censoring mechanisms and key data on trial follow-up. Great variation was also found regarding the types of time-to-event data that were available in trial publications for individual analyses (e.g., HRs, Kaplan-Meier curves or log-rank P values).

The reporting of some analytical aspects was more comprehensive in trial publications than in their including reviews. Connecting the data extracted for an individual trial with its including review showed, for example, that review authors rarely reported the types of time-to-event data they included in their reviews, even though the information was available in trial publications. Like their including reviews, trial publications only sporadically reported on characteristics relevant to the reliability of time-to-event outcomes. Trialists seldomly performed additional analyses to investigate the validity or robustness of time-to-event analyses.

A third meta-epidemiological study assessed the calculation of absolute effect measures in a sample of 96 oncological Cochrane reviews (11). Absolute effect estimates in evidence syntheses, for example, risk differences, are calculated based on a pooled relative effect measure, in form of the HR, and a baseline- or population risk.

The results of this meta-epidemiological study show that absolute effect estimates in current systematic reviews are often mislabeled, i.e., authors switch between reporting a single outcome as event (e.g., mortality) and its absence (e.g., overall survival), and often incorrectly calculated. This miscalculation resulted, in some cases, in a reversal of the actual effect direction. In addition to a considerable risk for misinterpretation by untrained users, the results of the meta-epidemiological study suggest problems of review authors in the interpretation of the effect measures from time-to-event analyses.

Addressing the hardships identified in the previous three meta-epidemiological studies, targeted guidance articles were developed according to the standards of the renowned GRADE Working Group.

One of these guidelines addressed informative censoring, which refers to the censoring of study participants through mechanisms associated with the outcome event (e.g., loss to follow-up or adverse events) and can be considered a time-to-event specific source of bias (2).

In addition to a basic, conceptual background and theoretical elaboration of the problem, the article uses an imputation approach based on the recalculation of individual participant data from a published Kaplan-Meier curve to show how informative censoring affects trial estimates. The article then provides authors of evidence syntheses with methods to detect a potential risk of bias in underlying studies, for example, by using reported information or the distribution of censoring over the trial observation period. Finally, authors are equipped with approaches to incorporate their judgments for an individual trial into risk of bias judgements for a body of evidence and subsequently into judgments about the certainty in the evidence using the GRADE approach.

A second GRADE guidance developed based on previously gathered empirical evidence addressed the calculation of absolute effect estimates for time-to-event outcomes (10). The article provides a theoretical background and illustrates multiple approaches to calculate different types of absolute effects for time-to-event outcomes, either directly, based on individual participant data, or indirectly, under use of the pooled HR and a baseline risk. The article also provides considerations about the optimal selection of variables for calculation, discusses limitations of the individual approaches and suggests alternative absolute effect estimates that might ease the communication of time-to-event outcomes in selected situations.

7.2. Strengths and limitations

Several strengths as well as limitations associated with the works in this dissertation require more detailed elaboration. Strengths and limitations related to the individual projects are also discussed in the respective publications (2, 10, 11, 162, 167).

7.2.1. Rigorous methodology

All meta-epidemiological studies contributing to this dissertation were performed based on a-priori developed protocols and followed methods that were piloted before the project start to assure their robustness and feasibility.

Noteworthy protocol deviations occurred, for example, when it was not possible to the recalculate meta-analyses of time-to-event outcomes in the course of the meta-epidemiological studies because of deficient reporting in the assessed reviews. Another deviation occurred, for instance, when it was not possible to assess the calculation of absolute effects for time-to-event outcomes in systematic reviews published in high impact factor journals, because no high-impact reviews reported absolute effect measures.

Despite such slight and justified deviations, the a-priori specified project plans were overall adhered to (176, 177).

The methods of the individual meta-epidemiological studies followed established standards for conducting systematic reviews and meta-epidemiological research (85, 89, 171). These standards include the systematic identification of the literature, a structured literature selection process according to a-priori defined in- and exclusion criteria and extraction items that were tested for robustness. All relevant steps were performed in duplicate by a second researcher or at least double-checked independently, and conflicts were resolved under support by a third researcher.

To ensure the transparency and replicability of the findings, all publications were reported in line with established requirements for systematic reviews and a specialized guideline for reporting meta-epidemiological research (85, 171).

7.2.2. Potential limitations associated with assessments of reported information

A central limitation of meta-epidemiological studies, including the studies in this dissertation, is their focus on reported information rather than directly investigating the conduct of trials or systematic reviews. In consequence, they might miss methodological errors that will often remain unpublished (23, 86).

For some of the items assessed during the meta-epidemiological studies in this dissertation, it will remain unclear whether the authors of trials and evidence syntheses truly did not perform a necessary step, conduct a certain analysis or assess a potential problem, or whether they simply did not report their performance. For example, review authors may not have specified which methods they used for the recalculation of time-to-event data from trial publications, because all their included trials reported HRs and CIs. Other authors may not have addressed informative censoring, because they did not suspect a problem in any of their included trials.

In difference to previous meta-epidemiological studies, the assessments performed within this dissertation examined reporting at both, the review level and the level of their included trials (23-25, 86, 144). In this way it was demonstrated, for example, that several methodological and reporting issues identified on trial level should have been addressed on review level but were not. Consequently, several problems evident in trials may translate to the review level.

In meta-epidemiological assessments, reporting is a surrogate for the trialists' and review authors' recognition of the issues addressed through the defined extraction items. Trial and review authors might omit certain information because they do not consider it as important in a respective context. Review authors, for example, might not report the types of trial analyses included in a meta-analysis, because only intention-to-treat analyses from their included trials were available. They might not report additional analyses for informative censoring because they did not suspect any informative censoring in their included trials.

This effect and the resulting loss of information is enhanced by strict wording limits of publishing journals which demand to prioritize the reported information. In this dissertation, these considerations primarily affect non-Cochrane reviews and trial publications. Cochrane reviews, on the other hand, are published with highly standardized reporting requirements and without a word limit (167).

7.2.3. Potential limitations associated with the selection of the literature

Cochrane reviews are often considered methodologically state-of-the-art and a gold standard amongst systematic reviews and are produced under support of a statistical methods unit. They constitute a considerably large source of data within this dissertation.

In absence of established guidance for systematic reviews with meta-analyses of time-to-event outcomes, the inclusion of a large proportion of Cochrane reviews in the first meta-epidemiological study was intended to identify possible best-practice examples (chapter 6.1). That was because the primary goal of the meta-epidemiological studies was to create an evidence base for future guidance and not to compare the quality of reporting amongst different types of reviews. In direct comparison, however, Cochrane reviews showed more complete reporting than most of the systematic reviews published in Core Clinical Journals. At least to some degree, this difference might be attributable to the absence of wording restrictions for Cochrane reviews (167).

It is reasonable to assume that the non-Cochrane reviews were also performed and published with greater rigor, or at least subject to higher requirements, than the general landscape of systematic reviews. That is because they were selected through the Core Clinical Journals filter, which limits the systematic search to the most relevant journals to physicians according to the U.S National Library of Medicine (170). As a result, several of the non-Cochrane reviews were published in journals with higher impact factors than the Cochrane Database of Systematic Reviews (167).

Because the assessed Cochrane and non-Cochrane reviews very likely represent a sample of systematic reviews with increased standard, many limitations identified in the meta-epidemiological studies might appear even more substantial in the general landscape of published systematic reviews.

7.2.4. Potential limitations associated with the applicability of results

Some limitations of the included meta-epidemiological studies relate to the applicability of their findings: first, most of the evidence presented in this dissertation was collected from systematic reviews of interventions. The eligibility criteria allowed to include systematic reviews on preventive questions as well, but other types of evidence syntheses which occasionally include time-to-event outcomes, particularly reviews of prognosis studies, were excluded. This is due to the fact that reviews of prognosis often do not include RCTs as their primary source of evidence and because they apply different requirements to the calculated parameters. Review authors might, for example, consistently favor adjusted estimates for certain prognosis questions.

Nonetheless, several of the reporting limitations revealed during this dissertation, such as limited information on informative censoring or competing events trial publications, also affect the validity of reviews of prognosis and other questions.

Second, most of the assessed systematic reviews addressed oncological questions and less often, for example, questions in cardiovascular research. Even though the meta-epidemiological study on absolute effects was limited to oncological systematic reviews, the other two studies did not impose any restrictions to clinical fields. The high proportion of oncological reviews reflects the great interest in time-to-event outcomes in oncology research.

Third, overall survival was by far the most frequently assessed type of outcome within the included meta-epidemiological studies. Despite the overreaching importance as an outcome to patients and care givers alike, overall survival as a study outcome has unique properties: it is particularly robust, its timing will often be known or traceable from death records, it is objective and it is not susceptible to competing events (178). These properties and the general importance of the outcome must be considered when interpreting some of the here presented findings. They might lead, for example, to a more rigorous and complete reporting of the outcome in systematic review and trial publications. In consequence some of the findings may overestimate the conditions for other types of outcomes.

7.2.5. Strengths and limitations associated with the presented guidance articles

The guidelines developed as part of this dissertation fulfill all requirements imposed by the GRADE Working Group (142, 143). Their development followed an evidence-based approach, oriented on distinct problems in evidence syntheses with time-to-event outcomes. The development process involved continuous discussions with a large group of experienced and internationally renowned scientists and required, after large group disputations and several open phases to comment on the articles, acceptance by the entire GRADE Working Group, which consist of more than 200 individuals.[26]

This systematic, interdisciplinary and rigorous approach guarantees that the given recommendations are accurate and applicable. Furthermore, it guarantees their wide reach. As of June 2023, the guidance on the calculation of absolute effects for time-to-event outcomes and the guidance on informative censoring have been cited 54 and 6 times according to the journal websites (2, 10).

Like their evidence base, provided through the meta-epidemiological studies, the applicability of the guidelines is somewhat limited to evidence from RCTs, particularly the GRADE guideline on informative censoring. That is because some of the suggested approaches to detect informative censoring in trial publications rely on randomization and its effect on the distribution of censoring in the assessed groups. Despite this restriction, several of the principles that are highlighted in the guidance on informative censoring, for instance, the conceptual reasoning as well as the suggestions for deriving decisions, are largely applicable to other types of evidence syntheses.

The guidance for the calculation of absolute effect measures is applicable to and advised for all types of evidence syntheses which include a pooled HR.

In the future, some of the suggestions in the guidelines might become obsolete, for example, because of methodological developments or because of changing reporting and publication practices. Under such circumstances, publications within the series of GRADE guidelines can be updated.

7.3. Similar works

Detailed discussions of similar works for each individual sub-project are included in the respective publications (2, 10, 11, 162, 167).

7.3.1. Previous research on methodological problems in trials and their including evidence syntheses

The two meta-epidemiological studies that addressed the characteristics, methods and reporting of time-to-event outcomes, are the first to assess such data on the level of systematic reviews and their included trials. In comparison to previous assessments, which address either the trial or review level, such an assessment allowed for more in-depths conclusions and enabled to show that reporting issues could translate from trial publications to review publications.

In a previous study, Kahale et al. 2020 (179) chose a similar two-level approach to assess how missing outcome data in RCTs is handled in meta-analyses of binary outcomes in systematic reviews. According to their findings, and consistent with the here presented results for time-to-

[26] Somebody at a GRADE meeting described the acceptance process of a GRADE guideline to me as "the hardest peer-review process there is", which might be true, at least for the clinical epidemiology literature.

event outcomes, an approach to missing outcome data was explained only in about a third of review publications and only in a small proportion considered in additional review analyses, such as sensitivity analyses.

Determining a potential risk of bias through missing outcome data for binary outcomes differs from considerations and approaches necessary for time-to-event data. This is because, as described previously in the background of this dissertation (chapter 3.2.1), individuals who are lost to follow-up in time-to-event analyses can simply be censored without reducing the total number of analyzed individuals. In consequence, individuals who are censored for potentially informative reasons cannot be distinguished from participants who provide end-of-study censored data, if not explicitly reported in trial publications.

For the here presented meta-epidemiological studies, only reported information on missing outcome data in time-to-event analyses was extracted. Some of the cases where missing outcome data was judged as “not reported” could indeed have censored affected participants (162). Specific approaches that allow to determine a risk of bias through informative censoring in absence of explicitly reported information are described in the GRADE guideline that is part of this dissertation (2).

7.3.2. Previous research on the reporting of time-to-event analyses in trials

Several previous meta-epidemiological assessments studied time-to-event analyses in study and trial publications, although none of these assessments focused on trials and studies actually included in meta-analyses (23-25, 144, 160, 161, 180, 181).

The results of these assessments support the findings of variable and sometime insufficient reporting of trial time-to-event analyses seen in the here included meta-epidemiological studies. They found, for example, that important data on follow-up, including the start and ending points as well as the overall duration, and data on censoring, including the reasons and number of participants, are often omitted in trial and study publications (23-25, 144, 161, 162). Also in accordance with the here presented results is the common observation that important data on statistical analyses, such as on model building and on time-to-event specific assumptions, are often not provided (24, 144, 162).

The meta-epidemiological studies presented in this dissertation extend these findings by demonstrating that several of the previously indicated problems also translate to review publications (162).

7.3.3. Previous research on time-to-event analysis challenges in trials

There are several methodological studies that address specific assumptions and methodological hardships associated with time-to-event analysis in trials and studies, in particular (non-) proportional hazards, informative censoring and competing events.

Rulli and colleagues in 2018 (44) assessed in how many of the currently published RCTs in oncology the proportional hazards assumption of the performed time-to-event analyses fails. For their assessment, they recalculated individual participant data from Kaplan-Meier plots of a sample of RCTs that compared interventions for lung cancer by using the approach of Guyot et al. 2012 (145). In almost every fifth of these trials, they found that the assumption of proportional hazards did not apply, particularly in trials comparing interventions with different treatment mechanisms (44).

With a similar approach, Royston and colleagues in 2018 (182) found a failure of the proportional hazards assumption in about a third of treatment comparisons with time-to-event outcomes in publications of phase III clinical trials published in four central medical journals (182).

The here presented meta-epidemiological studies revealed that only in a minor proportion of trials included in meta-analyses of time-to-event outcomes, the trialists assessed the assumption of proportional hazards. Of the trialists who assessed the assumption, several did not report or interpret their findings. None of the assessed systematic reviews reported an assessment of the assumption in trials included for meta-analysis.
As previously described (chapter 3.4.3), a failure of the proportional hazards assumption results in time dependency of the HR and might affect the certainty in pooled estimates from meta-analyses, for example, through introducing between-study heterogeneity. Literature on how a failure of the proportional hazards influences the results of meta-analyses is, however, currently lacking.

Rosen et al. 2020 (183) investigated and compared the distribution of censored individuals in trial groups over time, using an approach similar the method to detect potential informative censoring in trial publications described in the GRADE guideline that is part of this dissertation. They systematically included several oncological RCTs that were published in a high impact journal. Eligible RCTs reported overall and progression-free survival of the compared groups as Kaplan-Meier plots and presented censored observations in the groups over time.
By comparing the distribution of censored individuals in each plot, Rosen et al. 2020 (183) found a higher proportion of censored individuals in the comparator group at earlier follow-up, but a higher proportion of censored individuals in the experimental group during later follow-up in a considerable number of their assessed RCTs. Based on their observations, they concluded that a risk of bias may exist in several of the assessed RCTs, should individuals in the control group have been excluded from or left the group informatively and differentially to the experimental group (183).

The analysis of time-to-event outcomes that are susceptible to competing events requires, as previously described in chapter 3.2.3, certain adaptations of the statistical analysis. Regular Kaplan-Meier estimates with censoring of competing events, for example, will be biased and it is advised to use cumulative incidence functions instead. Several previous studies assessed the prevalence and the potential impact of ignoring competing events when analyzing susceptible time-to-event outcomes in RCTs.
Van Walraven and colleagues in 2016 (59), for instance, found a risk for overestimation of intervention effects in almost half of a random sample of 100 trials which reported Kaplan-Meier estimates for outcomes susceptible to competing events and were published in major clinical journals. For a large proportion of their assessable trials, they discovered a potential upwards bias through inadequate Kaplan-Meier analysis of at least 10%.
Another study by Schumacher and colleagues from 2016 (58) also found a potential risk of bias through inadequate consideration of competing events for almost half of trials published in the New England Journal of Medicine in 2015 and that used time-to-event analysis methods for an outcome susceptible to competing events.

These studies indicate that potential bias through inadequate handling of competing events is particularly prominent in trials addressing questions on cardiovascular diseases. Because oncology outcomes such as progression-, disease- and event-free survival include the competing event death by definition (see 3.2.3), the problem is less prominent in oncology trials. During the meta-epidemiological studies included in this dissertation, only a smaller proportion of the assessed systematic review outcomes, and accordingly of the assessed trials, were susceptible to competing events, which results from the circumstance that most reviews addressed oncological questions and the outcome overall survival.

Still, consistent with previous research, the meta-epidemiological studies of this dissertation found that competing event specific analyses were only seldomly reported for trials with outcomes susceptible to competing events (162). Similarly, only few of the systematic reviews which actually included time-to-event outcomes at risk of bias through competing events mentioned them in their reviews, in risk of bias assessments or discussions. None of these reviews performed additional analyses for competing events (167).

7.3.4. Previous research on the quality and reporting of evidence syntheses

Currently available assessments of systematic reviews address general review characteristics, their reporting and their quality, but there are no assessments focusing on reviews including time-to-event outcomes (86, 176, 177, 184).

The available assessments highlight substantial room for improvement, for example, regarding the consistent conduct of quality/ risk of bias assessments, the reporting of methods, the use of additional analyses, such as sensitivity analyses, and the consideration of trial characteristics with influence on the validity of meta-analyses (86, 177).

A systematic assessment of the methodological quality of systematic reviews in oncology published in high-impact journals and the Cochrane Database of Systematic Reviews confirmed the often limited reporting of relevant review methods and additional analyses (176).

These findings are in agreement with the conclusions of the meta-epidemiological study of systematic reviews in this dissertation, which also demonstrated problems in the reporting of general review methods, the conduct of additional, supporting analyses and the consideration of trial characteristics that affect the validity of meta-analyses (167).

7.3.5. Previous research on absolute effects for meta-analysis results

As recognized in the associated publication, the methodological study on absolute effect measures of time-to-event outcomes in cancer-related systematic reviews was one of the first studies to assess the correctness of absolute effects in evidence syntheses (11).

Previous research looked at how often absolute effects were used to communicate meta-analysis findings in general and found that, except for reviews which include a GRADE Summary of Findings table and in particular Cochrane reviews, only few systematic reviews actually include respective estimates. In line with the results of the here presented meta-epidemiological study, it found that critical information about the baseline risk, which is necessary for the calculation of absolute effect estimates from relative effects, is often absent (11, 185, 186).

An assessment of the number-needed-to-treat in publications assessing pharmacological interventions in high-impact journals, showed that meta-analyses were most frequent amongst the few articles which presented a number-needed-to-treat. Yet, over half of them calculated the measure with inappropriate methods (187).

An assessment of RCTs in leading medical journals revealed that absolute effects in form of the number-needed-to-treat were derived with erroneous calculation methods in almost half of cases, if calculated for time-to-event outcomes at all (188).

These findings are even more significant than the rate of 13% of Cochrane reviews with incorrect absolute effect measures identified in the meta-epidemiological study in this dissertation (187).

7.3.6. Available guidelines for time-to-event outcomes in evidence syntheses

More general aspects of evidence synthesis with time-to-event outcomes are described in the Cochrane Handbook, which also includes a chapter dedicated to the performing meta-analyses of time-to-event outcomes (89). Another central resource for the methods behind meta-analyses of time-to-event outcomes is the collection of recalculation methods for time-to-event trial data provided by Tierney and colleagues in 2007 (106). Unfortunately, a comprehensive guide on time-to-event specific hardships in evidence syntheses, how to identify problems and incorporate them in analyses and conclusions, is currently lacking.

As previously indicated, the explanation of the Risk of Bias 2.0 tool briefly mentions informative censoring amongst potential sources of bias due to missing outcome data (71).[27] A detailed explanation of the issue, its identification and incorporation into certainty of evidence ratings is, however, missing. The GRADE guidance on informative censoring provided as part of this dissertation seeks to fill this gap. It complements other related GRADE guidance dealing with rating potential bias through missing outcome data for binary and continuous outcomes (126). Likewise, the GRADE guidance on the calculation of absolute effect estimates, which was also developed within this dissertation, contributes to the slowly growing body of guidance for meta-analyses of time-to-event outcomes. A similar guidance has previously only been available for binary outcomes (136).

For other time-to-event specific hardships in meta-analyses guidance is either available in form of single conceptual articles or completely absent. Future efforts must fill these gaps.

7.4. Implications

Project specific implications are discussed in the respective publications (2, 10, 11, 162).

7.4.1. Implications for conductors of evidence syntheses

7.4.1.1. Guidelines for systematic review reporting

The projects included in this dissertation highlight the need for improvement in the conduct and reporting of current systematic reviews with meta-analyses of time-to-event outcomes (11, 162).

For systematic review reporting in general, the PRISMA guideline and its extensions provide a widely established standard (85). More rigorous adherence to its reporting items could improve several of the problematic reporting elements identified in the presented meta-epidemiological studies. The PRISMA guideline requires, for example, in-depth reporting of the assessed review outcomes, data sources, risk of bias assessments and GRADE certainty of evidence. Other PRISMA items include details on the quantitative synthesis, such as the pooled effect measures, meta-analytic techniques and used trial estimates, how absent trial summary estimates were calculated and which additional analyses were planned.

Time-to-event outcomes are associated with distinct difficulties, including challenging recalculation methods for summary data and unique assumptions. A reporting guideline that extends currently available PRISMA guidance by time-to-event specific items could improve review authors awareness of respective hardships and positively affect reporting. In form of the "*Transparent reporting of meta-analyses of time-to-event outcomes based on aggregate data from RCTs of interventions (META-TTE reporting guideline)*" such a guideline is currently under

[27] Informative censoring is briefly mentioned in the online explanation of the tool, available under: drive.google.com/file/d/19R9savfPdCHC8XLz2iiMvL_71lPJERWK/view (p. 42; published 22.08.2019; last accessed: 24.05.2024)

development and a product the here reported dissertation projects. The development of the guideline, which is based on an large survey of international experts, is registered at the EQUATOR networks webpage for respective projects (189). A protocol is registered under: https://osf.io/j5bmw (99).
Review authors should rigorously adhere to the currently available reporting standards and transparently report issues that affect time-to-event analyses until the guidance is implemented.

7.4.1.2. Guidance for meta-analyses of time-to-event outcomes and their certainty

Guidance focusing on specific methodological issues related to time-to-event outcomes in evidence syntheses is currently scarce, including appropriate meta-analysis methods and factors that affect the certainty of their results.
As previously mentioned, the Cochrane Handbook gives some methodological advice on choosing time-to-event outcomes as effect measures for evidence syntheses (89). Other resources for individual components of the evidence synthesis process exist, for example, through the established article by Tierney and colleagues from 2007 (106). Like several others, the article by Tierney et al. 2007 (106) focusses on the recalculation of time-to-event summary data from trial publications through various approaches (145-147, 151-155, 190).

Resulting from this dissertation, the two GRADE guidelines for informative censoring and absolute effect estimates provide some guidance for selected issues of time-to-event outcomes in evidence syntheses (2, 10). Yet, additional guidance for handling other specific difficulties of time-to-event analysis, such as proportional hazards and competing events, is needed (162, 167). Guidance should be developed following an evidence-based approach and could be implemented through the GRADE Working Group. In absence of guidance, review authors must, to their best knowledge, critically address and report trial characteristics that affect their certainty to inform their readers.

The meta-epidemiological study of systematic reviews in this dissertation shows that available methodological guidance seems to have been neglected in several cases (167). Informative censoring, for example, is mentioned in the Risk of Bias 2.0 tool's explanation, but was only seldomly discussed review publications (71). In a considerable number of reviews, the authors did not specify their effect of interest and the associated choice of trial analysis, such as intention-to-treat analyses if interested in the effect of assignment to an intervention. This reduces the transparency of their analytical proceeding considerably and limits the interpretability of the presented results (162, 167).

Another often neglected recommendation relates to the choice of the appropriate methods for meta-analysis of time-to-event outcomes. For random-effect meta-analyses, methodological guidance by Cochrane advises to prefer the Hartung-Knapp-Sidik-Jonkman method over the methods by DerSimonian Laird (109, 167). Yet, most of the assessed reviews used the second method and only two reviews used the method by Hartung-Knapp-Sidik-Jonkman (167). Particularly in presence of heterogeneity between trials and with a low number of trials in a meta-analysis, which was the case in multiple of the assessed systematic reviews, the Hartung-Knapp-Sidik-Jonkman method leads to more conservative conclusions. That is because DerSimonian-Laird meta-analyses tend to inflate type I errors and to produce (too) narrow confidence intervals (109).

Overall, review authors should more rigorously adhere to methodological guidance if it is available and those responsible for publication of systematic reviews may critically assess their adherence.

7.4.1.3. Recalculation of summary data from trial publications

As indicated previously, a considerable body of literature exists for the recalculation of time-to-event summary data from study and trial publications, highlighting the great interest, but also the great importance of this field of research (145-147, 151-155, 190). The available approaches (see also chapter 3.4.1), range from rather simplistic calculations based on reported data to recalculation of individual participant data from Kaplan-Meier plots and then fitting curves for the retrieved data. Yet, as also indicated in the meta-epidemiological studies included in this dissertation, there is a lack of comprehensive empirical evidence on which procedures to prefer and under which circumstances (162, 167).

A previous study compared direct (using, e.g., O-E, log(HR), etc.) and indirect methods (using, e.g., log-rank p-values, number of events, etc.) based on reported information as well as the Kaplan-Meier plot-based approach presented by Tierney et al. 2007 (106) and Parmar et al. 1998 (151). The study showed that direct methods should be preferred before indirect and finally plot-based approaches, which both tend to underestimate effects (157). This article does not incorporate novel, potentially more advanced recalculation methods which allow to derive individual participant data from curves and rely on digitization.

Two methodological studies compared different plot-based approaches based on published Kaplan-Meier curves from trial publications or on simulation of respective data. They determined the method by Guyot et al. 2012 (145) as most reliable with regard to error from the originally reported HRs (153, 154). The studies did not include direct or indirect calculation methods from reported data. How more novel Kaplan-Meier plot-based approaches compare to indirect methods, for example, is therefore currently unknown. This might be particularly relevant since even more advanced methods including digitization (and potentially soon automatization) are, as visible on the available literature, continuously being developed (146-149, 152, 159).

Herbert and colleagues (191) recently proposed a decision algorithm for the selection of recalculation methods for summary time-to-event data from study and trial publications, irrespective of their design. They begin with directly reported data, if such data is not available, they suggest to recalculate data from indirect measures, then the approach by Guyot et al. 2012 (145) and finally suggest calculating the HR from median survival times, which ignores censoring (191). Compared to using data calculated from directly reported measures of time-to-event outcomes only, their proposed algorithm increased available study and trial HRs by 122% (191).

A clear and generally established guidance on which recalculation methods to choose under which available data is needed. Such guidance should ideally include a hierarchy of approaches with regards to the lowest error to original trial effect estimates, like the algorithm proposed by Herbert et al. 2020 (191). Guidance could not only substantially support review authors, but also introduce necessary standards and potentially reduce the complexity that is still associated with some of the Kaplan-Meier plot-based recalculation methods, for example, through development of additional software.

In the current absence of generally accepted decision aids, review authors should use all available methods to maximize the information in their meta-analyses and, however they decide for a procedure, explain their decision. That is, because the exclusion of trials from meta-analyses because of supposedly unavailable time-to-event data introduces a risk for publication bias (99, 162, 167).

Because excluding trials with unavailable time-to-event data from meta-analyses imposes a risk for publication bias, it is important to discuss if binary data should be included in a given meta-analysis in case of absent time-to-event data for an otherwise eligible trial (162, 167). In 2022, Salika and colleagues (192) recalculated the meta-analyses of large sample of Cochrane reviews as binary odds ratios and as HRs. They found that odds ratios tended to be less conservative and to produce estimates which deviated stronger from the 1 than their HR-based equivalents in some instances. They also found a different behavior of both outcome types regarding between-study heterogeneity, particularly with high event probabilities in the compared groups. Overall, they concluded that binary meta-analysis of time-to-event outcomes might be feasible in situations with a low probability of the outcome event in included trials. If binary meta-analysis is necessary for certain time-to-event outcomes, they suggest the use of complementary log-log links to approximate HRs (162, 192).

7.4.1.4. Alternatives to meta-analyses of hazard ratios

Also given the complex interpretation of HRs (see also chapter 3.1.3), another strategy to maximize the use of time-to-event data from included trials is to select an alternative effect measure for meta-analysis.

Michiels et al. 2005 (150), for example, assessed how pooled median survival times perform as a surrogate for pooled HRs from trials where HRs and additional information are not available. They compared median survival times and HRs derived from individual participant data and concluded that, because both measures show opposing treatment effects in number of meta-analyses, meta-analysis of median survival is not a feasible alternative to its HR-based counterpart.

Building on the method by Michiels et al. 2005 (150) but in the setting of experimental studies, Hirst and colleagues in 2021 (193) found a similar proportion of diverging conclusions, although they reasoned that such divergences only occurred in case of small treatment effects and in absence of statistically significant differences between groups. They concluded that even though pooled HRs and pooled median survival times do not perform perfectly similar, median survival still might constitute an option for large meta-analyses of smaller, imprecise experimental studies, if necessary.

This discrepancy in conclusions could require further assessment. The here presented meta-epidemiological studies of systematic reviews and trials required the presence of a pooled HR for eligibility. Thus, it was not possible to determine whether some systematic reviews authors could have used median survival times because of missing time-to-event summary data in trial publications.

A currently more relevant alternative to the HR is the difference in the RMST. As previously highlighted in chapter 3.1.3, the RMST represents for each arm of a trial, the average time to an event, for example, survival, from a starting time point 0 to a defined time point. It is calculated as the area under the event or survival curve up to this defined time point (30, 46). The difference in the RMST allows to compare intervention effects between groups, can be statistically tested and allows the calculation of CIs (30).

Compared to HRs, the RMST brings two substantial advances. First, it is not associated with the proportional hazards assumption and associated statistical tests between groups remain valid under non-proportional hazards (30, 40, 47). Second, the interpretation of the RMST is more straightforward, particularly the difference in the RMST between groups ("*with treatment X you can expect on average 2 more years of survival as compared to treatment with Y*") (47). Following a randomized study design, Weir and colleagues in 2019 (194) randomized RCT authors and clinicians to three groups and presented them multiple RCT abstracts. The abstracts were only altered between groups in that effects were presented either as HR only, as difference in the RMST only or as HR together with a difference in the RMST. They found that, compared with the HR only group, treatment effects in the other two groups were consistently interpreted as lower. Furthermore, they found that HRs were often incorrectly interpreted as a reduction in absolute risk (194).

Calculation of the RMST requires individual participant level data which can be recalculated from aggregated data using Kaplan-Meier curves and the before illustrated approaches. Meta-analyses of the difference in the RMST can then be calculated using the individual participant data of each included trial (145, 195, 196). Since such an approach is significantly more complex than aggregate data meta-analysis based on the HR, the role of the difference in the RMST in current meta-analyses is rather limited, despite its advances.

7.4.1.5. Implications for calculating absolute effect measures

Absolute effect measures calculated based on the pooled HR are an important tool to comprehensively communicate the results of evidence syntheses with meta-analyses of time-to-event outcomes.

Several of the miscalculated absolute effects that were identified in the meta-epidemiological study of absolute effects in cancer-related Cochrane reviews were likely attributable to a mistake in the popular guideline development software GRADEpro GDT (11, 175). The software's function to automatically calculate absolute effect measures from a relative effect and a defined baseline risk did not recognize that authors might have included a baseline risk attributable to the frequency of the absence of an event (e.g., a survival proportion). When combined with a HR calculated based on the frequency of an event (e.g., deaths) in a population, this led to falsely opposing absolute effects.

Based on the results of the here presented meta-epidemiological study, the mistakes in the software were corrected (11). With the now available GRADE guideline for absolute effects of time-to-event outcomes it can be assumed, that previous hardships in the calculation of respective measures are diminished, and that more review authors might choose to present their pooled relative outcome effects together with an absolute effect measure. A general prerequisite is, of course, that evidence synthesis authors are aware of the guideline and use it when they conduct their reviews. It has therefore been included in the Cochrane Handbook (89).

For calculating absolute effect measures based on pooled HRs, the GRADE guideline currently proposes risk differences, natural frequencies, the number-needed-to-treat and median survival times (10). Even though the positive effect of absolute effect estimates to communicate meta-analysis results is generally accepted, it is currently unclear which type of these absolute effects to use (163, 166, 197).

In a large randomized experiment amongst individuals at risk of coronary heart disease, Carling and colleagues in 2009 (164) found that patients presented with natural frequencies to communicate a risk reduction through statins were more confident in their understanding of the information and with their decision, as compared to patients who were presented other measures for the same effect, including relative risk reductions, absolute risk reductions and the number needed-to-treat.

7.4.1.6. Meta-analyses based on individual participant data

This dissertation predominately addresses meta-analyses based on aggregate time-to-event data. Some of its components are also applicable to meta-analyses of individual participant data, in particular the GRADE guidance on the calculation of absolute effects.

Besides greater reliability, as previously discussed in chapter 3.3.2, these types of meta-analyses bring considerable advances over their aggregate data counterparts (101, 198). Meta-analyses of time-to-event outcomes often require recalculation of primary data from trial publications based on assumptions that can be circumvented by available individual participant data (101). Furthermore, individual participant data allows authors to perform more elaborate analyses, for example, imputation procedures to assess informative censoring, regression modelling to assess between-study heterogeneity and competing event analyses (32, 101, 199, 200). It allows to statistically assess the assumption of proportional hazards and, in case of a failure of the assumption, to perform meta-analysis of the difference in the RMST between groups or other appropriate analyses (195).

Finally, as described in the GRADE guideline for calculating of absolute effects, individual participant level data allows to calculate absolute effects, such as risk differences, directly and to account for the statistical uncertainty of the baseline risk, which is hardly possible with aggregate data (10).

Unfortunately, individual participant data is seldomly accessible to review authors and its retrieval often not feasible without immense effort. A trend towards data sharing with new data-sharing models for clinical trials could improve this situation (201).

7.4.2. Implications for trialists

7.4.2.1. Guidelines for trial reporting

Several implications for trialists who assess time-to-event outcomes arise from the projects conducted within this dissertation. Above all and particularly arising from the findings of the meta-epidemiological study of trials included in meta-analyses of time-to-event outcomes, is the need to improve the reporting of trials with time-to-event analyses. This is, as previously emphasized, in line with the implications of previous research (23-25, 144, 160, 161, 180, 181).

Following a systematic assessment of the reporting of survival analyses in medical journals, Abraira et al. 2013 (144) proposed a list of minimal reporting items for published time-to-event analyses. Their list includes, but is not limited to, clear outcome definitions, with the type of assessed events and circumstances of censoring, sample sizes and the number of events, a valid quantification of the follow-up time and details regarding the statistical analyses, including estimates, comprehensive Kaplan-Meier curves and regression model building.

Unfortunately, the results of the meta-epidemiological study that is part of this dissertation indicate that their suggestions are not comprehensively adhered to in current RCTs included in time-to-event meta-analyses (144, 162).

More recently, an extension to the established Consolidated Standards of Reporting Trials (CONSORT) became available. The CONSORT-Outcomes reporting guideline has great potential to become an influential standard for trial outcome reporting amongst trialists, editors and publishers (162, 197, 202). Some of the issue identified for RCTs with time-to-event analyses in the meta-epidemiological study of this dissertation, including a lack of detailed information for each individual outcome, missing outcome data and the analysis types, are mentioned in the CONSORT-Outcomes guideline and might improve in future trial publications.

For the reporting of specific aspects of time-to-event analysis in trial publications, in particular for Kaplan-Meier plots, some guidance is available. In their article from 2002, Pocock and colleagues (181) present key reporting items for Kaplan-Meier curves. They call, for example, for consistent reporting of the number of individuals at risk for given time points and a presentation of statistical uncertainty, such as a CI or SE, along the curves. A valuable addition to their suggestions, in line with the meta-epidemiological studies and the GRADE guideline on informative censoring of this dissertation, would be the presentation of censored observations over time on Kaplan-Meier curves and occasionally below for certain time points (figure 1).

Morris et al. 2019 (15) present and have tested various design options for Kaplan-Meier plots that allow to follow the state of trial participants over time, by presenting sufficient numbers of individuals at risk and censored observations. They also suggest displaying the statistical uncertainty of the plots over time (15). A more rigorous presentation of such data could significantly enhance the usefulness of Kaplan-Meier plots for the conductors of evidence syntheses, for example, allowing more distinct recalculation of time-to-event data from trials.

Unfortunately, not only the completeness of reporting appears to be a problem of Kaplan-Meier plots in trial publications, but also the correctness of their underlying data. Vervölgyi et al. 2011 (160) compared the follow-up information reported in the texts of RCTs published in four high-impact journals to the number of individuals lost to follow-up, which they retrieved from Kaplan-Meier plots. They found a deviation of both numbers in 15% of trial publications, sometimes to a degree that could have affected conclusions.

Additional, time-to-event analysis-focused reporting standards could positively influence the interpretability of trials with time-to-event analyses, could increase their usefulness in secondary analyses and increase the sustainability of clinical research in general.

7.4.2.2. Additional analyses for time-to-event outcomes in trials

Observations during the presented meta-epidemiological studies and the guideline articles suggest that trialists should make more use of additional analytical options to investigate the robustness of their time-to-event analyses. Multiple options are available and can be applied in meta-analyses when individual participant data is available, for example, to investigate potential informative censoring. Available approaches can also be straightforwardly adapted to assess issues such as treatment switching and competing events.

The most simplistic amongst these approaches is to impute survival times for censored individuals under realistic assumptions (32). As applied in the GRADE guideline for informative censoring, the easiest option is to select a time-frame of trial follow-up without end-of study censoring, for instance, the enrollment period or minimum follow-up, and to impute increased or decreased survival times for individuals censored during this time-frame as best- and worst-case scenarios (2, 32).

As already mentioned in 3.2.1, more advanced methods model situations if individuals had not been censored (or, e.g., “*had they not switched treatment*”) by using observed baseline-data and occasionally post-baseline data of participants (32). Prominent approaches include inverse probability censoring weighting, rank preserving structural failure time models as well as iterative parameter estimation (65, 67, 83, 203).

Inverse probability censoring weighting, for instance, assumes a relationship between the observable data and the censoring distribution. Observed participant characteristics are used to

reconstruct HRs by weighting individuals who stay on their allocated treatment with the inverse of their probability of being censored (67, 77). Depending on the similarity of baseline and post-baseline characteristics, the observations of individuals who provide complete data to an analysis are weighted with the inverse of their probability of being censored. Thereby they adjust for their censored counterparts: individuals with more similar characteristics to those who were censored, but who were not censored themselves, provide more data to the analysis.
The approaches require that censoring can be completely explained by the participant characteristics used to adjust for censoring. Besides their strong dependance on observable data, inverse censoring weighting and related methods assume constant treatment effects throughout follow-up, which might often be implausible, and appropriate sample sizes, including a sufficient number of individuals with complete observations (77, 84).

Critically considering the underlying assumptions, the outputs of such methods might be informative to authors of evidence syntheses in some situations. A trial sensitivity analysis adjusting for possible treatment switching, for example, could support review authors interested in the per protocol effect, the effect of adherence to an intervention, when judging the risk of bias due to deviations from the intended interventions.
The authors of a systematic review assessed during the here presented meta-epidemiological study on time-to-event outcomes in systematic reviews included a trial HR generated from a rank-preserving structural failure time model for competing events in a meta-analysis. Given their considerable assumptions, it is currently unclear how to deal with the outputs of such sensitivity analyses in evidence syntheses and further research is required.

7.4.2.3. Assessing safety outcomes as time-to-event outcomes

The outcomes examined during the meta-epidemiological studies of this dissertation were almost exclusively efficacy outcomes. Safety outcomes were analyzed using time-to-event analysis only in 3% (8/235) of the assessed trial publications and were instead often addressed using binary data analysis.

For a fair comparison of treatments, the observation times of the compared groups should be equal. In safety analyses of trials, however, the observation times sometimes differ substantially. They differ, for example, when observations of participants are ceased when they switch treatments, are incompliant or otherwise deviate from the trial protocol – or if one treatment is simply more effective and prolongs survival, thereby increasing the observation time for adverse events (2, 204). Through censoring, time-to-event analysis allows to account for variable follow-up times between groups and should be more frequently used by trialists to assess safety outcomes (2, 204).
Censoring of individuals who switch treatments, are noncompliant or otherwise drop out from the trial in time-to-event analyses of safety outcomes will result in a risk of bias, as it does in analyses of efficacy outcomes. Rather than simply adjusting for informative censoring as described before, it is urged to continue the collection of safety data even after protocol violations occur and to provide adequate intention-to-treat analyses for safety outcomes (204).

Often neglected, but particularly relevant in the analysis of safety outcomes are competing events. Think, for example, of the analysis of serious adverse events in the comparison of two treatments: a hypothetical experimental treatment improves progression-free survival (commonly referred to survival without progression or all-cause death) but is associated with a substantial risk of severe adverse events. Adverse events are assessable in the respective experimental trial arm because individuals survived long enough to observe them. In the control arm,

however, because of higher rates of the competing event death, individuals did not survive long enough to prove the absence (or presence) of later severe adverse events.
This introduces bias and reduces the power of safety analyses to identify relevant severe adverse effects (205). Because the event death constitutes a competing event for adverse events, competing event analysis, including cumulative incidence estimators, cause-specific hazard models and subdistributional hazard models (chapter 2.2.3), for safety analyses is suggested (26, 51, 205-207). In the trial publications included in the meta-epidemiological studies of this dissertation, such methods were not used.

7.4.3. Implications for evidence users

7.4.3.1. Interpreting quantitative effects

As emphasized in chapter 3.4.4, relative effect measures, including the HR, endanger overinterpretation of treatment effects. In the previously mentioned study by Carling et al. 2009 (164), patients presented with a relative effect for statin effectiveness were more likely to opt for the treatment than patients who were presented the same effect in form of one of the multiple different absolute effect measures (138, 163, 164). When interpreting quantitative effects of treatments, it is therefore important to look at absolute effects in addition to relative effects (163, 208).
Adequately interpreting illustrative absolute effect measures for time-to-event outcomes requires clarifying the time point to which an absolute effect applies and to consider the source of its associated baseline risk. Baseline risks might differ, for instance, between trial populations with potentially higher risk or observational studies of the general population (10).
The meta-epidemiological study on absolute effect measures in oncological Cochrane reviews of this dissertation showed incorrect calculation and missing additional information on absolute effects in a considerable number of Cochrane reviews. A second similar assessment in non-Cochrane reviews was foreseen but not possible, because non-Cochrane reviews often present no absolute effect at all (11). Even though the now available GRADE guidance on absolute effects may inform evidence synthesis authors and reduce their problems with absolute effects in the future, users must remain critical and check presented measures for plausibility.

When interpreting effect estimates for time-to-event outcomes, in particular the HR, additional caution is required towards the direction of the effect estimates. As previously mentioned in chapter 3.4.4, the direction of effects can become especially confusing in the oncological literature. Oncological survival outcomes are reported as events (e.g., all-cause mortality, time to progression) and as absence of events (e.g., overall survival, progression-free survival), sometimes interchangeably (11, 167). In other fields, confusion can arise, for example, when the same publication reports on positive outcomes (e.g., time to hospital discharge, time to child delivery) and negative outcomes (e.g., time to preeclampsia). This is because, as shown during the here presented meta-epidemiological studies, HRs are almost inevitably calculated for events and with the experimental treatment arm included in the numerator. For negative outcomes a HR < 1 will suggest a lower rate of events and thus a benefit in the experimental group, while for positive outcomes a HR > 1 indicates a benefit in the experimental group. If established treatment options are compared, it is often not perfectly clear which of the compared groups to consider as experimental group in the numerator, potentially leading to an inversion of the HR (10).
The authors of some of the systematic reviews assessed in the here presented meta-epidemiological studies included either a trial HR that was inappropriately inverted or a trial HR that should have been inverted as compared to other HRs included in their meta-analyses (162).

The most straightforward way to assess the direction of relative effects for evidence users are Kaplan-Meier or cumulative incidence curves, which show which group is favored for a respective outcome and allow to derive an appropriate interpretation of the associated HR. In situations where no plots are available, for example, in evidence syntheses, evidence users can simply compare the proportion of events in the respective comparator groups.

Specialized result presentation formats, such as GRADE Summary of Findings Tables or Evidence Profiles, allow evidence synthesis authors to communicate the quantitative strengths of effect estimates in different formats, such as numbers and symbols, and have been shown to improve the interpretability of effects for evidence users (139).

7.4.3.2. Interpreting the certainty of time-to-event outcomes

Besides critical consideration of the size and direction of presented quantitative effects, users of the evidence should pay attention to the (un)certainty associated with a given quantitative estimate. In evidence syntheses that report their results in form of GRADE Summary of Findings tables or Evidence Profiles, the degree of the certainty of evidence can be straightforwardly assessed from the respective symbols (e.g., ⊕⊕⊕⊕ represents high certainty). Additional footnotes explain the affected domain, as introduced in chapter 3.3.4, and the reasoning for rating down the certainty, if applicable (135).

Unfortunately, except for the GRADE guidance on informative censoring that is part of this dissertation, no further guidance for time-to-event specific issues that might affect one's certainty in a body of evidence is available to date. In absence of guidance, evidence users must be aware of the core assumptions of time-to-event analysis.
A given HR, for example, may not be representative for time points other than the reported follow-up time. It should not be extrapolated beyond the follow-up, particularly if comparing treatments of differential mechanisms, because of the assumption of proportional hazards (10, 40).

There is also a substantial lack of guidance on interpreting outcomes that are susceptible to competing events. In absence of guidance, evidence users should be aware that specific statistical analyses should be used for competing events and, if regular time to event analysis is applied, the resulting estimates might underly a specific interpretation. If possible, users may seek help by someone with statistical expertise, in particular because previous research has indicated that even authors of articles including competing event analyses do occasionally not interpret their outcomes adequately (61).

7.4.4. Implications for future research

7.4.4.1. Improving reviewer expertise with time-to-event outcomes

As previously discussed, methodological research has clearly demonstrated the problematic reporting of time-to-event analyses in trials. The meta-epidemiological studies in this dissertation show that several of the problems identified in trial publications can also be found in their including reviews (162, 167). It is unclear whether and if so to which degree the inadequate reporting in systematic reviews resulted from inadequate trial reporting and how much might be attributable to, for instance, lacking expertise of review authors.

As an implication of the meta-epidemiological study on absolute effects in Cochrane reviews (chapter 6.3), which showed problems in interpretation of time-to-event outcomes by review authors, Cochrane dedicated a share of its CRG Networks Innovation Fund 2019 to the development of additional resources on time-to-event outcomes in Cochrane reviews (174). A subsequent survey amongst Cochrane stakeholders, prepared by the doctoral candidate and accepted as a poster at the 2020 Cochrane colloquium, asked Cochrane review authors for necessary resources to improve their meta-analyses of time-to-event outcomes (209). According to the survey, authors primarily lacked materials on the assumptions underlying time-to-event analysis, the reconstruction of summary data from trials and the adequate interpretation of effects. The Cochrane review authors also claimed these issues as areas of their greatest uncertainty when conducting Cochrane reviews with time-to-event outcomes (209).

It is likely that missing expertise in addition to absent or too complex guidance constitute barriers for the adequate conduct and reporting of meta-analyses of time-to-event outcomes. Additional research, including qualitative research, could elicit central obstacles and identify potential for additional guidance and targeted training resources.

7.4.4.2. Improving guidance on informative censoring in meta-analyses

There is a substantial number of articles that discuss and quantitatively investigate censoring, including informative censoring, and its influence on the results and interpretability of clinical trials (6, 7, 33-35, 37, 38, 120, 210-215). Such assessments are, however, currently absent for the impact of censoring on meta-analyses.

In an earlier study, Vale et al. 2002 (216) assessed the effects of adjusting for censoring on the pooled effect sizes, trial weights and between-study heterogeneity of meta-analyses using odds ratios. To calculate censoring-adjusted odds ratios from trials, they estimated the number of individuals at risk for a specific follow-up time point that was adjusted by a constant rate of non-informative censoring, calculated time points specific odds ratios and pooled the respective estimates. They compared their so derived pooled time point specific odds ratios to their non-censoring adjusted counterparts, which they calculated using trial odds ratios that only used the number of events at a specific time point and the total number of trial participants, instead of the time point specific number of individuals at risk. They found a considerable impact of adjusting for censoring on effect estimates, with an average difference of 2,6% and up to 9% overestimation with unadjusted odds ratios. Furthermore, adjusting for censoring resulted in lower weights of trials with lower duration of follow-up and reduced between trial heterogeneity.
Despite these findings, there are currently no investigations on the impact of variable censoring patterns on HR-based meta-analyses, which incorporate censored data by default.

There is also a lack of literature on potential informative censoring and established methods to quantify the impact of informative censoring on aggregate data meta-analyses. Such assessments would, however, constitute an important background for risk of bias evaluations. While binary outcome analysis allows to perform sensitivity analyses by imputing reasonable values for individuals with missing data, a similar approach for time-to-event analyses would require individual participant level data, because of the time-dependency of censoring (2, 126). Methodological advancements that ease the recalculation of individual participant data from trial publications could substantially support evidence synthesis authors in assessing the influence of informative censoring on their meta-analyses.

The type of censoring addressed in this dissertation and the associated projects is limited to right-censored data. As explained in chapter 3.1.1, footnote 1, other types of censored data are interval-censored data and left-censored data. Because they are potentially less relevant in clinical trials with consistent assessment and observation of trial participants, their role in meta-analyses has not yet been emphasized in the literature. For distinct situations, they might be of relevance and at least require further discussion, for example, interval censoring as a potential source of between-study heterogeneity if trials with substantially different durations between participant follow-up assessments are combined (217).

7.4.4.3. Future research on proportional hazards in meta-analyses

Extensive methodological and theoretical literature on the assumption of proportional hazards in time-to-event analyses of trials is available (13, 40, 41, 43, 46, 50, 182, 218, 219). It is currently unclear, however, how to interpret and account for the assumption in meta-analyses and evidence syntheses. Future research on the concrete, quantitative effects of a failure of the proportional hazards assumption in included trials on meta-analyses, for instance, on between-study heterogeneity, is required. In addition, guidance on interpreting estimates and deriving decisions is needed, for example, whether to classify a failure of proportional hazards as a bias or consider it a source of heterogeneity.

Rulli et al. 2018 (44), who assessed the proportional hazards assumption in lung-cancer RCTs by recalculating individual participant time-to-event data from curves, suggest applying their approach in meta-analyses. They point out a risk of over- and underestimation of effects if trial analyses with a failed proportional hazards assumption are included in meta-analyses. They recommend performing either sensitivity analyses excluding respective trials or to lean onto alternative effect measure for meta-analysis, such as time point specific event probabilities or the RMST (44).

The difference in RMST as an alternative to the HR in situations with non-proportional hazards has been highlighted in the introduction of this dissertation (chapters 3.1.3, 3.2.3 and 7.4.1.4). Pooling of the difference in the RMST from multiple trials is possible but requires individual participant data so that its usefulness for meta-analyses based on aggregate data is currently limited (195, 196). Everest and colleagues in 2022 (196) compared RMST estimates derived from recalculated Kaplan-Meier curves to estimates from corresponding individual participant data. They found little deviation and bias, even in situation with Kaplan-Meier curves of lower quality. With methodological advances that ease the recalculation of individual participant data and more complete reporting in trial publications, as previously proposed, the RMST could become a more feasible effect measure for meta-analyses, that is robust and straightforward to interpret, particularly compared to the HR.

7.4.4.4. Future research on competing events in meta-analyses

Similar to the previously mentioned methodological issues, the body of literature for considering competing events in meta-analyses based on aggregate data is scarce, although some guidance exists for meta-analyses of individual participant data (200).

In their article, Bonofiglio et al. 2016 (220) describe a computationally advanced approach for meta-analysis of competing event data by calculating cumulative incidence function ratios from aggregate trial result data and pooling them across trials, assuming a constant hazard of the groups over time.

Irrespective of this work, it remains unclear which analytical procedures to apply in case of competing events in routine meta-analysis, in particular which trial estimates to use. The cause-specific HR, simply put the Cox proportional model HR with censored competing events,

is frequently used for meta-analysis under competing events. The results of such meta-analyses require a distinct interpretation which has currently has not been discussed in the literature. Pooling of Fine and Gray subdistributional HRs, sometimes viewed as the more appropriate regression model for time-to-event data under competing events for certain instances (chapter 3.2.3), might require even more rigorous discussion (22).
Although respective guidance is currently absent, in one of the systematic reviews assessed in the meta-epidemiological studies of this dissertation, review authors combined cause-specific HRs together with a HR from a Fine and Gray subdistributional hazards model in the same meta-analysis. They did not reason their decision (162).

Future research on competing events in meta-analyses should investigate the quantitative effects of competing events on pooled estimates, clarify alternative analysis approaches and provide guidance on how to incorporate competing events in risk of bias and certainty of the evidence ratings.

7.4.4.5. Future research on treatment switching in meta-analyses

An extensive body of methodological literature is available that discusses how treatment switching affects trial results, how to interpret it and how to quantitatively assess its influence on trial estimates.
Several articles propose approaches such as inverse probability censoring weighting or the rank preserving structural failure time to recalculate trial effects had participants not switched a specific treatment (chapter 7.4.2.2) (62, 67, 73-79, 81, 82, 221-223). As previously mentioned, these approaches require individual participant data and their value in meta-analyses has not been discussed yet.
For meta-analyses of aggregate data, sensitivity analyses by excluding affected trials or meta-regression allow to assess the robustness of results to treatment switching in included trials. Practical approaches to quantitatively investigate the impact of treatment switching on meta-analyses of aggregate data are currently absent and an implication for future methodological research.

Controversially discussed today is how to interpret treatment switching in trials in a body of evidence. Some have called for treating the reception of a comparator treatment in individuals initially allocated to a different treatment arm as a risk of bias, in particular for per protocol effects (71). Others argue that treatment switching constitutes a question of applicability or external validity (63). Recently, Wang et al. 2022 (118) performed a systematic assessment of all available risk of bias tools for RCTs. They listed all items of these tool and asked international methods experts to classify these items as issues relating to bias or to other concepts, such as applicability, reporting quality and precision. Most of the experts classified the item "*Whether there is crossover to the intervention*" as an issue of applicability and not of bias.
More theoretical discussion and methodological research could support settling this ongoing discourse and standardize certainty of evidence ratings of evidence syntheses affected by treatment switching.

7.4.4.6. Future research on adjusted and unadjusted trial analyses in meta-analyses

The two first reported meta-epidemiological studies of this dissertation found that the handling of covariate adjusted and unadjusted trial estimates differed substantially between the assessed systematic reviews: in general, the adjustment of eligible trial estimates was seldomly reported in review publications. If reported, most reviews stated to include either both, adjusted and unadjusted effects, or only adjusted effect estimates, while for distinct outcomes, they most often stated to include either unadjusted or adjusted estimates only. Review authors did not provide any reasonings for their preference. When looking at the adjustment status of the trial estimates in the respective trial publications, estimates included in the assessed meta-analyses were most often unadjusted (162, 167).

Some of these inconsistencies might result from the current absence of guidance on handling covariate adjusted estimates in meta-analyses. The Cochrane Handbook, as most central resource for the conduct of systematic reviews and meta-analyses, provides no information on whether to favor covariate unadjusted or adjusted estimates, or relevant factors, for meta-analysis (89). In the chapter for choosing and computing effect estimates, the authors simply suggest that direct recalculation of effect estimates from trial reports might be necessary in situations *"… when analyses have been performed to adjust for variables used in stratified randomization or minimization, or when analysis of covariance has been used to adjust for baseline measures of an outcome*" (Cochrane Handbook chapter 6.3. "Extracting effect estimates directly" (224)). This could be interpreted as a preference for unadjusted trial estimates.

In their algorithm for selecting methods to recalculate time-to-event summary data from RCTs, but also from non-randomized study publications, Hebert et al. 2022 (191) suggest preferring adjusted over unadjusted estimates, if possible. Although they included only a minor proportion of RCTs in their assessment of the algorithm, they reasoned their preference with an increased generalizability of results. Yet, the criterion of generalizability might not apply for meta-analyses of RCTs. That is, because in case of a valid randomization procedure in the underlying trials, potential observable covariates in addition to non-observable confounders should be appropriately balanced between the compared groups. Even visible differences between baseline characteristics of the compared groups of individual trials should not constitute a problem to the validity of the RCTs and their including meta-analyses, as long as the differences are explainable by chance and not by failure of the randomization procedure (225, 226).

Some argue that covariate adjustment in RCTs enhances statistical power, particularly with stratification and with strong prognostic baseline factors for the outcome of interest (226, 227). This requires an a-priori selection of applicable prognostic factors, best determined in the trial protocol (226). Pooling adjusted estimates from RCTs might increase the statistical power of the combined estimate and the ability to detect treatment differences if they exist, but could also result in increased between-study heterogeneity.

There is currently no literature available that critically discusses and/ or suggests whether to prefer adjusted or unadjusted RCT estimates for meta-analyses and if it is feasible to include adjusted together with unadjusted estimates in the same analysis (227). Additional guidance for covariate adjustment in evidence syntheses is required and should optimally be informed by meta-epidemiological research.

8. Conclusion

The here presented dissertation provides a comprehensive view on the characteristics, methods and challenges prevalent in current evidence syntheses that include time-to-event outcomes. Two meta-epidemiological assessments, of systematic reviews and their included trials, demonstrate inconsistent and sometimes deficient methods and reporting of meta-analyses of time-to-event outcomes, and allow the interpretation that deficiencies might translate from trials to their including meta-analyses (162, 167).
A third meta-epidemiological study demonstrated not only problems in the presentation of the results of meta-analyses of time-to-event outcomes in form of absolute effect estimates, but also indicated limited understanding of the HR by systematic review authors (11).

To overcome these shortcomings, this dissertation included two systematically developed GRADE guidance articles, for informative censoring and the calculation of absolute effect estimates, that target some of the identified challenges and which should enhance the quality of meta-analyses of time-to-event outcomes in the future (2, 10).
Nonetheless, several observed, but also unobserved challenges remain. Additional research on time-to-event specific and general methodological hardships in meta-analyses is required to inform additional guidance. Those who make decisions based on meta-analyses of time-to-event outcomes should critically assess the available evidence.

9. Supplementary information

9.1. Funding

The following articles and were published as final products of the project SK 146/2-1 (www.gepris.dfg.de/gepris/projekt/437030575) which was funded by a Sachbeihilfe of the German Research Foundation (Deutsche Forschungsgemeinschaft (DFG)):

- Goldkuhle M, Hirsch C, Iannizzi C, Bora AM, Bender R, van Dalen EC, Hemkens LG, Trivella M, Monsef I, Kreuzberger N, Skoetz N. Meta-epidemiological review identified variable reporting and handling of time-to-event analyses in publications of trials included in meta-analyses of systematic reviews. Journal of Clinical Epidemiology. 2023 Jul;159:174-189. doi: 10.1016/j.jclinepi.2023.05.023.

- Goldkuhle M, Hirsch C, Iannizzi C, Zorger AM, Bender R, van Dalen EC, et al. Exploring the characteristics, methods and reporting of systematic reviews with meta-analyses of time-to-event outcomes: A meta-epidemiological study. In review. 2024.

9.2. Conflicts of interest

The author of this thesis, Marius Goldkuhle, is part of the editorial group of Cochrane Haematology, based at the University and University Hospital of Cologne, Cologne, Germany. He is also a member of the GRADE Working Group since 2019 and a co-lead of the GRADE Time-to-event Project Group.

9.3. Prizes and awards associated with this dissertation

Cochrane's Bill Silverman prize 2020 for the article "Skoetz N, Goldkuhle M, Weigl A, Dwan K, Labonté V, Dahm P, et al. Methodological review showed correct absolute effect size estimates for time-to-event outcomes in less than one-third of cancer-related systematic reviews. Journal of Clinical Epidemiology. 2019;108:1-9.".

The German Network for Evidence-based Medicine's David Sackett prize 2021 received by Prof. Dr. Nicole Skoetz and Marius Goldkuhle for "Improving the methods of meta-analyses and clinical guidelines" including the articles:

- Skoetz N, Goldkuhle M, Weigl A, Dwan K, Labonté V, Dahm P, et al. Methodological review showed correct absolute effect size estimates for time-to-event outcomes in less than one-third of cancer-related systematic reviews. Journal of Clinical Epidemiology. 2019;108:1-9.

- Skoetz N, Goldkuhle M, van Dalen EC, Akl EA, Trivella M, Mustafa RA, et al. GRADE guidelines 27: how to calculate absolute effects for time-to-event outcomes in summary of findings tables and Evidence Profiles. Journal of Clinical Epidemiology. 2020;118:124-31.

- Goldkuhle M, Bender R, Akl EA, van Dalen EC, Nevitt S, Mustafa RA, et al. GRADE Guidelines: 29. Rating the certainty in time-to-event outcomes-Study limitations due to censoring of participants with missing data in intervention studies. Journal of Clinical Epidemiology. 2021;129:126-37".

9.4. Conference presentations related to this dissertation

1. **Goldkuhle M**, van Dalen EC, Macbeth F, Skoetz N. A survey of Cochrane editors revealed several problem areas related to time-to-event (meta-)analyses. Advances in Evidence Synthesis: special issue. Cochrane Database of Systematic Reviews 2020;(9 Suppl 1):354. doi: 10.1002/14651858.CD202001 (presentation)

2. Skoetz N, Van Dalen E, **Goldkuhle M**. Improving GRADE 'Summary of findings' tables for Cochrane Reviews: Detailed guidance for the calculation of absolute effect from time-to-event data. Cochrane Learning Live Webinar. December 4th 2018. Available: https://training.cochrane.org/resource/improving-grade-%E2%80%98summary-findings%E2%80%99-tables-cochrane-reviews-detailed-guidance-calculation (webinar)

3. **Goldkuhle M**, Kreuzberger N, Skoetz N. GRADE Summary of Findings Tabellen und Evidenzprofile: Detaillierte Anleitung für Time-to-event Variablen. EbM und Digitale Transformation in der Medizin. 20. Jahrestagung des Deutschen Netzwerks Evidenzbasierte Medizin. Berlin, 21.-23.03.2019. Düsseldorf: German Medical Science GMS Publishing House; 2019. Doc19ebmS1-W1-01, doi: 10.3205/19ebm152, urn:nbn:de:0183-19ebm1529 (workshop)

4. **Goldkuhle M**, Kreuzberger N, Bora AN, Hirsch C, Iannizzi C, Skoetz N. Eigenschaften, Methoden und Berichterstattung von systematischen Reviews mit Time-to-Event-Meta-Analysen: ein meta-epidemiologisches Review. Gesundheit und Klima – EbM für die Zukunft. 24. Jahrestagung des Netzwerks Evidenzbasierte Medizin. Potsdam, 22.-24.03.2023. Düsseldorf: German Medical Science GMS Publishing House; 2023. Doc23ebmV1-05, doi: 10.3205/23ebm005, urn:nbn:de:0183-23ebm0050 (presentation)

5. **Goldkuhle M**, Kreuzberger N, Hirsch C, Iannizzi C, Bora AM, Bender R, Van Dalen EC, Hemkens LG, Skoetz N. Characteristics, reporting, and methods of trials included in time-to-event meta-analyses of systematic reviews: A meta-epidemiological review. Abstracts accepted for the 27th Cochrane Colloquium, London, UK. Cochrane Database of Systematic Reviews 2023; (1 Suppl 1): 36567. https://doi.org/10.1002/14651858.CD202301 (presentation)

6. **Goldkuhle M**, Kreuzberger N, Hirsch C, Iannizzi C, Bora AM , Bender R , Van Dalen EC , Hemkens LG, Skoetz N. Exploring the characteristics, reporting, and methods of systematic reviews including time-to-event meta-analyses and outcome analyses: A meta-epidemiological review. Abstracts accepted for the 27th Cochrane Colloquium, London, UK. Cochrane Database of Systematic Reviews 2023; (1 Suppl 1): 36563. https://doi.org/10.1002/14651858.CD202301 (poster)

7. **Goldkuhle M**, Kreuzberger N, Hirsch C, Iannizzi C, Zorger AM, Bender R, Hemkens LGH, Skoetz N. Eigenschaften, Berichterstattung und Methoden von randomisierten Studien in Meta-Analysen zu Time-to-event-Endpunkten systematischer Reviews. Evidenzbasierte Politik und Gesundheitsversorgung – erreichbares Ziel oder Illusion?; 25. Jahrestagung des Deutschen Netzwerks Evidenzbasierte Medizin. Berlin, 13.-15.03.2024. Düsseldorf: German Medical Science GMS Publishing House; 2024. doi: 10.3205/24ebm007, urn:nbn:de:0183-24ebm0079 (presentation)

9.5. Acknowledgements

I am deeply grateful for the support, guidance and trust that Prof. Dr. Nicole Skoetz has given me over the past almost ten years.

I am thankful for all the wonderful colleagues with whom I have had the privilege of working, debating and celebrating over these years, and the co-authors of all the articles from whom I have learned so much.

Finally, I sincerely thank my tutors and referees, Prof. Dr. Thorsten Simon, Prof. Dr. Achim Tresch and Prof. Dr. Martin Hellmich, who have kindly guided me through the journey of this dissertation.

10. Bibliography

1. Sato Y, Gosho M, Nagashima K, Takahashi S, Ware JH, Laird NM. Statistical methods in the journal — an update. New England Journal of Medicine. 2017;376(11):1086-7.

2. Goldkuhle M, Bender R, Akl EA, van Dalen EC, Nevitt S, Mustafa RA, et al. GRADE guidelines: 29. rating the certainty in time-to-event outcomes - study limitations due to censoring of participants with missing data in intervention studies. Journal of Clinical Epidemiology. 2021;129:126-37.

3. Kleinbaum DG, Klein M. Survival analysis. 3 ed. New York: Springer-Verlag; 2012.

4. Bradburn MJ, Clark TG, Love SB, Altman DG. Survival analysis part II: multivariate data analysis – an introduction to concepts and methods. British Journal of Cancer. 2003;89(3):431-6.

5. Altman DG. Analysis of survival times. Practical statistics for medical research. London: Chapmann & Hall/CRC; 1991.

6. Lagakos SW. General right censoring and its impact on the analysis of survival data. Biometrics. 1979;35(1):139-56.

7. Leung K-M, Elashoff RM, Afifi AA. CENSORING ISSUES IN SURVIVAL ANALYSIS. Annual Review of Public Health. 1997;18(1):83-104.

8. Spruance SL, Reid JE, Grace M, Samore M. Hazard ratio in clinical trials. Antimicrobial Agents and Chemotherapy. 2004;48(8):2787-92.

9. Clark TG, Bradburn MJ, Love SB, Altman DG. Survival analysis part I: basic concepts and first analyses. British Journal of Cancer. 2003;89(2):232-8.

10. Skoetz N, Goldkuhle M, van Dalen EC, Akl EA, Trivella M, Mustafa RA, et al. GRADE guidelines 27: how to calculate absolute effects for time-to-event outcomes in summary of findings tables and Evidence Profiles. Journal of Clinical Epidemiology. 2020;118:124-31.

11. Skoetz N, Goldkuhle M, Weigl A, Dwan K, Labonté V, Dahm P, et al. Methodological review showed correct absolute effect size estimates for time-to-event outcomes in less than one-third of cancer-related systematic reviews. Journal of Clinical Epidemiology. 2019;108:1-9.

12. 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. The Lancet Oncology. 2017;18(12):1688-700.

13. Jachno K, Heritier S, Wolfe R. Are non-constant rates and non-proportional treatment effects accounted for in the design and analysis of randomised controlled trials? A review of current practice. BMC Medical Research Methodology. 2019;19(1):103.

14. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. Journal of the American Statistical Association. 1958;53(282):457-81.

15. Morris TP, Jarvis CI, Cragg W, Phillips PPJ, Choodari-Oskooei B, Sydes MR. Proposals on Kaplan-Meier plots in medical research and a survey of stakeholder views: KMunicate. BMJ Open. 2019;9(9):e030215.

16. Sachs MC, Brand A, Gabriel EE. Confidence bands in survival analysis. British Journal of Cancer. 2022;127(9):1636-41.

17. Yang S, Prentice R. Improved logrank-type tests for survival data using adaptive weights. Biometrics. 2010;66(1):30-8.

18. Cox DR. Regression models and life-tables. Journal of the Royal Statistical Society Series B (Methodological). 1972;34(2):187-220.

19. Bradburn MJ, Clark TG, Love SB, Altman DG. Survival analysis part III: multivariate data analysis – choosing a model and assessing its adequacy and fit. British Journal of Cancer. 2003;89(4):605-11.

20. Hernán MA. The hazards of hazard ratios. Epidemiology. 2010;21(1):13-5.

21. Kay R. An explanation of the hazard ratio. Pharmaceutical Statistics. 2004;3(4):295-7.

22. Austin PC, Lee DS, Fine JP. Introduction to the analysis of survival data in the presence of competing risks. Circulation. 2016;133(6):601-9.

23. Altman DG, De Stavola BL, Love SB, Stepniewska KA. Review of survival analyses published in cancer journals. British Journal of Cancer. 1995;72(2):511-8.

24. 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(5):e0154870.

25. Mathoulin-Pelissier S, Gourgou-Bourgade S, Bonnetain F, Kramar A. Survival end point reporting in randomized cancer clinical trials: a review of major journals. Journal of Clinical Oncology. 2008;26(22):3721-6.

26. Fine JP, Gray RJ. A proportional hazards model for the subdistribution of a competing Risk. Journal of the American Statistical Association. 1999;94(446):496-509.

27. Nardi A, Schemper M. Comparing Cox and parametric models in clinical studies. Statistics in Medicine. 2003;22(23):3597-610.

28. Le-Rademacher J, Wang X. Time-To-Event Data: An Overview and Analysis Considerations. Journal of Thoracic Oncology. 2021;16(7):1067-74.

29. Bender R, Augustin T, Blettner M. Generating survival times to simulate Cox proportional hazards models. Statistics in Medicine. 2005;24(11):1713-23.

30. Royston P, Parmar MKB. Restricted mean survival time: an alternative to the hazard ratio for the design and analysis of randomized trials with a time-to-event outcome. BMC Medical Research Methodology. 2013;13(1):152.

31. Wei LJ. The accelerated failure time model: A useful alternative to the cox regression model in survival analysis. Statistics in Medicine. 1992;11(14-15):1871-9.

32. Clark TG, Bradburn MJ, Love SB, Altman DG. Survival analysis part IV: further concepts and methods in survival analysis. British Journal of Cancer. 2003;89(5):781-6.

33. Prinja S, Gupta N, Verma R. Censoring in clinical trials: review of survival analysis techniques. Indian Journal of Community Medicine. 2010;35(2):217-21.

34. Lesko CR, Edwards JK, Moore RD, Lau B. Censoring for loss to follow-up in time-to-event analyses of composite outcomes or in the presence of competing risks. Epidemiology. 2019;30(6):817-24.

35. 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. Statistics in Medicine. 2014;33(27):4681-94.

36. Hernán MA, Hernández-Díaz S, Robins JM. A structural approach to selection bias. Epidemiology. 2004;15(5):615-25.

37. Siannis F. Applications of a parametric model for informative censoring. Biometrics. 2004;60(3):704-14.

38. Campigotto F, Weller E. Impact of informative censoring on the Kaplan-Meier estimate of progression-Free survival in phase II clinical trials. Journal of Clinical Oncology. 2014;32(27):3068-74.

39. Persson I, Khamis H. Bias of the Cox model hazard ratio. Journal of Modern Applied Statistical Methods. 2005;4(1):90-9.

40. Stensrud MJ, Aalen JM, Aalen OO, Valberg M. Limitations of hazard ratios in clinical trials. European Heart Journal. 2019;40(17):1378-83.

41. Stensrud MJ, Hernán MA. Why test for proportional hazards? JAMA. 2020;323(14):1401-2.

42. Howard G, Chambless LE, Kronmal RA. Assessing differences in clinical trials comparing surgical vs nonsurgical Therapy: using common (statistical) sense. JAMA. 1997;278(17):1432-6.

43. Royston P, Parmar MKB. An approach to trial design and analysis in the era of non-proportional hazards of the treatment effect. Trials. 2014;15(1):314.

44. Rulli E, Ghilotti F, Biagioli E, Porcu L, Marabese M, D'Incalci M, et al. Assessment of proportional hazard assumption in aggregate data: a systematic review on statistical methodology in clinical trials using time-to-event endpoint. British Journal of Cancer. 2018;119(12):1456-63.

45. Alexander BM, Schoenfeld JD, Trippa L. Hazards of hazard ratios — deviations from model assumptions in immunotherapy. New England Journal of Medicine. 2018;378(12):1158-9.

46. Trinquart L, Jacot J, Conner SC, Porcher R. Comparison of treatment effects measured by the hazard ratio and by the ratio of restricted mean survival times in oncology randomized controlled trials. Journal of Clinical Oncology. 2016;34(15):1813-9.

47. Royston P, Parmar MKB. The use of restricted mean survival time to estimate the treatment effect in randomized clinical trials when the proportional hazards assumption is in doubt. Statistics in Medicine. 2011;30(19):2409-21.

48. Schemper M, Wakounig S, Heinze G. The estimation of average hazard ratios by weighted Cox regression. Statistics in Medicine. 2009;28(19):2473-89.

49. Li H, Han D, Hou Y, Chen H, Chen Z. Statistical inference methods for two crossing survival curves: a comparison of methods. PLoS ONE. 2015;10(1):e0116774.

50. Grambsch PM, Therneau TM. Proportional hazards tests and diagnostics based on weighted residuals. Biometrika. 1994;81(3):515-26.

51. Prentice RL, Kalbfleisch JD, Peterson AV, Flournoy N, Farewell VT, Breslow NE. The analysis of failure times in the presence of competing risks. Biometrics. 1978;34(4):541-54.

52. Austin PC, Fine JP. Accounting for competing risks in randomized controlled trials: a review and recommendations for improvement. Statistics in Medicine. 2017;36(8):1203-9.

53. Gooley TA, Leisenring W, Crowley J, Storer BE. Estimation of failure probabilities in the presence of competing risks: new representations of old estimators. Statistics in Medicine. 1999;18(6):695-706.

54. Putter H, Fiocco M, Geskus RB. Tutorial in biostatistics: competing risks and multi-state models. Statistics in Medicine. 2007;26(11):2389-430.

55. Huebner M, Wolkewitz M, Enriquez-Sarano M, Schumacher M. Competing risks need to be considered in survival analysis models for cardiovascular outcomes. The Journal of Thoracic and Cardiovascular Surgery. 2017;153(6):1427-31.

56. Coemans M, Verbeke G, Döhler B, Süsal C, Naesens M. Bias by censoring for competing events in survival analysis. BMJ. 2022;378:e071349.

57. Manja V, AlBashir S, Guyatt G. Criteria for use of composite end points for competing risks—a systematic survey of the literature with recommendations. Journal of Clinical Epidemiology. 2017;82:4-11.

58. Schumacher M, Ohneberg K, Beyersmann J. Competing risk bias was common in a prominent medical journal. Journal of Clinical Epidemiology. 2016;80:135-6.

59. van Walraven C, McAlister FA. Competing risk bias was common in Kaplan-Meier risk estimates published in prominent medical journals. J Clin Epidemiol. 2016;69:170-3.e8.

60. Koller MT, Raatz H, Steyerberg EW, Wolbers M. Competing risks and the clinical community: irrelevance or ignorance? Statistics in Medicine. 2012;31(11-12):1089-97.

61. Austin PC, Fine JP. Practical recommendations for reporting Fine-Gray model analyses for competing risk data. Statistics in Medicine. 2017;36(27):4391-400.

62. Garas G, Markar SR, Malietzis G, Ashrafian H, Hanna GB, Zacharakis E, et al. Induced bias due to crossover within randomized controlled trials in surgical oncology: a meta-regression analysis of minimally invasive versus open surgery for the treatment of gastrointestinal cancer. Ann Surg Oncol. 2018;25(1):221-30.

63. Goldkuhle M, Guyatt GH, Kreuzberger N, Akl EA, Dahm P, van Dalen EC, et al. GRADE concept 4: rating the certainty of evidence when study interventions or comparators differ from PICO targets. Journal of Clinical Epidemiology. 2023;159:40-8.

64. Gaudino M, Fremes SE, Ruel M, Di Franco A, Di Mauro M, Chikwe J, et al. Prevalence and impact of treatment crossover in cardiac surgery randomized trials: a meta-epidemiologic study. Journal of the American Heart Association. 2019;8(21):e013711.

65. Latimer NR. Treatment switching in oncology trials and the acceptability of adjustment methods. Expert Review of Pharmacoeconomics & Outcomes Research. 2015;15(4):561-4.

66. Magill N, Knight R, McCrone P, Ismail K, Landau S. A scoping review of the problems and solutions associated with contamination in trials of complex interventions in mental health. BMC Medical Research Methodology. 2019;19(1):4.

67. Henshall C, Latimer NR, Sansom L, Ward RL. Treatment switching in cancer trial: issues and proposals. International Journal of Technology Assessment in Health Care. 2016;32(3):167-74.

68. Pinsky PF, Black A, Kramer BS, Miller A, Prorok PC, Berg C. Assessing contamination and compliance in the prostate component of the Prostate, Lung, Colorectal, and Ovarian (PLCO) Cancer Screening Trial. Clinical Trials. 2010;7(4):303-11.

69. Michiels H, Sotto C, Vandebosch A, Vansteelandt S. A novel estimand to adjust for rescue treatment in randomized clinical trials. Statistics in Medicine. 2021;40(9):2257-71.

70. Mansournia MA, Higgins JP, Sterne JA, Hernán MA. Biases in randomized trials: a conversation between trialists and epidemiologists. Epidemiology. 2017;28(1):54-9.

71. Sterne JAC, Savović J, Page MJ, Elbers RG, Blencowe NS, Boutron I, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. BMJ. 2019;366:l4898.

72. Higgins JP, Altman DG, Gøtzsche PC, Jüni P, Moher D, Oxman AD, et al. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. BMJ. 2011;343:d5928.

73. Hernán MA, Hernández-Díaz S. Beyond the intention-to-treat in comparative effectiveness research. Clin Trials. 2012;9(1):48-55.

74. Köhler M, Beckmann L, Grouven U, Guddat C, Vervölgyi V, S W. Treatment Switching in onkologischen Studien 2018 [cited 2023 01.09.2023]. Available from: https://www.iqwig.de/projekte/ga14-04.html.

75. Hernán MA, Robins JM. Per-protocol analyses of pragmatic trials. New England Journal of Medicine. 2017;377(14):1391-8.

76. Keene ON, Lynggaard H, Englert S, Lanius V, Wright D. Why estimands are needed to define treatment effects in clinical trials. BMC Medicine. 2023;21(1):276.

77. Ishak KJ, Proskorovsky I, Korytowsky B, Sandin R, Faivre S, Valle J. Methods for adjusting for bias due to crossover in oncology trials. Pharmacoeconomics. 2014;32(6):533-46.

78. Sullivan TR, Latimer NR, Gray J, Sorich MJ, Salter AB, Karnon J. Adjusting for treatment switching in oncology trials: a systematic review and recommendations for reporting. Value Health. 2020;23(3):388-96.

79. Latimer NR, White IR, Abrams KR, Siebert U. Causal inference for long-term survival in randomised trials with treatment switching: Should re-censoring be applied when estimating counterfactual survival times? Statistical Methods in Medical Research. 2019;28(8):2475-93.

80. Latimer NR, Henshall C, Siebert U, Bell H. Treatment switching: statistical and decision-making challenges and approaches. International Journal of Technology Assessment in Health Care. 2016;32(3):160-6.

81. Latimer NR, Abrams KR, Siebert U. Two-stage estimation to adjust for treatment switching in randomised trials: a simulation study investigating the use of inverse probability weighting instead of re-censoring. BMC Medical Research Methodology. 2019;19(1):69.

82. Latimer NR, Abrams KR, Lambert PC, Crowther MJ, Wailoo AJ, Morden JP, et al. Adjusting for treatment switching in randomised controlled trials - a simulation study and a simplified two-stage method. Statistical Methods in Medical Research. 2017;26(2):724-51.

83. Latimer NR, Abrams KR, Lambert PC, Crowther MJ, Wailoo AJ, Morden JP, et al. Adjusting survival time estimates to account for treatment switching in randomized controlled trials—an economic evaluation context: methods, limitations, and recommendations. Medical Decision Making. 2014;34(3):387-402.

84. Jönsson L, Sandin R, Ekman M, Ramsberg J, Charbonneau C, Huang X, et al. Analyzing overall survival in randomized controlled trials with crossover and implications for economic evaluation. Value in Health. 2014;17(6):707-13.

85. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. Systematic Reviews. 2021;10(1):89.

86. Page MJ, Shamseer L, Altman DG, Tetzlaff J, Sampson M, Tricco AC, et al. Epidemiology and reporting characteristics of systematic reviews of biomedical research: a cross-sectional study. PLoS Medicine. 2016;13(5):e1002028.

87. Elliott J, Lawrence R, Minx JC, Oladapo OT, Ravaud P, Tendal Jeppesen B, et al. Decision makers need constantly updated evidence synthesis. Nature. 2021;600(7889):383-5.

88. Mulrow CD. Systematic Reviews: Rationale for systematic reviews. BMJ. 1994;309(6954):597.

89. Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, et al. Cochrane Handbook for Systematic Reviews of Interventions: Cochrane; 2020. Available from: www.training.cochrane.org/handbook.

90. Lasserson TJ, Thomas J, Higgins JPT. Chapter 1: Starting a review. 2023. In: Cochrane Handbook for Systematic Reviews of Interventions version [Internet]. Cochrane. Available from: www.training.cochrane.org/handbook.

91. Sena ES, Currie GL, McCann SK, Macleod MR, Howells DW. Systematic reviews and meta-analysis of preclinical studies: why perform them and how to appraise them critically. Journal of Cerebral Blood Flow & Metabolism. 2014;34(5):737-42.

92. Guyatt GH, Oxman AD, Kunz R, Atkins D, Brozek J, Vist G, et al. GRADE guidelines: 2. framing the question and deciding on important outcomes. Journal of Clinical Epidemiology. 2011;64(4):395-400.

93. Blank O, von Tresckow B, Monsef I, Specht L, Engert A, Skoetz N. Chemotherapy alone versus chemotherapy plus radiotherapy for adults with early stage Hodgkin lymphoma. Cochrane Database Syst Rev. 2017;4(4):Cd007110.

94. Murad MH, Noor A, Mouaz A, Fares A. New evidence pyramid. Evidence Based Medicine. 2016;21(4):125.

95. Berlin JA, Golub RM. Meta-analysis as evidence: building a better pyramid. JAMA. 2014;312(6):603-6.

96. McInnes MDF, Moher D, Thombs BD, McGrath TA, Bossuyt PM, Clifford T, et al. Preferred reporting items for a systematic review and meta-analysis of diagnostic test accuracy studies: the PRISMA-DTA statement. JAMA. 2018;319(4):388-96.

97. Stewart LA, Clarke M, Rovers M, Riley RD, Simmonds M, Stewart G, et al. Preferred reporting items for a systematic review and meta-analysis of individual participant data: the PRISMA-IPD statement. JAMA. 2015;313(16):1657-65.

98. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Annals of Internal Medicine. 2018;169(7):467-73.

99. Goldkuhle M, Kreuzberger N, Bender R, Bora A, Burdett S, Hirsch C, et al. Transparent reporting of meta-analyses of time-to-event outcomes based on aggregate data from randomized trials of interventions (META-TTE reporting guideline)2023 22.08.2024 [cited 2024 22.08.2024]. Available from: https://osf.io/j5bmw.

100. Smith CT, Williamson PR, Marson AG. An overview of methods and empirical comparison of aggregate data and individual patient data results for investigating heterogeneity in meta-analysis of time-to-event outcomes. Journal of Evaluation in Clinical Practice. 2005;11(5):468-78.

101. Tierney JF, Fisher DJ, Burdett S, Stewart LA, Parmar MKB. Comparison of aggregate and individual participant data approaches to meta-analysis of randomised trials: An observational study. PLoS Med. 2020;17(1):e1003019.

102. Review Manager (RevMan). Version 5.4 ed: The Cochrane Collaboration; 2020.

103. Borenstein M, Hedges LV, Higgins JPT, Rothstein HR. A basic introduction to fixed-effect and random-effects models for meta-analysis. Research Synthesis Methods. 2010;1(2):97-111.

104. Deeks JJ, Higgins JPT, Altman DG. Chapter 10: analysing data and undertaking meta-analyses. 2023. In: Cochrane Handbook for Systematic Reviews of Interventions [Internet]. Cochrane. Available from: www.training.cochrane.org/handbook.

105. Higgins JP, Thompson SG, Spiegelhalter DJ. A re-evaluation of random-effects meta-analysis. Journal of the Royal Statistical Society Series A, (Statistics in Society). 2009;172(1):137-59.

106. 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.

107. DerSimonian R, Laird N. Meta-analysis in clinical trials. Controlled Clinical Trials. 1986;7(3):177-88.

108. DerSimonian R, Kacker R. Random-effects model for meta-analysis of clinical trials: an update. Contemporary Clinical Trials. 2007;28(2):105-14.

109. Veroniki A, Jackson D, Viechtbauer W, Bender R, Knapp G, Kuss O, et al. Recommendations for quantifying the uncertainty in the summary intervention effect and estimating the between-study heterogeneity variance in random-effects meta-analysis. Cochrane Database of Systematic Reviews. 2015:25-7.

110. IntHout J, Ioannidis JPA, Borm GF. The Hartung-Knapp-Sidik-Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird method. BMC Medical Research Methodology. 2014;14(1):25.

111. Langan D, Higgins JPT, Jackson D, Bowden J, Veroniki AA, Kontopantelis E, et al. A comparison of heterogeneity variance estimators in simulated random-effects meta-analyses. Research Synthesis Methods. 2019;10(1):83-98.

112. Higgins JPT, Thompson SG. Quantifying heterogeneity in a meta-analysis. Statistics in Medicine. 2002;21(11):1539-58.

113. Guyatt GH, Oxman AD, Kunz R, Woodcock J, Brozek J, Helfand M, et al. GRADE guidelines: 7. rating the quality of evidence--inconsistency. Journal of Clinical Epidemiology. 2011;64(12):1294-302.

114. 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). Journal of Clinical Epidemiology. 2011;64(4):407-15.

115. Higgins JPT, Altman DG, Sterne JAC. Chapter 8: assessing risk of bias in included studies. 2011. In: Cochrane Handbook for Systematic Reviews of Interventions [Internet]. The Cochrane Collaboration. Available from: www.handbook.cochrane.org.

116. Wood L, Egger M, Gluud LL, Schulz KF, Jüni P, Altman DG, et al. Empirical evidence of bias in treatment effect estimates in controlled trials with different interventions and outcomes: meta-epidemiological study. BMJ. 2008;336(7644):601.

117. Page MJ, Higgins JPT, Clayton G, Sterne JAC, Hróbjartsson A, Savović J. Empirical evidence of study design biases in randomized trials: systematic review of meta-epidemiological studies. PLoS One. 2016;11(7):e0159267.

118. Wang Y, Ghadimi M, Wang Q, Hou L, Zeraatkar D, Iqbal A, et al. Instruments assessing risk of bias of randomized trials frequently included items that are not addressing risk of bias issues. J Clin Epidemiol. 2022;152:218-25.

119. Schuster NA, Hoogendijk EO, Kok AAL, Twisk JWR, Heymans MW. Ignoring competing events in the analysis of survival data may lead to biased results: a nonmathematical illustration of competing risk analysis. Journal of Clinical Epidemiology. 2020;122:42-8.

120. Templeton AJ, Amir E, Tannock IF. Informative censoring — a neglected cause of bias in oncology trials. Nature Reviews Clinical Oncology. 2020;17(6):327-8.

121. GRADE Working Group. Grading quality of evidence and strength of recommendations. BMJ. 2004;328(7454):1490.

122. Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P, et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ. 2008;336(7650):924.

123. The GRADE Working Group 2023 [cited 2024 06.06.2024]. Available from: https://www.gradeworkinggroup.org/.

124. Balshem H, Helfand M, Schünemann HJ, Oxman AD, Kunz R, Brozek J, et al. GRADE guidelines: 3. rating the quality of evidence. Journal of Clinical Epidemiology. 2011;64(4):401-6.

125. Hultcrantz M, Rind D, Akl EA, Treweek S, Mustafa RA, Iorio A, et al. The GRADE Working Group clarifies the construct of certainty of evidence. Journal of Clinical Epidemiology. 2017;87:4-13.

126. 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. Journal of Clinical Epidemiology. 2017;87:14-22.

127. Guyatt GH, Oxman AD, Kunz R, Woodcock J, Brozek J, Helfand M, et al. GRADE guidelines: 8. rating the quality of evidence--indirectness. Journal of Clinical Epidemiology. 2011;64(12):1303-10.

128. Guyatt GH, Oxman AD, Kunz R, Brozek J, Alonso-Coello P, Rind D, et al. GRADE guidelines 6. rating the quality of evidence--imprecision. Journal of Clinical Epidemiology. 2011;64(12):1283-93.

129. Schünemann HJ, Neumann I, Hultcrantz M, Brignardello-Petersen R, Zeng L, Murad MH, et al. GRADE guidance 35: update on rating imprecision for assessing contextualized certainty of evidence and making decisions. Journal of Clinical Epidemiology. 2022;150:225-42.

130. Zeng L, Brignardello-Petersen R, Hultcrantz M, Mustafa RA, Murad MH, Iorio A, et al. GRADE Guidance 34: update on rating imprecision using a minimally contextualized approach. Journal of Clinical Epidemiology. 2022;150:216-24.

131. Dwan K, Gamble C, Williamson PR, Kirkham JJ. Systematic review of the empirical evidence of study publication bias and outcome reporting bias - an updated review. PLoS One. 2013;8(7):e66844.

132. Guyatt GH, Oxman AD, Montori V, Vist G, Kunz R, Brozek J, et al. GRADE guidelines: 5. rating the quality of evidence--publication bias. Journal of Clinical Epidemiology. 2011;64(12):1277-82.

133. Higgins JPT, Savović J, Page MJ, Elbers RG, Sterne JAC. Chapter 8: assessing risk of bias in a randomized trial. 2023. In: Cochrane Handbook for Systematic Reviews of Interventions [Internet]. Cochrane. Available from: www.training.cochrane.org/handbook.

134. Guyatt GH, Oxman AD, Sultan S, Glasziou P, Akl EA, Alonso-Coello P, et al. GRADE guidelines: 9. rating up the quality of evidence. Journal of Clinical Epidemiology. 2011;64(12):1311-6.

135. Guyatt G, Oxman AD, Akl EA, Kunz R, Vist G, Brozek J, et al. GRADE guidelines: 1. Introduction—GRADE evidence profiles and summary of findings tables. Journal of Clinical Epidemiology. 2011;64(4):383-94.

136. Guyatt GH, Oxman AD, Santesso N, Helfand M, Vist G, Kunz R, et al. GRADE guidelines: 12. preparing Summary of Findings tables - binary outcomes. Journal of Clinical Epidemiology. 2013;66(2):158-72.

137. Guyatt GH, Thorlund K, Oxman AD, Walter SD, Patrick D, Furukawa TA, et al. GRADE guidelines: 13. preparing Summary of Findings tables and evidence profiles - continuous outcomes. Journal of Clinical Epidemiology. 2013;66(2):173-83.

138. Gigerenzer G, Gaissmaier W, Kurz-Milcke E, Schwartz LM, Woloshin S. Helping doctors and patients make sense of health statistics. Psychological Science in the Public Interest. 2007;8(2):53-96.

139. Akl EA, Maroun N, Guyatt G, Oxman AD, Alonso-Coello P, Vist GE, et al. Symbols were superior to numbers for presenting strength of recommendations to health care consumers: a randomized trial. Journal of Clinical Epidemiology. 2007;60(12):1298-305.

140. Rosenbaum SE, Glenton C, Nylund HK, Oxman AD. User testing and stakeholder feedback contributed to the development of understandable and useful Summary of Findings tables for Cochrane reviews. Journal of Clinical Epidemiology. 2010;63(6):607-19.

141. Rosenbaum SE, Glenton C, Oxman AD. Summary-of-findings tables in Cochrane reviews improved understanding and rapid retrieval of key information. Journal of Clinical Epidemiology. 2010;63(6):620-6.

142. McGowan J, Akl EA, Coello PA, Brennan S, Dahm P, Davoli M, et al. Update on the JCE GRADE series and other GRADE article types. Journal of Clinical Epidemiology. 2021;140:163-4.

143. Schünemann HJ, Brennan S, Akl EA, Hultcrantz M, Alonso-Coello P, Xia J, et al. The development methods of official GRADE articles and requirements for claiming the use of GRADE – A statement by the GRADE guidance group. Journal of Clinical Epidemiology. 2023;159:79-84.

144. Abraira 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. Journal of Clinical Epidemiology. 2013;66(12):1340-6.e5.

145. Guyot P, Ades A, Ouwens MJ, Welton NJ. Enhanced secondary analysis of survival data: reconstructing the data from published Kaplan-Meier survival curves. BMC Medical Research Methodology. 2012;12(1):1-13.

146. Irvine AF, Waise S, Green EW, Stuart B. A non-linear optimisation method to extract summary statistics from Kaplan-Meier survival plots using the published P value. BMC Medical Research Methodology. 2020;20(1):269.

147. Liu N, Zhou Y, Lee JJ. IPDfromKM: reconstruct individual patient data from published Kaplan-Meier survival curves. BMC Medical Research Methodology. 2021;21(1):111.

148. Liu Z, Rich B, Hanley JA. Recovering the raw data behind a non-parametric survival curve. Systematic Reviews. 2014;3(1):151.

149. Messori A, Damuzzo V, Rivano M, Cancanelli L, Di Spazio L, Ossato A, et al. Application of the IPDfromKM-Shiny method to compare the efficacy of novel treatments aimed at the same disease condition: a report of 14 analyses. Cancers (Basel). 2023;15(6).

150. Michiels S, Piedbois P, Burdett S, Syz N, Stewart L, Pignon JP. Meta-analysis when only the median survival times are known: a comparison with individual patient data results. Int J Technol Assess Health Care. 2005;21(1):119-25.

151. Parmar MKB, Torri V, Stewart L. Extracting summary statistics to perform meta-analyses of the published literature for survival endpoints. Statistics in Medicine. 1998;17(24):2815-34.

152. Rogula B, Lozano-Ortega G, Johnston KM. A method for reconstructing individual patient data from Kaplan-Meier survival curves that incorporate marked censoring times. MDM Policy & Practice. 2022;7(1):23814683221077643.

153. Saluja R, Cheng S, Delos Santos KA, Chan KKW. Estimating hazard ratios from published Kaplan-Meier survival curves: A methods validation study. Research Synthesis Methods. 2019;10(3):465-75.

154. Wan X, Peng L, Li Y. A Review and Comparison of Methods for Recreating Individual Patient Data from Published Kaplan-Meier Survival Curves for Economic Evaluations: A Simulation Study. PLOS ONE. 2015;10(3):e0121353.

155. Wei Y, Royston P. Reconstructing time-to-event data from published Kaplan-Meier curves. Stata Journal 2017;17(4):786-802.

156. Williamson PR, Smith CT, Hutton JL, Marson AG. Aggregate data meta-analysis with time-to-event outcomes. Statistics in Medicine. 2002;21(22):3337-51.

157. Hirooka T, Hamada C, Yoshimura I. A note on estimating treatment effect for time-to-event data in a literature-based meta-analysis. Methods of Information in Medicine. 2009;48(2):104-12.

158. Abdallah DY. Potential bias in the indirect methods for extracting summary statistics in literature-based meta-analyses: an empirical evaluation. PeerJ PrePrints. 2013;1(e142v1).

159. Hoyle MW, Henley W. Improved curve fits to summary survival data: application to economic evaluation of health technologies. BMC Medical Research Methodology. 2011;11(1):139.

160. 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.

161. Zhu X, Zhou X, Zhang Y, Sun X, Liu H, Zhang Y. Reporting and methodological quality of survival analysis in articles published in Chinese oncology journals. Medicine (Baltimore). 2017;96(50):e9204.

162. Goldkuhle M, Hirsch C, Iannizzi C, Bora AM, Bender R, van Dalen EC, et al. Meta-epidemiological review identified variable reporting and handling of time-to-event analyses in publications of trials included in meta-analyses of systematic reviews. Journal of Clinical Epidemiology. 2023(159):174-89.

163. Busse JW, Guyatt GH. Copresentation of relative and absolute effects is essential to promote optimal interpretability of treatment effects. Journal of Clinical Epidemiology. 2015;68(3):355-6.

164. Carling CLL, Kristoffersen DT, Montori VM, Herrin J, Schünemann HJ, Treweek S, et al. The effect of alternative summary statistics for communicating risk reduction on decisions about taking statins: a randomized trial. PLOS Medicine. 2009;6(8):e1000134.

165. Schünemann HJ, Higgins JPT, Vist GE, Glasziou P, Akl EA, Skoetz N, et al. Chapter 14: completing ‘Summary of findings’ tables and grading the certainty of the evidence. 2023. In: Cochrane Handbook for Systematic Reviews of Interventions [Internet]. Cochrane. Available from: www.training.cochrane.org/handbook.

166. Altman DG, Andersen PK. Calculating the number needed to treat for trials where the outcome is time to an event. BMJ. 1999;319(7223):1492.

167. Goldkuhle M, Hirsch C, Iannizzi C, Zorger AM, Bender R, van Dalen EC, et al. Exploring the characteristics, methods and reporting of systematic reviews with meta-analyses of time-to-event outcomes: a meta-epidemiological study. BMC Medical Research Methodology. 2024;24(1):291.

168. Goldkuhle M, Kreuzberger N, Bora AM, Hirsch C, Iannizzi C, Skoetz N, editors. Eigenschaften, Methoden und Berichterstattung von systematischen Reviews mit Time-to-Event-Meta-Analysen: ein meta-epidemiologisches Review. Gesundheit und Klima – EbM für die Zukunft 24 Jahrestagung des Netzwerks Evidenzbasierte Medizin; 2023; Potsdam, 22.-24.03.2023: Düsseldorf: German Medical Science GMS Publishing House.

169. Goldkuhle M, Kreuzberger N, Hirsch C, Iannizzi C, Bora AM, Bender R, et al., editors. Exploring the characteristics, reporting, and methods of systematic reviews including time-to-

event meta-analyses and outcome analyses. Abstracts accepted for the 27th Cochrane Colloquium; 2023; London, UK: Cochrane Database of Systematic Reviews.

170. Akl EA, Kahale LA, Agarwal A, Al-Matari N, Ebrahim S, Alexander PE, et al. Impact of missing participant data for dichotomous outcomes on pooled effect estimates in systematic reviews: a protocol for a methodological study. Systematic Reviews. 2014;3(1):137.

171. Murad MH, Wang Z. Guidelines for reporting meta-epidemiological methodology research. BMJ Evidence-based Medicine. 2017;22(4):139-42.

172. Goldkuhle M, Kreuzberger N, Hirsch C, Iannizzi C, Bora AM, Bender R, et al., editors. Characteristics, reporting, and methods of trials included in time-to-event meta-analyses of systematic reviews: A meta-epidemiological review. Abstracts accepted for the 27th Cochrane Colloquium; 2023; London, UK: Cochrane Database of Systematic Reviews.

173. Goldkuhle M, Kreuzberger N, Hirsch C, Iannizzi C, Zorger AM, Bender R, et al., editors. Eigenschaften, Berichterstattung und Methoden von randomisierten Studien in Meta-Analysen zu Time-to-event-Endpunkten systematischer Reviews. Evidenzbasierte Politik und Gesundheitsversorgung – erreichbares Ziel oder Illusion?; 25 Jahrestagung des Deutschen Netzwerks Evidenzbasierte Medizin; 2024; Berlin, 13.- 15.03.2024: Düsseldorf: German Medical Science GMS Publishing House.

174. Skoetz N, van Dalen EC, Macbeth F, Goldkuhle M, Tudur Smith C, Opiyo N. Application for the CRG Networks Innovation Fund 2019: Instructions for inclusion and presentation of time-to-event outcomes in Cochrane Intervention Reviews – development of training resources. Cochrane Community, CRG Networks Innovation Fund 2019: The Cochrane Collaboration; 2019 [cited 2024 13.06]. Available from: https://community.cochrane.org/sites/default/files/uploads/inline-files/4_Nicole%20Skoetz%20-%20Time%20to%20event%20data%20-%20cancer.pdf.

175. GRADEpro GDT: GRADEpro Guideline Development Tool. McMaster University and Evidence Prime; 2024.

176. Goldkuhle M, Narayan VM, Weigl A, Dahm P, Skoetz N. A systematic assessment of Cochrane reviews and systematic reviews published in high-impact medical journals related to cancer. BMJ Open. 2018;8(3):e020869.

177. Moher D, Tetzlaff J, Tricco AC, Sampson M, Altman DG. Epidemiology and reporting characteristics of systematic reviews. PLoS Med. 2007;4(3):e78.

178. Delgado A, Guddati AK. Clinical endpoints in oncology - a primer. American Journal of Cancer Research. 2021;11(4):1121-31.

179. Kahale LA, Khamis AM, Diab B, Chang Y, Lopes LC, Agarwal A, et al. Meta-analyses proved inconsistent in how missing data were handled across their included primary trials: A methodological survey. Clinical Epidemiology. 2020;12:527-35.

180. Kuitunen I, Nikkilä A, Ponkilainen VT, Uimonen MM, Lohi O. Survival analysis and Cox proportional hazards model reporting in pediatric leukemia studies—a systematic review. SN Comprehensive Clinical Medicine. 2022;5(1):24.

181. Pocock SJ, Clayton TC, Altman DG. Survival plots of time-to-event outcomes in clinical trials: good practice and pitfalls. The Lancet. 2002;359(9318):1686-9.

182. Royston P, Choodari-Oskooei B, Parmar MKB, Rogers JK. Combined test versus logrank/Cox test in 50 randomised trials. Trials. 2019;20(1):172.

183. Rosen K, Prasad V, Chen EY. Censored patients in Kaplan–Meier plots of cancer drugs: An empirical analysis of data sharing. European Journal of Cancer. 2020;141:152-61.

184. Wayant C, Page MJ, Vassar M. Evaluation of Reproducible Research Practices in Oncology Systematic Reviews With Meta-analyses Referenced by National Comprehensive Cancer Network Guidelines. JAMA Oncology. 2019;5(11):1550-5.

185. Agarwal A, Johnston BC, Vernooij RW, Carrasco-Labra A, Brignardello-Petersen R, Neumann I, et al. Authors seldom report the most patient-important outcomes and absolute effect measures in systematic review abstracts. Journal of Clinical Epidemiology. 2017;81:3-12.

186. Alonso-Coello P, Carrasco-Labra A, Brignardello-Petersen R, Neumann I, Akl EA, Vernooij RW, et al. Systematic reviews experience major limitations in reporting absolute effects. Journal of Clinical Epidemiology. 2016;72:16-26.

187. Mendes D, Alves C, Batel-Marques F. Number needed to treat (NNT) in clinical literature: an appraisal. BMC Medicine. 2017;15(1):112.

188. Hildebrandt M, Vervölgyi E, Bender R. Calculation of NNTs in RCTs with time-to-event outcomes: a literature review. BMC Med Res Methodol. 2009;9:21.

189. Skoetz N, Goldkuhle M. EQUATOR Network: Reporting guidelines under development for systematic reviews (META-TTE – Transparent reporting of meta-analyses of time-to-event outcomes based on aggregate data from randomised trials of interventions (registered 27 April 2023)): UK EQUATOR Centre at the Centre for Statistics in Medicine (CSM), NDORMS, University of Oxford; 2023 [updated 10.06.2024; cited 2024 13.06]. Available from: https://www.equator-network.org/library/reporting-guidelines-under-development/reporting-guidelines-under-development-for-systematic-reviews/#META.

190. Tudur C, Williamson PR, Khan S, Best LY. The value of the aggregate data approach in meta-analysis with time-to-event outcomes. Journal of the Royal Statistical Society: Series A (Statistics in Society). 2001;164(2):357-70.

191. Hebert AE, Kreaden US, Yankovsky A, Guo D, Li Y, Lee SH, et al. Methodology to standardize heterogeneous statistical data presentations for combining time-to-event oncologic outcomes. PLoS One. 2022;17(2):e0263661.

192. Salika T, Turner RM, Fisher D, Tierney JF, White IR. Implications of analysing time-to-event outcomes as binary in meta-analysis: empirical evidence from the Cochrane Database of Systematic Reviews. BMC Medical Research Methodology. 2022;22(1):73.

193. Hirst TC, Sena ES, Macleod MR. Using median survival in meta-analysis of experimental time-to-event data. Systematic Reviews. 2021;10(1):292.

194. Weir IR, Marshall GD, Schneider JI, Sherer JA, Lord EM, Gyawali B, et al. Interpretation of time-to-event outcomes in randomized trials: an online randomized experiment. Annals of Oncology. 2019;30(1):96-102.

195. Wei Y, Royston P, Tierney JF, Parmar MK. Meta-analysis of time-to-event outcomes from randomized trials using restricted mean survival time: application to individual participant data. Statistics in Medicine. 2015;34(21):2881-98.

196. Everest L, Blommaert S, Tu D, Pater JL, Hay A, Cheung MC, et al. Validating restricted mean survival time estimates from reconstructed Kaplan-Meier data against original trial individual patient data from trials conducted by the Canadian Cancer Trials Group. Value in Health. 2022;25(7):1157-64.

197. Schulz KF, Altman DG, Moher D, the CG. CONSORT 2010 Statement: updated guidelines for reporting parallel group randomised trials. BMC Medicine. 2010;8(1):18.

198. Duchateau L, Pignon JP, Bijnens L, Bertin S, Bourhis J, Sylvester R. Individual patient-versus literature-based meta-analysis of survival data: time to event and event rate at a particular time can make a difference, an example based on head and neck cancer. Controlled Clinical Trials. 2001;22(5):538-47.

199. Smith CT, Williamson PR, Marson AG. Investigating heterogeneity in an individual patient data meta-analysis of time to event outcomes. Statistics in Medicine. 2005;24(9):1307-19.

200. Meddis A, Latouche A, Zhou B, Michiels S, Fine J. Meta-analysis of clinical trials with competing time-to-event endpoints. Biomedical Journal. 2020;62(3):712-23.

201. Rydzewska LHM, Stewart LA, Tierney JF. Sharing individual participant data: through a systematic reviewer lens. Trials. 2022;23(1):167.

202. Butcher NJ, Monsour A, Mew EJ, Chan A-W, Moher D, Mayo-Wilson E, et al. Guidelines for reporting outcomes in trial reports: the CONSORT-outcomes 2022 extension. JAMA. 2022;328(22):2252-64.

203. Rimawi M, Hilsenbeck SG. Making sense of clinical trial aata: is inverse probability of censoring weighted analysis the answer to crossover bias? Journal of Clinical Oncology. 2012;30(4):453-8.

204. Bender R, Beckmann L, Lange S. Biometrical issues in the analysis of adverse events within the benefit assessment of drugs. Pharmaceutical Statistics. 2016;15(4):292-6.

205. Allignol A, Beyersmann J, Schmoor C. Statistical issues in the analysis of adverse events in time-to-event data. Pharmaceutical Statistics. 2016;15(4):297-305.

206. Nishikawa M, Tango T, Ogawa M. Non-parametric inference of adverse events under informative censoring. Statistics in Medicine. 2006;25(23):3981-4003.

207. Latouche A, Allignol A, Beyersmann J, Labopin M, Fine JP. A competing risks analysis should report results on all cause-specific hazards and cumulative incidence functions. Journal of Clinical Epidemiology. 2013;66(6):648-53.

208. Saad ED, Zalcberg JR, Péron J, Coart E, Burzykowski T, Buyse M. Understanding and communicating measures of treatment effect on survival: can we do better? JNCI: Journal of the National Cancer Institute. 2018;110(3):232-40.

209. Goldkuhle M, van Dalen E, Macbeth F, Skoetz N. A survey of Cochrane editors revealed several problem areas related to time-to-event (meta-)analyses. Advances in Evidence Synthesis: special issue Cochrane Database of Systematic Reviews. 2020((9 Suppl 1)):353.

210. Huang X, Wolfe RA. A frailty model for informative censoring. Biometrics. 2002;58(3):510-20.

211. 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.

212. Lee S-Y, Wolfe RA. A Simple Test for Independent Censoring under the Proportional Hazards Model. Biometrics. 1998;54(3):1176-82.

213. 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(3):779-88.

214. Siannis F. Sensitivity analysis for multiple right censoring processes: investigating mortality in psoriatic arthritis. Statistics in Medicine. 2011;30(4):356-67.

215. Siannis F, Copas J, Lu G. Sensitivity analysis for informative censoring in parametric survival models. Biostatistics (Oxford, England). 2005;6(1):77-91.

216. Vale CL, Tierney JF, Stewart LA. Effects of adjusting for censoring on meta-analyses of time-to-event outcomes. Int J Epidemiol. 2002;31(1):107-11.

217. Amzal B, Wiecek W, Obadia T, Benzaghou F. Interval-Censored Survival Data Analysis: Learnings From Phase Iii Trial In Prostate Cancer. Value in Health. 2016;19(3):A98-A9.

218. Ananthakrishnan R, Green S, Previtali A, Liu R, Li D, LaValley M. Critical review of oncology clinical trial design under non-proportional hazards. Critical Reviews in Oncology/Hematology. 2021;162:103350.

219. Royston P, Parmar MKB. Augmenting the logrank test in the design of clinical trials in which non-proportional hazards of the treatment effect may be anticipated. BMC Medical Research Methodology. 2016;16(1):16.

220. Bonofiglio F, Beyersmann J, Schumacher M, Koller M, Schwarzer G. Meta-analysis for aggregated survival data with competing risks: a parametric approach using cumulative incidence functions. Research Synthesis Methods. 2016;7(3):282-93.

221. Deng L, Hsu C-Y, Shyr Y. Power and sample sizes estimation in clinical trials with treatment switching in intention-to-treat analysis: a simulation study. BMC Medical Research Methodology. 2023;23(1):49.

222. Haslam A, Prasad V. When is crossover desirable in cancer drug trials and when is it problematic? Annals of Oncology. 2018;29(5):1079-81.

223. Hernán MA, Robins JM. Estimating causal effects from epidemiological data. J Epidemiol Community Health. 2006;60(7):578-86.

224. Higgins JPT, Li T, Deeks JJ. Chapter 6: choosing effect measures and computing estimates of effect. 2023. In: Cochrane HandbookCochrane Handbook for Systematic Reviews of Interventions [Internet]. Cochrane. Available from: www.training.cochrane.org/handbook.

225. Senn S. Seven myths of randomisation in clinical trials. Statistics in Medicine. 2013;32(9):1439-50.

226. Holmberg MJ, Andersen LW. Adjustment for baseline characteristics in randomized clinical trials. JAMA. 2022;328(21):2155-6.

227. Voils CI, Crandell JL, Chang Y, Leeman J, Sandelowski M. Combining adjusted and unadjusted findings in mixed research synthesis. J Eval Clin Pract. 2011;17(3):429-34.

11. Appendix

11.1. Paper 1: Characteristics, methods and reporting of systematic reviews that include meta-analyses of time-to-event outcomes

11.1.1. Work shares

Authors: **Goldkuhle M**, Hirsch C, Iannizzi C, Bora AM, Bender R, Van Dalen EC, Hemkens LG, Monsef I, Kreuzberger N, Skoetz N

Contributions by the doctoral student:
Ahead of this sub-project and before the beginning of the meta-epidemiological study, the doctoral student was involved in acquiring the necessary funding for the project, which were ultimately obtained by Prof. Skoetz from the German Research Foundation (DFG): He developed the first draft of the grant application, outlined the project and established the project schedule. Both were subsequently reviewed and revised by Prof. Skoetz before submission.
After successful application for funding, the meta-epidemiological study for this sub-project was designed under the doctoral student's lead and his administration. He was responsible for the conceptualization of the sub-project and the development of all applied methods. All necessary tools and resources, such as data extraction forms and the analysis plan, were developed by the doctoral student and then reviewed by Nina Kreuzberger and Prof. Skoetz. Accordingly, the project protocol was published under the doctoral students' last authorship in a publicly available registry (osf.io/6825g/).

During the course of the sub-project, the doctoral student selected relevant literature and, in form of the data extraction from systematic review publications, collected the relevant study data. The methodological standards of the project required a duplicate, independent performance of all relevant project steps, in particular the literature selection and data extraction, which were thus additionally also supported by other project participants. Conflicts were resolved by a third independent project participant. The doctoral student analyzed the collected data and was responsible for their presentation and interpretation. These steps were reviewed by other project members for quality assurance.
Finally, the doctoral student prepared the publication of the study, integrated all comments of the co-authors and revised the article according to the received peer review, again coordinating the advice of the co-authors.

Co-author contributions:
Prof. Skoetz supervised the appropriate and adequate performance of the meta-epidemiological study. She supported the doctoral student in the conceptualization of the sub-project and in the development of its underlying methods. Furthermore, she resolved potential conflicts during literature selection and data extraction and revised the doctoral students drafts for the publication. Nina Kreuzberger (Cologne), Caroline Hirsch (Cologne), Claire Iannizzi (Cologne) and Ana-Mihaela Bora (Cologne) performed the required independent and duplicate review of the data extraction process, supported the data analysis and provided important comments on the manuscript for publication. Nina Kreuzberger screened the relevant literature independently and in duplicate with the doctoral student. She was involved in the project's conceptualization and the development of the methodology through providing important comments on the doctoral student's original drafts. Therefore, she is the first author in the publicly available

study protocol. Prof. Ralf Bender (Cologne), Dr. Elvira van Dalen (Utrecht, Netherlands), and PD Dr. Lars Hemkens (Basel, Switzerland) also provided important comments for the conceptualization of the project and revised the drafts of the publication. The systematic database searches were conducted by the information specialist Ina Monsef (Cologne).

11.1.2. Publication appendix

Appendix

Appendix A1: Complete search strategy for non-Cochrane reviews

Medline on February 8th, 2021

#	Searches
1	"time-to-event".tw,kf.
2	"log rank".tw,kf.
3	survival.tw,kf.
4	hazard.tw,kf.
5	Kaplan-meier estimate/
6	kaplan-meier.tw,kf.
7	(method* adj1 (product* or limit*)).tw,kf.
8	(cumulative* adj1 incidence*).tw,kf.
9	outcome expectation.tw,kf.
10	(cox adj2 (model* or proportional*)).tw,kf.
11	proportional hazards models/
12	or/1-11
13	(randomi?ed or placebo or randomly).ab.
14	meta analysis.mp,pt.
15	12 and 13 and 14
16	limit 15 to dt=20170101-20200801

Appendix A2: List of extraction items

Extraction items for reviews

#	Feld	Options	Description
1	ID of the assessed reviews		Individual number of review
2	Trial level extraction?	Yes; No	
3	Type of assessed review	Cochrane review; Non-Cochrane review	
4	Last name of first author		
5	Publication year		
6	If non-Cochrane review: Journal?		Please insert the full name of the journal that published the assessed review
7	Review update?		Is this a review update?
8	Medical field	Infections and parasitic diseases; Neoplasms (oncological studies irrespective of the field are sorted into this category); Diseases of the blood, blood-forming organs and the immune mechanism; Endocrine, nutritional and metabolic diseases; Mental and behavioral disorders; Diseases of the nervous system; Diseases of the eye and adnexa; Diseases of the ear and mastoid process; Diseases of the circulatory system; Diseases of the respiratory system; Diseases of the digestive system; Diseases of the skin and subcutaneous tissue; Diseases of the musculoskeletal system and connective tissue; Diseases of the genitourinary system; Pregnancy, childbirth and the puerperium	
9	Multiple comparisons?	Yes; No	
10	Comments on review information	Comments on general review information	
11	Medical condition		Medical condition of the assessed population e.g., Melanoma, 1st line, NSCLC, Squamous, 1st line, Brain Tumor WHO Grading II, 1st line "Not specified": if not particular stage of disease is defined, an eligibility criterion or otherwise identifiable
12	Clinical stage		Clinical stage of condition of assessed population E.g., primary occurrence of disease, primary recurrence of disease, multiple recurrences of disease, chronic disease etc.
13	What was the assessed age group?	Adults; Pediatric; Both; Not reported	Adults if participants included with upper age limit in adult population (e.g., ≥18 years, ≥16 years) Pediatric if limited to youth population (e.g., <18 years) Both if explicitly not age limit applied Interpretation: „Adults“ = Adults or not otherwise reported (besides explicit reporting, age in included trials and relevance of disease in population feasible criteria for judgement)
14	Experimental intervention	Biologics/ drug; Surgical procedure; Medical devices; Behavioral intervention; Exercise intervention; Screening; Radiation; Absence of intervention; Other (please specify)	If possible, choose experimental intervention reported in first Summary of Findings table, otherwise use experimental intervention specified in abstract
15	Specify experimental intervention		
16	Control intervention	Placebo; No treatment; Usual care; Biologics/ drugs; Surgical procedure; Medical devices; Behavioral intervention; Exercise intervention; Screening; Radiation; Other (please specify); Best supportive care/ Optimal medical care; Observation	If possible, choose control intervention reported in first Summary of Findings table, otherwise use control intervention specified in abstract
17	Specify control intervention		
18	Comparator treatment considered?	Yes (Control group received intervention); Yes (Experimental group received control); Yes (Both possible); No	Reception of a comparator treatment in either intervention group explicitly considered?
19	Planned outcome number		Number of planned outcomes "Not clear" = if counting of outcomes not possible "Not reported" = no planned outcomes reported e.g., overall survival, progression-free survival, ... If "adverse events" with no further specification, please consider it as a single outcome, otherwise, count specification together with "Adverse events: ..., Adverse events: ..."

20	Planned TTE outcome number		Number of planned time-to-event outcomes "Not clear" = if counting of outcomes not possible "Not reported" = no planned outcomes reported e.g., overall survival, progression-free survival, ... If "adverse events" with no further specification, please consider it as a single outcome, otherwise, count specification together with "Adverse events: ..., Adverse events: ..."
21	List of planned TTE outcomes		Please list the names of all outcomes that were to be assessed in the review as time-to-event outcomes Names only, no definition, e.g., overall survival, progression free survival; Outcomes initially planned to be assessed as time-to-event but analyzed as binary should here be counted as TTE
22	Planned follow-up of review		Specify time-frame if it was mentioned (including exact durations and minimum durations; please copy and paste a respective sentence or describe appropriately) Enter "Not reported" if no time-frame or any specification for follow-up duration was mentioned Enter "Unclear" otherwise
23	Comments on PICO or time-frame		Comments regarding the review PICO or time-frame
24	Studies in quantitative analysis		Number of studies included in quantitative analysis If number included reported without specification use this number If number included in "qualitative synthesis" and "quantitative synthesis"/ meta-analysis are reported separately, please choose the number included in "quantitative synthesis"/ meta-analysis If necessary type: "Unclear" "Not reported"
25	Total population in review		Size (number) of the total population included in quantitative analysis? If number included reported without specification use this number If number included in "qualitative synthesis" and "quantitative synthesis"/ meta-analysis are reported separately, please choose the number included in "quantitative synthesis"/ meta-analysis If necessary type: "Unclear" "Not reported"
26	Experimental population in review		Size (number) of the experimental population included in the review? Only if number explicitly reported (number of participants in arms from forest plot relevant on review outcome level) If number included reported without specification use this number If number included in "qualitative synthesis" and "quantitative synthesis"/ meta-analysis are reported separately, please choose the number included in "quantitative synthesis"/ meta-analysis If necessary type: "Unclear" "Not reported"
27	Control population in review		Size (number) of the control population included in the review? Only if number explicitly reported (number of participants in arms from forest plot relevant on review outcome level) If number included reported without specification use this number If number included in "qualitative synthesis" and "quantitative synthesis"/ meta-analysis are reported separately, please choose the number included in "quantitative synthesis"/ meta-analysis

			If necessary type: "Unclear" "Not reported"
28	Number of outcomes analyzed		Number of all outcomes analyzed Number of main meta-analyses, disregard narrative, sensitivity and subgroup analyses If necessary type: "Unclear" "Not reported"
29	Number of TTE outcomes analyzed		Number of outcomes analyzed as TTE outcomes Number of main meta-analyses only, disregard narrative, sensitivity and subgroup analyses; Outcomes that were initially planned to be assessed as TTE outcomes but analyzed as binary outcomes (e.g., due to data issues) in the review should NOT be counted here. If necessary type: "Unclear" "Not reported"
30	Comments regarding the sample size		
31	HR / log(HR) in analysis	HR/ log(HR) not further specified; HR/ log(HR) (Cox model); HR/ log(HR) (log-rank); HR/ log(HR) (parametric model); HR/ log(HR) (Calculated from Kaplan Meier); other (specify); unclear; not reported	What types of hazard ratios/ log(HR)s did the review authors plan to include in their time-to-event meta-analyses?
32	Methods to obtain TTE data	HR and confidence intervals; HR together with other information (e.g., events in each arm, total events, etc.); O & E or hazard rates on research and control arm; O-E together with logrank V; P-value together with additional information (e.g., events, total events, etc.); Survival curves; Specified particular set of methods (e.g., Tierney 2008, Cochrane Handbook, etc.); Unclear; Not reported; Other (specify); log(HR) and standard error; Median survival times; Time point specific survival times; IPD (recalculated or from publication)	Methods the review authors specified to obtain time-to-event data from trials to pool time-to-event outcomes
33	Comment comparative effect measures		Please make a comment if the authors specified other methods to calculate HRs (e.g., parametric models)
34	Types of analysis for meta-analyses	ITT; mITT; PP; As treated; Unclear; Not reported	Types of analysis preferred to be included in main meta-analyses of review mITT = "justified" exclusion of participants post randomization (e.g., ineligible participants; exclusion before reception of study treatment Explicit preference or inclusion of presence/ absence of ITT in RoB assessments analyses can also give a hint on the target analysis of authors
35	Types of analyses in meta-analyses	ITT; mITT; PP; As treated; Unclear; Not reported; Other (specify); Not reported for all trials; Included trial(s) did not report type of analysis	Was it explicitly reported which analyses from trials were included in meta-analyses of this review? Assessment on review level - if there is any specification of included analyses on review outcome level, please mark this item as „not reported“ and use respective field on review outcome level It should be clear and explicitly reported which type of analysis was included from which trial (analyses reported only for individual trials are counted as "not reported for all trials"
36	Adjusted, unadjusted or both types of effects in meta-analyses	Adjusted only; Unadjusted only; Hierarchical (unadjusted before adjusted); Hierarchical (adjusted before unadjusted); Both; Unclear; Not reported	Did the review authors specify to include adjusted, unadjusted and/or both types of effects in the main meta-analyses of this review?
37	Handling of adjusted and unadjusted HR in meta-analysis	Combined in meta-analysis; Sensitivity analysis (two separate analyses); Other (please specify); Mentioned as RoB criterion in methods; Not applicable (unadjusted/ adjusted not mentioned); Unclear; Not reported	How did the authors plan to handle adjusted and unadjusted HRs in meta-analyses of this review?
38	Stratified effects eligible?	Yes; No; Unclear; Not reported	Please indicate whether stratified effect measures (e.g., stratified HRs) were eligible for meta-analysis? Please choose yes only, if stratified effect measures where explicitly mentioned. Stratification can be seen as a form of adjustment. Not to be confused with randomization stratification or subgroup-analyses (analysis-results per stratum).
39	Comments on handling of adjusted and unadjusted hazard ratios		
40	Methods to pool TTE data	Inverse variance; Peto (fixed-effects) model; Other (specify); Unclear; Not reported	Which methods were planned to pool time-to-event data? (irrespective whether reported in forest plot; data in forest plots is relevant for the assessment on review outcome level)
41	Handling of heterogeneity	Random-effects meta-analysis performed; Subgroup analyses; No pooling if too heterogeneous; Other; Unclear; Not reported	How did the authors intent to handle heterogeneity between studies?

			(Refers to planned handling of heterogeneity)
42	TTE meta-analysis models	fixed effects only; random effects only; mixed (either one as sensitivity analysis); fixed; and if not possible random effects; unclear; not reported	Model type for time-to-event meta-analyses
43	Other pooled TTE outcome measures	none; median survival time; restricted mean survival time; rank preserving structural failure; other (please specify); relative risk	Pooled analysis with other outcome measures besides HR for time-to-event outcomes? e.g., pooling of median survival times; standard error of the combined log(MST)
44	Comments on meta-analytic methods		
45	Dealing with varying follow-up	Sensitivity analyses (e.g., studies with shorter/longer follow-up time in separate analysis); Exclusion of studies with divergent follow-up time; Mentioned as RoB criterion in methods; Other (please specify); Not applicable, pre-defined timing as inclusion criterion; Unclear; Not reported	How did the authors intent to deal with varying follow-up times between the included trials?
46	Comments on treating follow-up times		Comments regarding the treatment of variable follow-up times
47	Handling of competing events	Subgroup analysis (e.g., according to competing event rate); Mentioned as RoB criterion in methods; Exclusion of trials above a certain rate of competing events; Other (please specify); Not applicable, no outcomes with potential competing events; Unclear; Not reported	How did the review authors plan to deal with competing events in time-to-event analyses of included trials?
48	Comments on treating competing events		Comments regarding the treatment of competing events in the meta-analysis for this outcome
49	Handling of MOD	Recalculation where possible; Single imputation; Multiple imputation; Meta-regression; Sensitivity analyses (according to rate of missing values); Mentioned as RoB criterion in methods; Contact with authors; Other (please specify); Unclear; Not reported	How did the authors intent to deal with missing outcome data in the included trials
50	Comments on the treatment of missing outcome data		Comments on the treatment of missing outcome data
51	Handling of non-administrative censoring	Sensitivity analysis (e.g., according to rate of censoring); Exclusion of trials (e.g., according to rate of censoring); Single imputation; Multiple imputation; Meta-regression; Mentioned as RoB criterion in methods; Other (please specify); Unclear; Not reported	How did the review authors intend to deal with censoring for non-administrative reasons (informative censoring) in the time-to-event analyses of the included trials? Example for administrative reasons: end-of-study censoring
52	Comments on treating non-administrative censoring		Comments on treating non-administrative censoring
53	Handling of comparator treatments in participants	Complies with review PICO; No handling mentioned, probably ITT; Intention-to-treat; Per protocol; As treated; Sensitivity analysis (e.g., According to rate of participants); Mentioned as RoB criterion in methods; Other (please specify); Unclear; Not reported	How did the review authors intend to deal with the reception of comparator treatments in trial participants?
54	Comments on dealing with participants receiving comparator treatments		Comments regarding the treatment of trial participants receiving comparator treatments
55	Proportional hazards assessment	Recalculation of IPD; Use of trial level tests; Use of other trial level data (e.g., visual inspection of survival curves); Both: trial level tests, if not provided inspection of curve; Unclear; Not reported;	How did the authors intend to assess the proportionality o hazards in included trials?
56	Handling non-proportional hazards	No action, all trials included regardless of prop. hazard assumption; Sensitivity analysis (e.g., according to degree of non-proportionality); Other (please specify); Not applicable (No assessment of proportionality reported); Unclear; Not reported	How did the authors intend to deal with non-proportional hazards in included trials?
57	Comments on proportionality of hazards		Comments on proportionality of hazards
58	RoB tool	no RoB assessment; RoB 1, study level; RoB 1, outcome level; RoB 2; Other	Which tool did the authors intend to use to assess bias in included trials?
59	TTE specific RoB assessment?	Yes; No; Unclear; Not applicable (no RoB assessment)	Did the authors intend to assess TTE specific trial characteristics in their RoB assessment?
60	If yes or unclear, please specify (RoB)		If yes or unclear, please specify (RoB)
61	Use of GRADE	Yes; No	
62	TTE specific GRADE assessment?	Yes; No; Unclear; Not applicable (no GRADE assessment)	Did the authors intend to include time-to-event specific aspects in their GRADE rating?
63	If yes or unclear, please specify (GRADE)		If yes or unclear, please specify (GRADE)
64	Summary of findings tables for TTE outcomes?	Yes; No	
65	Comments on risk of bias and GRADE		Comments on risk of bias and GRADE
66	Competing events discussed?	In results; In discussion; Not reported; Not applicable	Did the authors discuss competing events and/or their potential impact on review results?
67	Heterogenous outcome definitions discussed?	In results; In discussion; Not reported; Not applicable	Did the authors discuss heterogeneity of outcome definitions among included studies and/or its potential impact on review results?
68	Difference in follow-up times of trials discussed?	In results; In discussion; Not reported; Not applicable	Did the authors discuss differential trial follow-up for outcomes and/or their potential impact on review results?

69	MOD discussed?	In results; In discussion; Not reported; Not applicable	Did the authors discuss missing outcome data in included trials and/or its potential impact on review results?
70	Non-administrative censoring discussed?	In results; In discussion; Not reported; Not applicable	Did the authors discuss censoring for non-administrative reasons (informative censoring) and/or its potential impact on review results?
71	Reception of comparator treatments discussed?	In results; In discussion; Not reported; Not applicable	Did the authors discuss the reception of comparator treatments in trial participants and/or its potential impact on review results?
72	Adjusted or unadjusted estimates discussed?	In results; In discussion; Not reported; Not applicable	Did the authors discuss the inclusion of adjusted and unadjusted estimates for the same outcomes and/or its potential impact on review results?
Abbreviations: HR = hazard ratio, ITT, = intention to treat, mITT = modified intention to treat, MOD = missing outcome data, MST = mean survival time, PICO = population-intervention-comparator-outcomes, RoB = risk of bias			

Extraction items for individual review time-to-event outcomes

#	Item	Options	Description
1	Review ID		Review ID (number) assessable from Excel
2	Review outcome		Outcome ID (number) assessable from Excel Please use review description (without definition)
3	Primary outcome?	Yes; No; Not applicable (No primary/ secondary outcome defined)	Was this outcome specified as a primary outcome of the review?
4	Complete outcome definition		Please provide the complete definition of the outcome
5	Composite outcome	Yes; No; Unclear; Not reported	Does this outcome include multiple outcome events? for example progression free-survival - progression, relapse and death from any cause
6	Composite events described?	Yes; No; Not applicable ("composite outcome" unclear or not reported)	Were the outcome events composing this composite outcome described?
7	All-cause mortality part of outcome?	Yes; No; Unclear	Was all-cause mortality a component of the assessed outcome?
8	Competing events possible?	Yes; No; Unclear	Are competing events possible by definition of the outcome? Choose yes, e.g., if overall mortality was not part of the defined outcome
9	Reported as event or absence of event?	Event; Absence of event; Both (with reasoning); Unclear (both without reasoning, e.g., switching of reporting)	Was the outcome reported as event (e.g., death, relapse) or absence of event (e.g., overall survival, progression-free survival)? Refers to the description of the outcome overall (e.g., in the methods section, headlines, etc.) (Presentation of the pooled results is assessed in a subsequent item)
10	Follow-up pre-specified?	Time-specific (12 months, 2 year, 10 year, ...); Longest follow-up; Minimum duration of follow-up required; Maximum duration of follow-up specified; Not reported	Was a duration of follow-up included in definition or otherwise pre-specified for the outcome?
11	Start of outcome assessment defined	Yes; No; Not applicable (e.g., outcome not defined); Unclear	Was the start of outcome assessment for this outcome included in the outcome definition or otherwise prespecified in the statistical methods section?
12	Outcome assessment start	Randomization; Enrollment; Allocated treatment; Previous treatment (e.g., surgery); Other (specify); Not applicable (e.g., start of follow-up not reported); Unclear	What was the defined or otherwise reported start time-point of outcome assessment for this outcome? MG: Nachträglich eingefügt: Für bereits extrahierte: Wenn auf Basis der extrahieren Definitionen nicht nachverfolgbar bitte offen lassen (ich sollte es extrahiert haben)
13	Field for commenting on outcome		Field for commenting on the outcome (e.g., what was unclear, whether name changes occurred during the reviews (OS to All cause mortality), other irregularities, etc.)
14	Types of analyses ELIGIBLE for meta-analysis	ITT; mITT; PP; Other (please specify); Unclear; Not reported	Please indicate what types of analyses (e.g., ITT, PP) were ELIGIBLE in the assessed meta-analysis
15	Types of analyses INCLUDED in meta-analysis	ITT; mITT; PP; Other (please specify); Unclear; Not reported	Please indicate what types of analyses (e.g., ITT, PP) were INCLUDED in the assessed meta-analysis (as reported by the review authors)
16	Comments on eligible or included analyses		
17	Varying follow-up times handling specifically reported	Yes; No	Was handling of varying follow-up specifically reported for this outcome?
18	Handling of varying follow-up times	Sensitivity analysis (studies with shorter/ longer follow-up time in separate analysis); Exclusion of studies with divergent follow-up time; Other (please specify); Pre-defined timing as inclusion criterion; Not applicable	Please indicate whether and how the authors dealt with varying follow-up times between the included trials for this meta-analysis
19	Comments on variable follow-up times in		

	trials included in meta-analysis		
20	Competing events handling specifically reported	Yes; No	Was handling of competing events specifically reported for this outcome? ("No" also when competing events not possible)
21	Handling of competing events	Sensitivity analysis (e.g., according to competing event rate); Exclusion of studies above a certain rate of competing events; Other (please specify); No outcome with potential competing events	Please indicate how the authors reported to or treated competing events in the meta-analysis of this outcome
22	Commenting on competing events in the meta-analysis		Comments on competing events in the meta-analysis of this outcome
23	Missing data handling specifically reported	Yes; No	Was handling of missing data specifically reported for this outcome
24	Handling of MOD	Recalculation where possible; Contact with authors; Single imputation; Multiple imputation; Meta-regression; Sensitivity analyses (according to rate of missing values); Hierarchy of the above (please specify); Other or multiple (please specify)	How did the authors treat missing outcome data in the meta-analysis of this outcome?
25	Comments on MOD in the meta-analysis		Comments on missing outcome data in the meta-analysis of this outcome
26	Censoring handling specifically reported	Yes; No	Was handling of censoring for non-administrative reasons reported specifically for this outcome? "Censoring" should explicitly be mentioned for a judgement
27	Treatment of censoring for non-administrative reasons	Sensitivity analysis (e.g., according to rate of censoring); Exclusion of trials (e.g., according to rate of censoring); Single imputation; Multiple imputation; Meta-regression; Other or multiple (please specify)	Please indicate how the authors treated censoring of participants for non-administrative reasons (informative censoring) in the meta-analysis of this outcome
28	Comments on treatment of non-administrative censoring		Comments on the treatment of non-administrative (informative) censoring in the meta-analysis of this outcome
29	Comparator treatments handling specifically reported	Yes; No	Was handling the reception of comparator treatments specifically reported for this outcome?
30	Treatment of comparator treatments	Complies with review PICO; No handling mentioned, ITT; Per protocol analysis; Sensitivity analysis (e.g., According to rate of participants); Other or multiple (please specify)	How did the authors treat the reception of comparator treatments in trial participants in the meta-analysis of this outcome
31	Comments on comparator treatments in meta-analysis		Comments on the treatment of trial participants receiving comparator treatments in the meta-analysis for this outcome
32	Absolute effect measures reported?	Yes; No; Absolute effects explicitly not calculated ("e.g., not calculable because of TTE outcome")	Where absolute effect measures based on the outcomes of this meta-analysis calculated and provided?
33	Type of absolute effects	Natural frequencies; Risk difference; NNT; Median survival or difference in median survival; Not applicable	What type of absolute effects where calculated and reported?
34	Baseline risk applicable for events or absence of events	Event; Absence of event; Unclear; Not applicable	If a baseline risk (e.g., control group risk) was used for the calculation of absolute effects (e.g., in a SoF table), was it applicable for events or absence of events Event (e.g., death (mortality), relapse, etc.): Number of individuals with composite outcome (e.g., "relapse or death" (falsely: PFS)) HIGHER than single event outcome (e.g., "mortality" (falsely: OS)) (when composite is including the single event outcome) (e.g., "Relapse or death"/ PFS: 200 vs. "Mortality"/OS: 100) Absence of event (e.g., OS, PFS, EFS): Number of individuals with composite outcome (e.g., PFS (falsely: "relapse or death")) LOWER than single event outcome (e.g., OS (falsely: "mortality")) (when composite is including the single event outcome) (e.g., "Relapse or death"/ PFS: 800 vs. "Mortality"/OS: 900) Unclear: No reported or conflicting information which type of baseline risk was chosen (e.g., baseline risks from multiple sources so that the above rule does not apply, same baseline risk for different outcomes (e.g., OS=PFS), no comparison possible)
35	Description of outcome event or direction changed for calculation	No (no changes); Yes (description changed with reasoning); Yes (description changed without reasoning); Yes (HR inverted, with reasoning); Yes (HR inverted, without reasoning); Unclear; Not applicable	Was the description of the type of outcome event (absence of event to event (overall survival to all cause mortality) or the direction of the effect estimator (e.g., by inversion) changed for the calculation of absolute effects?
36	Absolute effects correct?	Yes; No; Correct calculation but wrong labeling; Unclear (e.g., not clear what the baseline risk is applicable for or unclear whether the HR corresponds to absence of events or events); Not applicable	Were absolute effects calculated correctly for this outcome? (false labelling refers to e.g., mortality (event) labelled as overall survival (event-free)) Please check the SoF table first
37	Absolute effects comment		
38	HR type specifically reported	Yes; No	Were the type or characteristics of the summarized hazard ratio (e.g., Cox, parametric model, log-rank, etc.) specifically described for this outcome?
39	Summarized HR	HR/ log(HR) not specified; HR/ log(HR) (Cox model); HR/ log(HR) (log-rank); HR/ log(HR) (parametric model); HR/ log(HR)	Type of HR summarized in this analysis?

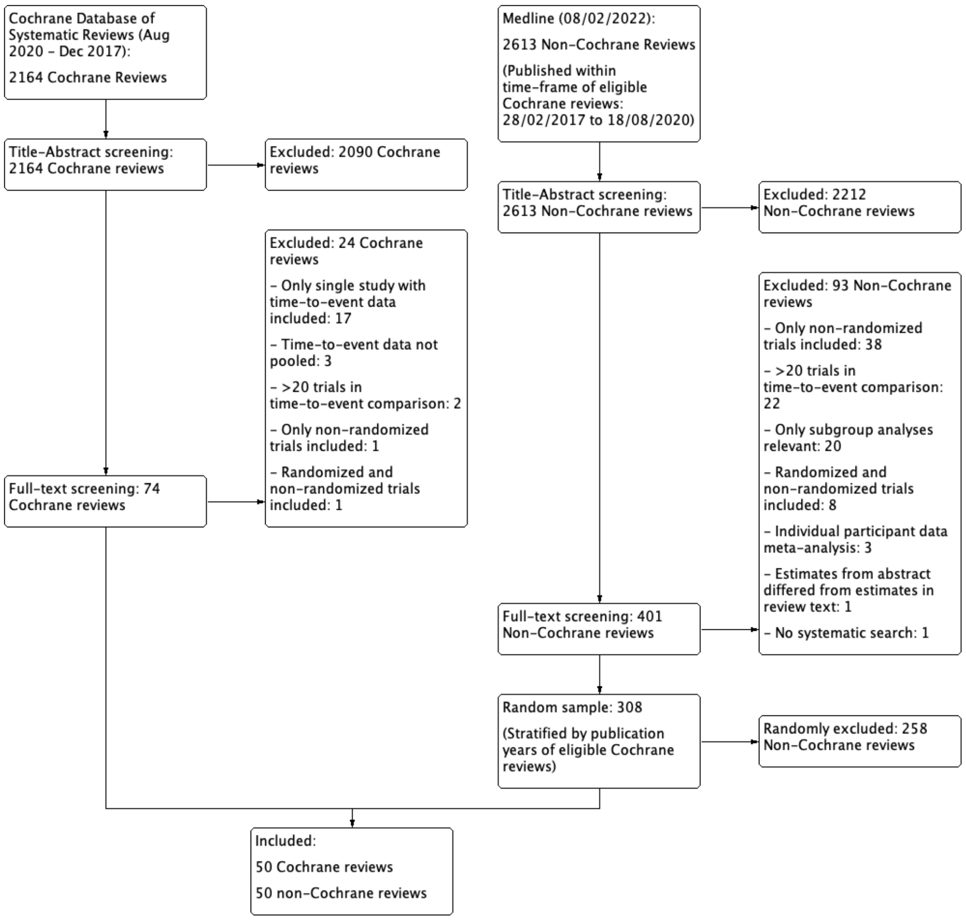
		(Calculated from Kaplan Meier); Other (specify); Unclear; Not reported	
40	Comments on HR		Comments on the hazard ratio that was included in the meta-analysis
41	Recalculation of HRs specifically reported	Yes; No	Did the authors report on whether and/or how they recalculated HRs specifically for this outcome? Choose "no" of only review level or no information was reported
42	Retrieval of TTE data	HR and confidence intervals; HR together with other information (e.g., events in each arm, total events, etc.); O & E or hazard rates on research and control arm; O-E together with logrank V; P-value together with additional information (e.g., events, total events, etc.); Survival curves; Other (specify); Not specified for this outcome; Unclear; Median survival times; Time-point specific survival times; IPD (recalculated or from publication); Reported, but method not clear	How did the authors obtain time-to-event data to pool for this particular outcome?
43	Method to pool TTE data	Inverse variance method; Peto (fixed-effects) method; HKSJ (random-effects); Other (specify); Unclear; Not reported	Which method was used to pool time-to-event data in this meta-analysis?
44	Model used for meta-analysis	Fixed effects; Random effects; Both, one as sensitivity analysis; Unclear; Not reported	Which model was used for this meta-analysis?
45	Heterogeneity handling specifically reported	Yes; No	Did the authors report on how they dealt with heterogeneity between studies specifically for this outcome Choose "no" of only review level or no information was reported
46	Dealing with heterogeneity	Random-effects MA; Subgroup analyses; Other (specify); Unclear; No heterogeneity	How did the authors deal with heterogeneity between studies in the performed meta-analysis?
47	Comment on handling heterogeneity		Comments on the handling of heterogeneity for this particular outcome
48	Proportional hazard handling specifically reported	Yes; No	Did the authors specifically report dealing with proportional hazards for this outcome? Choose "no" of only review level or no information was reported
49	Test for proportionality of hazards	Recalculation of IPD; Use of trial level tests; Use of other trial level data (e.g., visual inspection of survival curves); Both: trial level tests, if not provided inspection of curve; No test of proportional hazards done; Unclear	Please indicate whether and through which test the review authors assessed the proportionality of hazards in trials included in this meta-analysis
50	Non-proportionality of hazards indicated by tests?	Yes; No; Not applicable; Unclear	Did the test for proportionality that review authors performed indicate a problem with non-proportionality of hazards? (According to the review authors)
51	Dealing with (non-)proportionality of hazards	No action, all studies included regardless of prop. hazard assumption; Sensitivity analysis (e.g., According to degree of non-proportionality); Other (please specify); Unclear; Not reported; Not applicable	Please indicate whether and how the review authors dealt with the (non-)proportionality of hazards in the trials included in this meta-analysis
52	Comment on proportionality of hazards in trials included in this meta-analysis		
53	Adjusted and unadjusted HRs specifically reported	Yes; No	Did the authors report on the inclusion of adjusted and unadjusted hazard ratios specifically for this outcome? Choose "no" of only review level or no information was reported
54	Adjusted or unadjusted HRs included?	Adjusted only; Unadjusted only; Hierarchical selection (unadjusted before adjusted); Hierarchical (adjusted before unadjusted); Both; Unclear	Were adjusted or unadjusted hazard ratios included in this meta-analysis?
55	Dealing with with adjusted and unadjusted HR	Combined in same meta-analysis; Sensitivity analysis (two separate analyses); Other (please specify); Not applicable	Please indicate how the review authors dealt with adjusted and unadjusted hazard ratios in the conducted meta-analyses
56	Comments on handling adjusted and unadjusted HRs		Comments on the treatment of adjusted and unadjusted hazard ratios in this meta-analysis
57	Number of studies		Number of studies included in this meta-analysis If not directly reported (e.g., in SoF), please recalculate from forest plot
58	Participants in experimental arm		Total number of participants in experimental arm of the assessed meta-analysis If not directly reported (e.g., in SoF), please recalculate from forest plot Empty field = "Not reported/ Unclear"
59	Participants in control arm		Total number of participants in control arm of the assessed meta-analysis If not directly reported (e.g., in SoF), please recalculate from forest plot Empty field = "Not reported/ Unclear"
60	Total participants		Total number of participants in control arm of the assessed meta-analysis Empty field = "Not reported/ Unclear"

61	Comments on sample size and number of studies		
62	Pooled HR		What was the pooled hazard ratio for this outcome?
63	Lower 95% CI		What was the lower 95% CI of the HR for this outcome?
64	Upper 95% CI		What was the upper 95% CI of the HR for this outcome?
65	Chi²		What was the value of the Chi² statistic in the meta-analysis for this outcome? Empty field = "Not reported/ Unclear"
66	I²		What was the value of the I² statistic for this outcome? Empty field = "Not reported/ Unclear"
67	HR for events or absence of events	Event; Absence of event; Inconsistently (e.g., "OS" in abstract or meta-analysis and "risk of death"/ "mortality" in results for the same HR); Reasonable variation (e.g., "mortality" in results and inverted to "OS" in SoF); Unclear	Was the pooled hazard ratio reported as applicable for events or for absence of events (in the abstract, results section and (where applicable) SoF) Negative events (e.g., death), when HR <1 reported as beneficial Absence of negative event (e.g., OS, PFS) when HR >1 reported as beneficial Positive event when HR >1 reported as beneficial Absence of positive event when >1 HR reported as beneficial
68	HR <1 an increased or decreased risk of event in experimental group	Increased risk; Decreased risk; Unclear	Does a hazard ratio <1 indicate an increased or decreased risk of the event in the group that is assessed as experimental group in the review publication?
69	HRs inverted?	Yes; Unclear; Not reported	Did the authors describe that hazard ratios from trial publications were inverted to correspond to the direction of the hazard ratio in this meta-analysis?
70	Comments on meta-analysis results		Field for comments on the results of the meta-analysis
71	Outcome in SoF	Yes; No; Not applicable	Was the assessed outcome presented in a summary of findings table?
72	Overall GRADE rating	High; Moderate; Low; Very low; Not applicable	What was the overall GRADE rating for the assessed outcome?
73	Study limitations rating	0; 0.5; 1; 2; Not applicable	Did the review authors rate down for study limitations?
74	TTE-specific study limitations	Yes (specify); No; Unclear; Not applicable	Did the review authors considerer TTE-specific study limitations in their GRADE assessment? "Not applicable" = No GRADE rating or not downgraded for this particular domain (e.g., naive inclusion of competing events or informative censoring)
75	Imprecision rating	0; 0.5; 1; 2; Not applicable	Did the review authors rate down for imprecision in their GRADE assessment?
76	TTE-specific imprecision	Yes (specify); No; Unclear; Not applicable	Did the review authors consider TTE-specific sources of imprecision in their GRADE assessment? "Not applicable" = No GRADE rating or not downgraded for this particular domain (e.g., too short follow-up period; high rates of censoring; low number of overall events)
77	Indirectness rating	0; 0.5; 1; 2; Not applicable	Did the review authors rate down for indirectness?
78	TTE-specific indirectness	Yes (specify); No; Unclear; Not applicable	Did the review authors consider TTE-specific sources of indirectness? "Not applicable" = No GRADE rating or not downgraded for this particular domain (e.g., inadequate handling of participants who receive a comparator intervention or the duration of follow-up)
79	Inconsistency rating	0; 0.5; 1; 2; Not applicable	Did the review authors rate down for inconsistency?
80	TTE-specific inconsistency	Yes (specify); No; Unclear; Not applicable	Did the review authors consider TTE-specific sources of inconsistency "Not applicable" = No GRADE rating or not downgraded for this particular domain (e.g., proportional hazards in some of the included studies and non-proportional in others)
81	Comments on TTE-specific aspects of the review authors' GRADE rating		
82	Competing events discussed?	Yes; No; Not applicable	Were competing events mentioned in the discussion?
83	Competing events discussion		Comment on how competing events were discussed
84	Outcome heterogeneity discussed?	Yes; No; Not applicable	Was heterogeneity between outcome definitions mentioned in the discussion for this outcome?

85	Outcome heterogeneity discussion		Comment on how heterogeneity was discussed
86	Varying follow-up times discussed?	Yes; No; Not applicable	Were difference in the follow-up times of trials mentioned in the discussion for this outcome?
87	Follow-up discussion		Comment on how varying follow-up among included trials was discussed
88	Adjusted or unadjusted HRs discussed?	Yes; No; Not applicable	Was the inclusion of adjusted and/or unadjusted hazards ratios in this meta-analysis mentioned in the discussion
89	Adjusted and unadjusted discussion		Comment on how the inclusion of adjusted and/ or unadjusted hazard ratios was discussed
90	MOD discussed?	Yes; No; Not applicable	Was missing outcome data in the included trials mentioned in the discussion for this outcome?
91	MOD discussion		Comment on how missing outcome data was discussed
92	Non-administrative censoring discussed?	Yes; No; Not applicable	Was non-administrative censoring (informative censoring) in the included trials mentioned in the discussion for this outcome?
93	Censoring discussion		Comment on whether censoring was discussed, in particular censoring for non-administrative reasons (informative censoring)
94	Reception of comparator treatments discussed?	Yes; No; Not applicable	Was the reception of comparator treatments in trial participants in the included trials mentioned in the discussion for this outcome?
95	Comparator treatments discussion		Comment on how the reception of comparator treatments was discussed
96	Other comments		Comments on the consideration of time-to-event specific aspects in the review discussion

Abbreviations: CI = confidence interval, HR = hazard ratio, ITT, = intention to treat, mITT = modified intention to treat, MOD = missing outcome data, MST = mean survival time, PICO = population-intervention-comparator-outcomes, RoB = risk of bias

Appendix A3: Flow-diagram


Cochrane Database of Systematic Reviews (Aug 2020 – Dec 2017):
2164 Cochrane Reviews
Title–Abstract screening: 2164 Cochrane reviews
Excluded: 2090 Cochrane reviews
Excluded: 24 Cochrane reviews
– Only single study with time–to–event data included: 17
– Time–to–event data not pooled: 3
– >20 trials in time–to–event comparison: 2
– Only non–randomized trials included: 1
– Randomized and non–randomized trials included: 1
Full–text screening: 74 Cochrane reviews
Medline (08/02/2022):
2613 Non–Cochrane Reviews
(Published within time–frame of eligible Cochrane reviews: 28/02/2017 to 18/08/2020)
Title–Abstract screening: 2613 Non–Cochrane reviews
Excluded: 2212 Non–Cochrane reviews
Excluded: 93 Non–Cochrane reviews
– Only non–randomized trials included: 38
– >20 trials in time–to–event comparison: 22
– Only subgroup analyses relevant: 20
– Randomized and non–randomized trials included: 8
– Individual participant data meta–analysis: 3
– Estimates from abstract differed from estimates in review text: 1
– No systematic search: 1
Full–text screening: 401 Non–Cochrane reviews
Random sample: 308
(Stratified by publication years of eligible Cochrane reviews)
Randomly excluded: 258 Non–Cochrane reviews
Included:
50 Cochrane reviews
50 non–Cochrane reviews

Appendix A4: Search results

Between December 2017 and August 2020, we identified 2164 CR (for summary see Flow diagram in A3). After screening titles and abstracts, we assessed the full texts of 74 CR. Fifty finally eligible CR were published from 28/02/2017 to 18/08/2020. For that timeframe, our search strategy identified 2613 records from Core Clinical Journals, of which we selected 401 for full-text screening. Finally, we drew our random sample of 50 nCR from a total sample of 308 eligible reviews.

Appendix A5: List of included reviews

Cochrane reviews

1. Ameratunga M, Pavlakis N, Wheeler H, Grant R, Simes J, Khasraw M. Anti-angiogenic therapy for high-grade glioma. Cochrane Database of Systematic Reviews. 2018(11).
2. Arora M, Harvey LA, Glinsky JV, Nier L, Lavrencic L, Kifley A, et al. Electrical stimulation for treating pressure ulcers. Cochrane Database of Systematic Reviews. 2020(1).
3. Bala MM, Celinska-Lowenhoff M, Szot W, Padjas A, Kaczmarczyk M, Swierz MJ, et al. Antiplatelet and anticoagulant agents for secondary prevention of stroke and other thromboembolic events in people with antiphospholipid syndrome. Cochrane Database of Systematic Reviews. 2017(10).
4. Blank O, von Tresckow B, Monsef I, Specht L, Engert A, Skoetz N. Chemotherapy alone versus chemotherapy plus radiotherapy for adults with early stage Hodgkin lymphoma. Cochrane Database of Systematic Reviews. 2017(4).
5. Bui KT, Willson ML, Goel S, Beith J, Goodwin A. Ovarian suppression for adjuvant treatment of hormone receptor-positive early breast cancer. Cochrane Database of Systematic Reviews. 2020(3).
6. Bulsara VM, Worthington HV, Glenny AM, Clarkson JE, Conway DI, Macluskey M. Interventions for the treatment of oral and oropharyngeal cancers: surgical treatment. Cochrane Database of Systematic Reviews. 2018(12).
7. Chan DLH, Segelov E, Wong RSH, Smith A, Herbertson RA, Li BT, et al. Epidermal growth factor receptor (EGFR) inhibitors for metastatic colorectal cancer. Cochrane Database of Systematic Reviews. 2017(6).
8. Chin V, Nagrial A, Sjoquist K, O'Connor CA, Chantrill L, Biankin AV, et al. Chemotherapy and radiotherapy for advanced pancreatic cancer. Cochrane Database of Systematic Reviews. 2018(3).
9. Chionh F, Lau D, Yeung Y, Price T, Tebbutt N. Oral versus intravenous fluoropyrimidines for colorectal cancer. Cochrane Database of Systematic Reviews. 2017(7).
10. Claassen YHM, van der Valk MJM, Breugom AJ, Frouws MA, Bastiaannet E, Liefers GJ, et al. Survival differences with immediate versus delayed chemotherapy for asymptomatic incurable metastatic colorectal cancer. Cochrane Database of Systematic Reviews. 2018(11).
11. Coleridge SL, Bryant A, Lyons TJ, Goodall RJ, Kehoe S, Morrison J. Chemotherapy versus surgery for initial treatment in advanced ovarian epithelial cancer. Cochrane Database of Systematic Reviews. 2019(10).
12. Dalal A, Eskin-Schwartz M, Mimouni D, Ray S, Days W, Hodak E, et al. Interventions for the prevention of recurrent erysipelas and cellulitis. Cochrane Database of Systematic Reviews. 2017(6).
13. Egger SJ, Willson ML, Morgan J, Walker HS, Carrick S, Ghersi D, et al. Platinum-containing regimens for metastatic breast cancer. Cochrane Database of Systematic Reviews. 2017(6).
14. El Moheb M, Nicolas J, Khamis AM, Iskandarani G, Akl EA, Refaat M. Implantable cardiac defibrillators for people with non-ischaemic cardiomyopathy. Cochrane Database of Systematic Reviews. 2018(12).
15. Fisher SA, Cutler A, Doree C, Brunskill SJ, Stanworth SJ, Navarrete C, et al. Mesenchymal stromal cells as treatment or prophylaxis for acute or chronic graft-versus-host disease in haematopoietic stem cell transplant (HSCT) recipients with a haematological condition. Cochrane Database of Systematic Reviews. 2019(1).
16. Frost JA, Webster KE, Bryant A, Morrison J. Lymphadenectomy for the management of endometrial cancer. Cochrane Database of Systematic Reviews. 2017(10).
17. Galaal K, Donkers H, Bryant A, Lopes AD. Laparoscopy versus laparotomy for the management of early stage endometrial cancer. Cochrane Database of Systematic Reviews. 2018(10).

18. Haun MW, Estel S, Rücker G, Friederich HC, Villalobos M, Thomas M, et al. Early palliative care for adults with advanced cancer. Cochrane Database of Systematic Reviews. 2017(6).
19. Hickey BE, James ML, Daly T, Soh FY, Jeffery M. Hypofractionation for clinically localized prostate cancer. Cochrane Database of Systematic Reviews. 2019(9).
20. Høeg BL, Bidstrup PE, Karlsen RV, Friberg AS, Albieri V, Dalton SO, et al. Follow-up strategies following completion of primary cancer treatment in adult cancer survivors. Cochrane Database of Systematic Reviews. 2019(11).
21. Hwang EC, Sathianathen NJ, Jung JH, Kim MH, Dahm P, Risk MC. Single-dose intravesical chemotherapy after nephroureterectomy for upper tract urothelial carcinoma. Cochrane Database of Systematic Reviews. 2019(5).
22. Janmaat VT, Steyerberg EW, van der Gaast A, Mathijssen RHJ, Bruno MJ, Peppelenbosch MP, et al. Palliative chemotherapy and targeted therapies for esophageal and gastroesophageal junction cancer. Cochrane Database of Systematic Reviews. 2017(11).
23. Jeffery M, Hickey BE, Hider PN. Follow-up strategies for patients treated for non-metastatic colorectal cancer. Cochrane Database of Systematic Reviews. 2019(9).
24. Jung JH, Risk MC, Goldfarb R, Reddy B, Coles B, Dahm P. Primary cryotherapy for localised or locally advanced prostate cancer. Cochrane Database of Systematic Reviews. 2018(5).
25. Khan L, Soliman H, Sahgal A, Perry J, Xu W, Tsao MN. External beam radiation dose escalation for high grade glioma. Cochrane Database of Systematic Reviews. 2020(5).
26. Kindts I, Laenen A, Depuydt T, Weltens C. Tumour bed boost radiotherapy for women after breast-conserving surgery. Cochrane Database of Systematic Reviews. 2017(11).
27. Küley-Bagheri Y, Kreuzer KA, Monsef I, Lübbert M, Skoetz N. Effects of all-trans retinoic acid (ATRA) in addition to chemotherapy for adults with acute myeloid leukaemia (AML) (non-acute promyelocytic leukaemia (non-APL)). Cochrane Database of Systematic Reviews. 2018(8).
28. Kunath F, Jensen K, Pinart M, Kahlmeyer A, Schmidt S, Price CL, et al. Early versus deferred standard androgen suppression therapy for advanced hormone-sensitive prostate cancer. Cochrane Database of Systematic Reviews. 2019(6).
29. Lee A, Arasaratnam M, Chan DH, Khasraw M, Howell VM, Wheeler H. Anti-epidermal growth factor receptor therapy for glioblastoma in adults. Cochrane Database of Systematic Reviews. 2020(5).
30. Majumdar A, Roccarina D, Thorburn D, Davidson BR, Tsochatzis E, Gurusamy KS. Management of people with early- or very early-stage hepatocellular carcinoma. Cochrane Database of Systematic Reviews. 2017(3).
31. Morrison J, Thoma C, Goodall RJ, Lyons TJ, Gaitskell K, Wiggans AJ, et al. Epidermal growth factor receptor blockers for the treatment of ovarian cancer. Cochrane Database of Systematic Reviews. 2018(10).
32. Norman G, Christie J, Liu Z, Westby MJ, Jefferies JM, Hudson T, et al. Antiseptics for burns. Cochrane Database of Systematic Reviews. 2017(7).
33. O'Carrigan B, Wong MHF, Willson ML, Stockler MR, Pavlakis N, Goodwin A. Bisphosphonates and other bone agents for breast cancer. Cochrane Database of Systematic Reviews. 2017(10).
34. Pasquali S, Hadjinicolaou AV, Chiarion Sileni V, Rossi CR, Mocellin S. Systemic treatments for metastatic cutaneous melanoma. Cochrane Database of Systematic Reviews. 2018(2).
35. Patil CG, Pricola K, Sarmiento JM, Garg SK, Bryant A, Black KL. Whole brain radiation therapy (WBRT) alone versus WBRT and radiosurgery for the treatment of brain metastases. Cochrane Database of Systematic Reviews. 2017(9).
36. Rai BP, Bondad J, Vasdev N, Adshead J, Lane T, Ahmed K, et al. Robotic versus open radical cystectomy for bladder cancer in adults. Cochrane Database of Systematic Reviews. 2019(4).
37. Rosenberg JE, Jung JH, Edgerton Z, Lee H, Lee S, Bakker CJ, et al. Retzius-sparing versus standard robotic-assisted laparoscopic prostatectomy for the treatment of clinically localized prostate cancer. Cochrane Database of Systematic Reviews. 2020(8).

38. Saeaib N, Peeyananjarassri K, Liabsuetrakul T, Buhachat R, Myriokefalitaki E. Hormone replacement therapy after surgery for epithelial ovarian cancer. Cochrane Database of Systematic Reviews. 2020(1).

39. Sathianathen NJ, Philippou YA, Kuntz GM, Konety BR, Gupta S, Lamb AD, et al. Taxane-based chemohormonal therapy for metastatic hormone-sensitive prostate cancer. Cochrane Database of Systematic Reviews. 2018(10).

40. Schmidt S, Kunath F, Coles B, Draeger DL, Krabbe LM, Dersch R, et al. Intravesical Bacillus Calmette-Guérin versus mitomycin C for Ta and T1 bladder cancer. Cochrane Database of Systematic Reviews. 2020(1).

41. Sim EHA, Yang IA, Wood-Baker R, Bowman RV, Fong KM. Gefitinib for advanced non-small cell lung cancer. Cochrane Database of Systematic Reviews. 2018(1).

42. Skoetz N, Will A, Monsef I, Brillant C, Engert A, von Tresckow B. Comparison of first-line chemotherapy including escalated BEACOPP versus chemotherapy including ABVD for people with early unfavourable or advanced stage Hodgkin lymphoma. Cochrane Database of Systematic Reviews. 2017(5).

43. Tosello G, Torloni MR, Mota BS, Neeman T, Riera R. Breast surgery for metastatic breast cancer. Cochrane Database of Systematic Reviews. 2018(3).

44. Tsao MN, Xu W, Wong RKS, Lloyd N, Laperriere N, Sahgal A, et al. Whole brain radiotherapy for the treatment of newly diagnosed multiple brain metastases. Cochrane Database of Systematic Reviews. 2018(1).

45. Vasconcellos VF, Marta GN, da Silva EMK, Gois AFT, de Castria TB, Riera R. Cisplatin versus carboplatin in combination with third-generation drugs for advanced non-small cell lung cancer. Cochrane Database of Systematic Reviews. 2020(1).

46. Vellayappan BA, Soon YY, Ku GY, Leong CN, Lu JJ, Tey JCS. Chemoradiotherapy versus chemoradiotherapy plus surgery for esophageal cancer. Cochrane Database of Systematic Reviews. 2017(8).

47. Vernooij RWM, Lancee M, Cleves A, Dahm P, Bangma CH, Aben KKH. Radical prostatectomy versus deferred treatment for localised prostate cancer. Cochrane Database of Systematic Reviews. 2020(6).

48. Vogel N, Schandelmaier S, Zumbrunn T, Ebrahim S, de Boer WEL, Busse JW, et al. Return-to-work coordination programmes for improving return to work in workers on sick leave. Cochrane Database of Systematic Reviews. 2017(3).

49. Zaheed M, Wilcken N, Willson ML, O'Connell DL, Goodwin A. Sequencing of anthracyclines and taxanes in neoadjuvant and adjuvant therapy for early breast cancer. Cochrane Database of Systematic Reviews. 2019(2).

50. Zhu J, Li R, Tiselius E, Roudi R, Teghararian O, Suo C, et al. Immunotherapy (excluding checkpoint inhibitors) for stage I to III non-small cell lung cancer treated with surgery or radiotherapy with curative intent. Cochrane Database of Systematic Reviews. 2017(12).

Non-Cochrane reviews

1. Al-Khatib SM, Fonarow GC, Joglar JA, Inoue LYT, Mark DB, Lee KL, et al. Primary Prevention Implantable Cardioverter Defibrillators in Patients With Nonischemic Cardiomyopathy: A Meta-analysis. JAMA Cardiol. 2017;2(6):685-8.

2. Alba AC, Foroutan F, Duero Posada J, Battioni L, Schofield T, Alhussein M, et al. Implantable cardiac defibrillator and mortality in non-ischaemic cardiomyopathy: an updated meta-analysis. Heart. 2018;104(3):230-6.

3. Alnimer Y, Hindi Z, Katato K. The Effect of Perioperative Bevacizumab on Disease-Free and Overall Survival in Locally Advanced HER-2 Negative Breast Cancer: A Meta-Analysis. Breast Cancer. 2018;12:1178223418792250.

4. Arnott C, Li Q, Kang A, Neuen BL, Bompoint S, Lam CSP, et al. Sodium-Glucose Cotransporter 2 Inhibition for the Prevention of Cardiovascular Events in Patients With Type 2 Diabetes Mellitus: A Systematic Review and Meta-Analysis. J Am Heart Assoc. 2020;9(3):e014908.
5. Balitsky AK, Karkar A, McCurdy A, Rochwerg B, Mian HS. Maintenance therapy in transplant ineligible adults with newly-diagnosed multiple myeloma: A systematic review and meta-analysis. Eur J Haematol. 2020;105(5):626-34.
6. Berger MD, Trelle S, Buchi AE, Jegerlehner S, Ionescu C, Lamy de la Chapelle T, et al. Impact on survival through consolidation radiotherapy for diffuse large B-cell lymphoma: a comprehensive meta-analysis. Haematologica. 2020;06:18.
7. Cai W, Yuan Y, Ge W, Fan Y, Liu X, Wu D, et al. EGFR Target Therapy Combined with Gemox for Advanced Biliary Tract Cancers: a Meta-analysis based on RCTs. Journal of Cancer. 2018;9(8):1476-85.
8. Caparica R, Bruzzone M, Hachem GE, Ceppi M, Lambertini M, Glasberg J, et al. Adjuvant chemotherapy in biliary tract cancer patients: A systematic review and meta-analysis of randomized controlled trials. Crit Rev Oncol Hematol. 2020;149:102940.
9. Caparica R, Bruzzone M, Poggio F, Ceppi M, de Azambuja E, Lambertini M. Anthracycline and taxane-based chemotherapy versus docetaxel and cyclophosphamide in the adjuvant treatment of HER2-negative breast cancer patients: a systematic review and meta-analysis of randomized controlled trials. Breast Cancer Res Treat. 2019;174(1):27-37.
10. Changal K, Masroor S, Elzanaty A, Patel M, Mir T, Khan S, et al. Meta-Analysis Comparing Multiple Arterial Grafts Versus Single Arterial Graft for Coronary-Artery Bypass Grafting. Am J Cardiol. 2020;130:46-55.
11. Chen S, Hu B, Li H. A meta-analysis of nivolumab for the treatment of advanced non-small-cell lung cancer. Onco Targets Ther. 2018;11:7691-7.
12. Ciccarese C, Iacovelli R, Bria E, Mosillo C, Bimbatti D, Fantinel E, et al. Second-line therapy for metastatic urothelial carcinoma: Defining the best treatment option among immunotherapy, chemotherapy, and antiangiogenic targeted therapies. A systematic review and meta-analysis. Semin Oncol. 2019;46(1):65-72.
13. Dong YW, Shi YQ, He LW, Cui XY, Su PZ. Effects of metformin on survival outcomes of pancreatic cancer: a meta-analysis. Oncotarget. 2017;8(33):55478-88.
14. Engel S, Awerbuch A, Kwon D, Picado O, Yechieli R, Yakoub D, et al. Optimal radiation dosing in concurrent neoadjuvant chemoradiation for resectable esophageal cancer: a meta-analysis. Journal of Gastrointestinal Oncology. 2020;10(3):391-9.
15. Giacoppo D, Colleran R, Cassese S, Frangieh AH, Wiebe J, Joner M, et al. Percutaneous Coronary Intervention vs Coronary Artery Bypass Grafting in Patients With Left Main Coronary Artery Stenosis: A Systematic Review and Meta-analysis. JAMA Cardiol. 2017;2(10):1079-88.
16. Giugliano D, Maiorino MI, Bellastella G, Longo M, Chiodini P, Esposito K. GLP-1 receptor agonists for prevention of cardiorenal outcomes in type 2 diabetes: An updated meta-analysis including the REWIND and PIONEER 6 trials. Diabetes Obes Metab. 2019;21(11):2576-80.
17. Haller PM, Sulzgruber P, Kaufmann C, Geelhoed B, Tamargo J, Wassmann S, et al. Bleeding and ischaemic outcomes in patients treated with dual or triple antithrombotic therapy: systematic review and meta-analysis. Eur Heart J Cardiovasc Pharmacother. 2019;5(4):226-36.
18. Han D, Wang G, Sun L, Ren X, Shang W, Xu L, et al. Comparison of irinotecan/platinum versus etoposide/platinum chemotherapy for extensive-stage small cell lung cancer: A meta-analysis. Eur J Cancer Care (Engl). 2017;26(6).
19. Han S, Hong Y, Liu T, Wu N, Ye Z. The efficacy and safety of paclitaxel and carboplatin with versus without bevacizumab in patients with non-small-cell lung cancer: a systematic review and meta-analysis. Oncotarget. 2018;9(18):14619-29.
20. Hao C, Tian J, Liu H, Li F, Niu H, Zhu B. Efficacy and safety of anti-PD-1 and anti-PD-1 combined with anti-CTLA-4 immunotherapy to advanced melanoma: A systematic review and meta-analysis of randomized controlled trials. Medicine (Baltimore). 2017;96(26):e7325.

21. Hu H, Zhu Q, Luo XS, Yang XW, Wang HD, Guo CY. Efficacy of PD-1/PD-L1 inhibitors against pretreated advanced cancer: a systematic review and meta-analysis. Oncotarget. 2018;9(14):11846-57.

22. Jang HJ, Kim HS, Kim JH, Lee J. The Effect of Statin Added to Systemic Anticancer Therapy: A Meta-Analysis of Randomized, Controlled Trials. Journal of Clin Med. 2018;7(10):04.

23. Kumar A, Reljic T, Hamadani M, Mohty M, Kharfan-Dabaja MA. Antithymocyte globulin for graft-versus-host disease prophylaxis: an updated systematic review and meta-analysis. Bone Marrow Transplant. 2019;54(7):1094-106.

24. Landre T, Des Guetz G, Chouahnia K, Duchemann B, Assie JB, Chouaid C. First-line angiogenesis inhibitor plus erlotinib versus erlotinib alone for advanced non-small-cell lung cancer harboring an EGFR mutation. J Cancer Res Clin Oncol. 2020;146(12):3333-9.

25. Leal F, Ferreira FP, Sasse AD. FOLFOXIRI Regimen for Metastatic Colorectal Cancer: A Systematic Review and Meta-Analysis. Clin Colorectal Cancer. 2017;16(4):405-9.e2.

26. Liu JW, Chen C, Loh EW, Chu CC, Wang MY, Ouyang HJ, et al. Tyrosine kinase inhibitors for advanced or metastatic thyroid cancer: a meta-analysis of randomized controlled trials. Curr Med Res Opin. 2018;34(5):795-803.

27. Ma H, Wu X, Tao M, Tang N, Li Y, Zhang X, et al. Efficacy and safety of bevacizumab-based maintenance therapy in metastatic colorectal cancer: A meta-analysis. Medicine (Baltimore). 2019;98(50):e18227.

28. Matuschek C, Bolke E, Haussmann J, Mohrmann S, Nestle-Kramling C, Gerber PA, et al. The benefit of adjuvant radiotherapy after breast conserving surgery in older patients with low risk breast cancer- a meta-analysis of randomized trials. Radiat. 2017;12(1):60.

29. Mauri D, Zarkavelis G, Filis P, Tsali L, Zafeiri G, Papadaki A, et al. Postoperative chemotherapy with single-agent fluoropyrimidines after resection of colorectal cancer liver metastases: a meta-analysis of randomised trials. ESMO open. 2018;3(4):e000343.

30. Miyashita M, Hattori M, Takano T, Toyama T, Iwata H. Risks and benefits of bevacizumab combined with chemotherapy for advanced or metastatic breast cancer: a meta-analysis of randomized controlled trials. Breast Cancer. 2020;27(3):347-54.

31. Montagnani F, Di Leonardo G, Pino M, Perboni S, Ribecco A, Fioretto L. Protracted Inhibition of Vascular Endothelial Growth Factor Signaling Improves Survival in Metastatic Colorectal Cancer: A Systematic Review. Journal of Translational Internal Medicine. 2017;5(1):18-26.

32. Natori A, Ethier JL, Amir E, Cescon DW. Capecitabine in early breast cancer: A meta-analysis of randomised controlled trials. Eur J Cancer. 2017;77:40-7.

33. Ottaiano A, Capozzi M, De Divitiis C, De Stefano A, Botti G, Avallone A, et al. Gemcitabine mono-therapy versus gemcitabine plus targeted therapy in advanced pancreatic cancer: a meta-analysis of randomized phase III trials. Acta Oncol. 2017;56(3):377-83.

34. Palmerini T, Serruys P, Kappetein AP, Genereux P, Riva DD, Reggiani LB, et al. Clinical outcomes with percutaneous coronary revascularization vs coronary artery bypass grafting surgery in patients with unprotected left main coronary artery disease: A meta-analysis of 6 randomized trials and 4,686 patients. Am Heart J. 2017;190:54-63.

35. Poggio F, Ceppi M, Lambertini M, Bruzzi P, Ugolini D, Bighin C, et al. Concurrent versus sequential adjuvant chemo-endocrine therapy in hormone-receptor positive early stage breast cancer patients: a systematic review and meta-analysis. Breast. 2017;33:104-8.

36. Pula A, Stawiski K, Braun M, Iskierka-Jazdzewska E, Robak T. Efficacy and safety of B-cell receptor signaling pathway inhibitors in relapsed/refractory chronic lymphocytic leukemia: a systematic review and meta-analysis of randomized clinical trials. Leuk Lymphoma. 2018;59(5):1084-94.

37. Ramos-Esquivel A, van der Laat A, Rojas-Vigott R, Juarez M, Corrales-Rodriguez L. Anti-PD-1/anti-PD-L1 immunotherapy versus docetaxel for previously treated advanced non-small cell

lung cancer: a systematic review and meta-analysis of randomised clinical trials. ESMO open. 2017;2(3):e000236.

38. Raphael J, Chan K, Karim S, Kerbel R, Lam H, Santos KD, et al. Antiangiogenic Therapy in Advanced Non-small-cell Lung Cancer: A Meta-analysis of Phase III Randomized Trials. Clin Lung Cancer. 2017;18(4):345-53.e5.

39. Reljic T, Kumar A, Klocksieben FA, Djulbegovic B. Treatment targeted at underlying disease versus palliative care in terminally ill patients: a systematic review. BMJ Open. 2017;7(1):e014661.

40. Roviello G, Corona SP, D'Angelo A, Rosellini P, Nobili S, Mini E. Immune Checkpoint Inhibitors in Pre-Treated Gastric Cancer Patients: Results from a Literature-Based Meta-Analysis. Int J Molec Sc. 2020;21(2):10.

41. Schmitt AM, Herbrand AK, Fox CP, Bakunina K, Bromberg JEC, Cwynarski K, et al. Rituximab in primary central nervous system lymphoma-A systematic review and meta-analysis. Hematol Oncol. 2019;16:16.

42. Shui L, Wu YS, Lin H, Shui P, Sun Q, Chen X. Triplet Chemotherapy (FOLFOXIRI) Plus Bevacizumab Versus Doublet Chemotherapy (FOLFOX/FOLFIRI) Plus Bevacizumab in Conversion Therapy for Metastatic Colorectal Cancer: a Meta-Analysis. Cell Physiol Biochem. 2018;48(5):1870-81.

43. Stavrakis S, Asad Z, Reynolds D. Implantable Cardioverter Defibrillators for Primary Prevention of Mortality in Patients With Nonischemic Cardiomyopathy: A Meta-Analysis of Randomized Controlled Trials. J Cardiovasc Electrophysiol. 2017;28(6):659-65.

44. Tang X, He J, Li B, Zheng Y, Li K, Zou S, et al. Efficacy and Safety of Gefitinib in Patients with Advanced Head and Neck Squamous Cell Carcinoma: A Meta-Analysis of Randomized Controlled Trials. J Oncol. 2019;2019:6273438.

45. Tringali A, Hassan C, Rota M, Rossi M, Mutignani M, Aabakken L. Covered vs. uncovered self-expandable metal stents for malignant distal biliary strictures: a systematic review and meta-analysis. Endoscopy. 2018;50(6):631-41.

46. Vidal L, Gurion R, Shargian L, Dreyling M, Gafter-Gvili A. Bendamustine for patients with indolent B cell lymphoproliferative malignancies including chronic lymphocytic leukaemia - an updated meta-analysis. Br J Haematol. 2019;186(2):234-42.

47. Wang J, Xu B, Wang W, Zhai X, Chen X. Efficacy and safety of fulvestrant in postmenopausal patients with hormone receptor-positive advanced breast cancer: a systematic literature review and meta-analysis. Breast Cancer Res Treat. 2018;171(3):535-44.

48. Wang X, Bao Z, Zhang X, Li F, Lai T, Cao C, et al. Effectiveness and safety of PD-1/PD-L1 inhibitors in the treatment of solid tumors: a systematic review and meta-analysis. Oncotarget. 2017;8(35):59901-14.

49. Xu L, Yan N, Li Z, Luo L, Wu X, Liu Q, et al. A comparison of fulvestrant plus a targeted agent with fulvestrant alone in hormone receptor-positive advanced breast cancer that progressed on prior endocrine therapy: a meta-analysis. Onco Targets Ther. 2018;11:8389-98.

50. Zhong S, Qie S, Yang L, Yan Q, Ge L, Wang Z. S-1 monotherapy versus S-1 combination therapy in gemcitabine-refractory advanced pancreatic cancer: A meta-analysis (PRISMA) of randomized control trials. Medicine (Baltimore). 2017;96(30):e7611.

Appendix-Table A6: Extended characteristics of reviews

Category	Item	Review			Review time-to-event outcomes		
		Overall (N = 100)	Cochrane (n = 50)	Non-Cochrane (n = 50)	Overall (N = 217)	Cochrane (n = 93)	Non-Cochrane (n = 124)
Publication							
Publication year	2017	36% (36)	36% (18)	36% (18)			
	2018	28% (28)	28% (14)	28% (14)			
	2019	18% (18)	18% (9)	18% (9)			
	2020	18% (18)	18% (9)	18% (9)			
Journal Impact Factor	Median (IQR)		11.87	4.41 (3.33 - 6.18)			
	Mean (Range)		11.87	6.37 (1.817 - 35.86)			
Review update	Yes	27% (27)	50% (25)	4% (2)			
Multiple review comparisons	Yes	28% (28)	44% (22)	12% (6)			
Population							
Medical field	Neoplasms	82% (82)	86% (43)	78% (39)			
	Diseases of the circulatory system	11% (11)	4% (2)	18% (9)			
	Diseases of the skin and subcutaneous tissue	3% (3)	6% (3)	0% (0)			
	Diseases of blood, -forming organs and immune mechanism	2% (2)	2% (1)	2% (1)			
	Other	2% (2)	2% (1)	2% (1)			
Medical condition	Breast cancer	13% (13)	10% (5)	16% (8)			
	Colorectal cancer	9% (9)	8% (4)	10% (5)			
	Non-small cell lung cancer	8% (8)	6% (3)	10% (5)			
	Prostate cancer	6% (6)	12% (6)	0% (0)			
	Non-ischemic cardiomyopathy	4% (4)	2% (1)	6% (3)			
	Other	3% (3)	60% (30)	58% (29)			
Clinical stage	Advanced/ Second or third line	36% (36)	30% (15)	42% (21)			
	Early/ First line	30% (30)	38% (19)	22% (11)			
	No restriction	12% (12)	16% (8)	8% (4)			
	Not reported	13% (13)	10% (5)	16% (8)			
	Not applicable	9% (9)	6% (3)	12% (6)			
Age group	Adults	96% (96)	98% (49)	94% (47)			
	Both	1% (1)	2% (1)	0% (0)			
	Not reported	3% (3)	0% (0)	6% (3)			
Interventions							
Experimental intervention	Biologics/ drug	58% (58)	40% (20)	76% (38)			
	Surgical procedure	10% (10)	14% (7)	6% (3)			
	Radiation	5% (5)	10% (5)	0% (0)			
	Biologics/ drug, Surgical procedure	4% (4)	6% (3)	2% (1)			
	Medical devices	4% (4)	0% (0)	8% (4)			
	Other	3% (3)	30% (15)	8% (4)			
Comparator intervention	Biologics/ drugs	42% (42)	28% (14)	56% (28)			
	Surgical procedure	9% (9)	8% (4)	10% (5)			
	Best supportive/ Optimal medical care	7% (7)	8% (4)	6% (3)			
	Observation	5% (5)	2% (1)	8% (4)			
	Placebo, No treatment	5% (5)	10% (5)	0% (0)			

Category	Item	Review			Review time-to-event outcomes		
		Overall (N = 100)	**Cochrane (n = 50)**	**Non-Cochrane (n = 50)**	**Overall (N = 217)**	**Cochrane (n = 93)**	**Non-Cochrane (n = 124)**
	Radiation	5% (5)	10% (5)	0% (0)			
	Other	28% (28)	36% (18)	20% (10)			
Comparisons	Biologics/ drugs	37% (37)	24% (12)	50% (25)			
	Surgical procedures	7% (7)	8% (4)	6% (3)			
	Biologics/ drug vs. Best supportive/ Optimal medical care	4% (4)	4% (2)	4% (2)			
	Biologics/ drug vs. Observation	4% (4)	0% (0)	8% (4)			
	Biologics/ drug vs. Biologics/ Drugs, Placebo	3% (3)	2% (1)	4% (2)			
	Biologics/ drug vs. Placebo	3% (3)	0% (0)	6% (3)			
	Biologics/ drug vs. Placebo, No treatment	3% (3)	6% (3)	0% (0)			
	Radiation	3% (3)	6% (3)	0% (0)			
	Biologics/ drug (schedule alteration)	2% (2)	2% (1)	2% (1)			
	Follow-up strategies	2% (2)	4% (2)	0% (0)			
	Other	32% (32)	44% (22)	20% (10)			
Comparator treatment considered?	Yes	2% (2)	2% (1)	2% (1)			
Outcomes - Planned							
Planned outcome number	Median (IQR)	5 (4 - 8)	7 (5 - 8)	4 (3 - 5)			
	Mean (Range)	5.79 (1 - 15)	6.82 (3 - 12)	4.67 (1 - 15)			
Planned TTE outcome number	Median	2 (2 - 2)	2 (2 - 3)	2 (2 - 2)			
	Mean	2.39 (1 - 12)	2.17 (1 - 4)	2.62 (1 - 12)			
Planned TTE outcomes	Overall survival, ACM or death from any cause	89% (89)	88% (44)	90% (45)			
	Progression-free survival	44% (44)	36% (18)	52% (26)			
	Disease-free survival	13% (13)	16% (8)	10% (5)			
	Myocardial infarction	5% (5)	0% (0)	10% (5)			
	Stroke	5% (5)	0% (0)	10% (5)			
	Cardiovascular mortality	4% (4)	2% (1)	6% (3)			
	Time to death from prostate cancer	4% (4)	8% (4)	0% (0)			
	Cardiac death	4% (4)	2% (1)	6% (3)			
	Other	3% (3)	62% (31)	58% (29)			
	Unclear	2% (2)	2% (1)	2% (1)			
	Not reported	1% (1)	0% (0)	2% (1)			
	Not applicable	2% (2)	2% (1)	2% (1)			
Number of outcomes analyzed	Median	5 (3 - 6)	5 (4 - 6.75)	4 (2.25 - 5)			
	Mean	4.83 (1 - 12)	5.34 (1 - 12)	4.32 (1 - 12)			
Number of TTE outcomes analyzed	Median	2 (1 - 2)	2 (1 - 2)	2 (2 - 2)			
	Mean	2.23 (1 - 12)	1.92 (1 - 4)	2.54 (1 - 12)			
TTE outcomes in methods different from analyzed	Yes	11% (11)	20% (10)	2% (1)			
Reasons for difference in mentioned and analyzed TTE outcomes	Outcome not assessed in trial(s)	7% (7)	14% (7)	0% (0)			
	Time-to-event data not available in trial(s)	2% (2)	2% (1)	0% (0)			
	Not pooled due to clinical heterogeneity	1% (1)	2% (1)	0% (0)			
	Not reported	1% (1)	0% (0)	2% (1)			
	Not applicable	89% (89)	80% (40)	98% (49)			
TTE outcomes analyzed	Overall survival, ACM or death from any cause	89% (89)	84% (42)	94% (47)	41% (89)	45% (42)	38% (47)
	Progression-free survival	39% (39)	26% (13)	52% (26)	18% (39)	14% (13)	21% (26)

Category	Item	Review			Review time-to-event outcomes		
		Overall (N = 100)	**Cochrane (n = 50)**	**Non-Cochrane (n = 50)**	**Overall (N = 217)**	**Cochrane (n = 93)**	**Non-Cochrane (n = 124)**
	Disease-free survival	10% (10)	10% (5)	10% (5)	5% (10)	5% (5)	4% (5)
	Myocardial infarction	6% (6)	0% (0)	12% (6)	3% (6)	0% (0)	5% (6)
	Stroke	5% (5)	0% (0)	10% (5)	2% (5)	0% (0)	4% (5)
	Cardiac death	4% (4)	2% (1)	6% (3)	2% (4)	1% (1)	2% (3)
	Cardiovascular mortality	3% (3)	2% (1)	4% (2)	1% (3)	1% (1)	2% (2)
	Major adverse cardiac events (MACE)	3% (3)	0% (0)	6% (3)	1% (3)	0% (0)	2% (3)
	Other	3% (3)	62% (31)	54% (27)	1% (3)	33% (31)	22% (27)
Sample size							
Number of included studies in reviews and meta-analyses	Median	5 (4 - 8)	5 (3 - 10)	6 (4 - 8)	4 (3 - 7)	4 (2 - 6)	5 (4 - 7)
	Mean	6.69 (2 - 24)	7.12 (2 - 24)	6.26 (2 - 19)	5.25 (2 - 19)	5.05 (2 - 19)	5.40 (2 - 19)
Total population in review or meta-analysis	Median	1722 (978 - 4390)	1415 (572 - 4022)	1866 (1395 - 4526)	811 (308 - 2876)	711 (177 - 2327)	1042 (698 - 3173)
	Mean	3877 (170 - 56004)	2795 (170 - 13216)	4911 (343 - 56004)	5745 (181 - 38723)	2656 (181 - 13949)	8985 (482 - 38723)
	Not reported	12% (12)	14% (7)	10% (5)	23% (49)	8% (7)	34% (42)
Experimental population in review or meta-analysis	Median	765 (373 - 1539)	223 (159 - 641)	937 (645 - 1804)	791 (290 - 2197)	715 (181 - 1772)	981 (645 - 3076)
	Mean	1235 (90 - 5039)	402 (90 - 951)	1529 (238 - 5039)	1840 (81 - 12373)	1722.64 (81 - 12373)	1963 (238 - 5039)
	Not reported	77% (77)	88% (44)	66% (33)	62% (135)	55% (51)	68% (84)
Control population in review or meta-analysis	Median	765 (373 - 1539)	223 (159 - 641)	937 (645 - 1804)	2025 (879 - 4822)	1407 (446 - 3657)	4394 (1580 - 9302)
	Mean	1128 (80 - 4278)	432 (80 - 1145)	1374 (244 - 4278)	1575 (80 - 6403)	1377 (80 - 6403)	1783 (244 - 4278)
	Not reported	77% (77)	88% (44)	66% (33)	62% (135)	55% (51)	68% (84)

Abbreviations: ACM = all-cause mortality, TTE = time to event

Appendix-Table A7: Extended characteristics of included time-to-event review outcomes

Category	Item	Review			Review time-to-event outcomes					
					Overall			All-cause mortality/ Overall survival (n = 89)	Combined, including all-cause mortality (n = 63)	Not including all-cause mortality (n = 65)
		Overall (N = 100)	Cochrane (n = 50)	Non-Cochrane (n = 50)	Overall (N = 217)	Cochrane (n = 93)	Non-Cochrane (n = 124)			
TTE outcomes as primary outcomes	Yes	69% (69)	92% (46)	46% (23)	39% (85)	62% (58)	22% (27)	55% (49)	29% (18)	28% (18)
	No	3% (3)	6% (3)	0% (0)	33% (72)	37% (34)	31% (38)	16% (14)	41% (26)	49% (32)
	Not applicable	28% (28)	2% (1)	54% (27)	28% (60)	1% (1)	48% (59)	29% (26)	30% (19)	23% (15)
TTE outcomes that were primary outcomes	Overall survival, ACM or death from any cause	71% (49)	76% (35)	61% (14)	58% (49)	60% (35)	52% (14)	100% (49)	0% (0)	0% (0)
	Progression-free survival	20% (14)	13% (6)	35% (8)	16% (14)	10% (6)	30% (8)	0% (0)	71% (12)	11% (2)
	Disease-free survival	6% (4)	7% (3)	4% (1)	5% (4)	5% (3)	4% (1)	0% (0)	24% (4)	0% (0)
	Time to complete healing	3% (2)	4% (2)	0% (0)	2% (2)	3% (2)	0% (0)	0% (0)	0% (0)	11% (2)
	All-cause death, myocardial infarction or stroke	1% (1)	0% (0)	4% (1)	1% (1)	0% (0)	4% (1)	0% (0)	6% (1)	0% (0)
	Any thromboembolic event	1% (1)	2% (1)	0% (0)	1% (1)	2% (1)	0% (0)	0% (0)	0% (0)	5% (1)
	Cardiac death	1% (1)	2% (1)	0% (0)	1% (1)	2% (1)	0% (0)	0% (0)	6% (1)	0% (0)
	Cardiovascular mortality	1% (1)	2% (1)	0% (0)	1% (1)	2% (1)	0% (0)	0% (0)	0% (0)	5% (1)
	Other	1% (1)	20% (9)	13% (3)	1% (1)	16% (9)	11% (3)	0% (0)	0% (0)	5% (1)
Reviews providing definitions of TTE outcomes	For all outcomes	48% (48)	74% (37)	22% (11)						
	For ≥1 outcome	7% (7)	6% (3)	8% (4)						
	For no outcome	45% (45)	20% (10)	70% (35)						
TTE outcome definition per outcome	Yes				48% (104)	83% (77)	22% (27)	51% (45)	56% (35)	37% (24)
	No				52% (113)	17% (16)	78% (97)	49% (44)	44% (28)	63% (41)
Composite TTE outcomes	Yes	39% (39)	42% (21)	36% (18)	22% (47)	24% (22)	20% (25)	0% (0)	56% (35)	18% (12)
	No	97% (97)	98% (49)	96% (48)	65% (140)	69% (64)	61% (76)	100% (89)	0% (0)	78% (51)
	Unclear/ Not reported	29% (29)	12% (6)	46% (23)	14% (30)	8% (7)	19% (23)	0% (0)	44% (28)	3% (2)
Composite events described	Yes	87% (34)	90% (19)	83% (15)	87% (41)	91% (20)	84% (21)	0% (NA)	91% (32)	75% (9)
All-cause mortality part of outcome definition	Yes	90% (90)	86% (43)	94% (47)	57% (124)	65% (60)	52% (64)	100% (89)	56% (35)	0% (0)
	No	29% (29)	40% (20)	18% (9)	29% (62)	28% (26)	29% (36)	0% (0)	0% (0)	95% (62)
	Unclear	31% (31)	14% (7)	48% (24)	14% (31)	8% (7)	19% (24)	0% (0)	44% (28)	5% (3)
Death as competing event possible	Yes	29% (29)	40% (20)	18% (9)	29% (64)	28% (26)	31% (38)	1% (1)	2% (1)	98% (64)
	No	90% (90)	86% (43)	94% (47)	56% (121)	62% (58)	51% (63)	99% (88)	52% (33)	186% (121)
	Unclear	31% (31)	16% (8)	46% (23)	15% (32)	10% (9)	19% (23)	0% (0)	46% (29)	49% (32)
Reviews reporting outcomes as events of absence of events	Absence of event only	61% (61)	48% (24)	74% (37)						
	Event only	24% (24)	26% (13)	22% (11)						
	Both (with reasoning)	5% (5)	10% (5)	0% (0)						
	Mixed	2% (2)	4% (2)	0% (0)						
	At least one unclear (without reasoning)	8% (8)	12% (6)	4% (2)						
Outcome reporting as events or absence of event	Absence of event				54% (118)	53% (49)	56% (69)	71% (63)	81% (51)	6% (4)
	Event				37% (81)	30% (28)	43% (53)	19% (17)	13% (8)	86% (56)
	Both (with reasoning)				5% (10)	11% (10)	0% (0)	7% (6)	2% (1)	5% (3)
	Unclear (both without reasoning)				4% (8)	6% (6)	2% (2)	3% (3)	5% (3)	3% (2)
Reviews including follow-up start in outcome definitions	Randomization	32% (32)	48% (24)	16% (8)						
	Allocated treatment	2% (2)	2% (1)	2% (1)						
	Enrollment	2% (2)	2% (1)	0% (0)						
	Mixed for different outcomes	2% (2)	2% (1)	0% (0)						

Category	Item	Review			Review time-to-event outcomes					
					Overall			All-cause mortality/ Overall survival (n = 89)	Combined, including all-cause mortality (n = 63)	Not including all-cause mortality (n = 65)
		Overall (N = 100)	Cochrane (n = 50)	Non-Cochrane (n = 50)	Overall (N = 217)	Cochrane (n = 93)	Non-Cochrane (n = 124)			
	At least one not applicable (e.g., start not reported)	62% (62)	46% (23)	82% (41)						
Follow-up start included in outcome definition	Randomization				32% (70)	57% (53)	14% (17)	37% (33)	33% (21)	25% (16)
	Enrollment				3% (6)	6% (6)	0% (0)	4% (4)	3% (2)	0% (0)
	Allocated treatment				2% (4)	2% (2)	2% (2)	1% (1)	2% (1)	3% (2)
	Multiple time points (e.g., "enrollment or treatment")				1% (2)	2% (2)	0% (0)	1% (1)	2% (1)	0% (0)
	Not applicable (e.g., start of follow-up not reported)				62% (135)	32% (30)	85% (105)	56% (50)	60% (38)	72% (47)

Abbreviations: ACM = all-cause mortality, TTE = time to event

Appendix-Table A8: Analysis principles and adjustment status of trial analyses included in time-to-event meta-analyses of included reviews

Category	Item	Review			Review time-to-event outcomes		
		Overall (N = 100)	Cochrane (n = 50)	Non-Cochrane (n = 50)	Overall (N = 217)	Cochrane (n = 93)	Non-Cochrane (n = 124)
Types of analyses eligible in reviews	ITT	42% (42)	72% (36)	12% (6)			
	Not reported	58% (58)	28% (14)	88% (44)			
Types of analyses eligible for outcome analyses	ITT	1% (1)	0% ()	2% (1)	1% (2)	0% (0)	2% (2)
	Not reported	99% (99)	100% (50)	98% (49)	99% (215)	100% (93)	98% (122)
Types of analyses included in reviews	ITT	21% (21)	24% (12)	18% (9)			
	Included trial(s) did not report type of analysis	6% (6)	10% (5)	2% (1)			
	mITT	4% (4)	8% (4)	0% (0)			
	PP	2% (2)	4% (2)	0% (0)			
	Other	2% (2)	4% (2)	0% (0)			
	Not reported for all trials	16% (16)	28% (14)	4% (2)			
	Not reported for any trial	63% (63)	48% (24)	78% (39)			
Types of analyses included in outcome analyses	ITT	2% (2)	2% (1)	2% (1)	2% (5)	2% (2)	2% (3)
	Not reported for all trials	1% (1)	2% (1)	0% (0)	0% (1)	1% (1)	0% (0)
	Not reported	97% (97)	96% (48)	98% (49)	97% (211)	97% (90)	98% (121)
Unadjusted/ adjusted HRs eligible in reviews	Both	4% (4)	8% (4)	0% (0)			
	Hierarchical (adjusted before unadjusted)	4% (4)	6% (3)	2% (1)			
	Unadjusted only	2% (2)	2% (1)	2% (1)			
	Adjusted only	1% (1)	2% (1)	0% (0)			
	Hierarchical (unadjusted before adjusted)	1% (1)	2% (1)	0% (0)			
	Unclear	6% (6)	10% (5)	2% (1)			
	Not reported	82% (82)	90% (45)	94% (47)			
Dealing with unadjusted/ adjusted HRs	Included in interpretation of heterogeneity	2% (2)	4% (2)	0% ()			
	Combined in meta-analysis	1% (1)	0% ()	2% (1)			
	Unclear	7% (7)	12% (6)	2% (1)			
	Not reported	8% (8)	14% (7)	2% (1)			
	Not applicable (unadjusted/ adjusted not mentioned)	82% (82)	70% (35)	94% (47)			
Stratified HRs eligible in reviews	Yes	1% (1)	2% (1)	0% (0)			
	Unclear	1% (1)	2% (1)	0% (0)			
	No	98% (98)	96% (48)	100% (50)			
Unadjusted/ adjusted HRs eligible in outcome analyses	Unadjusted only	2% (2)	4% (2)	0% (0)	1% (2)	2% (2)	0% (0)
	Adjusted only	1% (1)	2% (1)	0% (0)	0% (1)	1% (1)	0% (0)
	Not reported	97% (97)	94% (47)	100% (50)	99% (214)	97% (90)	100% (124)
Dealing with unadjusted/ adjusted HRs	Only unadjusted/adjusted HRs included in analysis	2% (2)	4% (2)	0% (0)	1% (2)	2% (2)	0% (0)
	Unadjusted HRs recalculated	1% (1)	2% (1)	0% (0)	0% (1)	1% (1)	0% (0)
	Not applicable	97% (97)	94% (47)	100% (50)	99% (214)	97% (90)	100% (124)
Unadjusted/ adjusted HRs discussed in reviews	In discussions	2% (2)	2% (1)	2% (1)			
	In results	1% (1)	2% (1)	0% (0)			
	Not applicable	1% (1)	0% (0)	2% (1)			
	Not reported	96% (96)	96% (48)	96% (48)			

Category	**Item**	**Review**			**Review time-to-event outcomes**		
		Overall (N = 100)	**Cochrane (n = 50)**	**Non-Cochrane (n = 50)**	**Overall (N = 217)**	**Cochrane (n = 93)**	**Non-Cochrane (n = 124)**
Unadjusted/ adjusted in discussion for individual outcome	No	100% (100)	100% (50)	100% (50)	100% (217)	100% (93)	100% (124)
Abbreviations: HR = hazard ratio, ITT = intention to treat, mITT = modified intention to treat							

Appendix-Table A9: Extended information on time-to-event specific methods and time-to-event data acquirement in included reviews

Category	Item	Review			Review time-to-event outcomes					
					Overall			All-cause mortality/ Overall survival (n = 89)	Combined, including all-cause mortality (n = 63)	Not including all-cause mortality (n = 65)
		Overall (N = 100)	Cochrane (n = 50)	Non-Cochrane (n = 50)	Overall (N = 217)	Cochrane (n = 93)	Non-Cochrane (n = 124)			
HR type eligible in reviews	HR/ log(HR) not further specified	91% (91)	94% (47)	88% (44)						
	Cox model HR/ log(HR)	2% (2)	4% (2)	0% (0)						
	Cox model HR, log-rank test and KM-Curve	1% (1)	2% (1)	0% (0)						
	Not reported	6% (6)	0% (0)	12% (6)						
HR types eligible per outcome	HR/ log(HR) from Cox model	1% (1)	2% (1)	0% (0)	0% (1)	1% (1)	0% (0)	1% (1)	0% (0)	0% (0)
	HR/ log(HR) from median survival times and CI	1% (1)	2% (1)	0% (0)	0% (1)	1% (1)	0% (0)	1% (1)	0% (0)	0% (0)
	Unclear	1% (1)	2% (1)	0% (0)	1% (3)	3% (3)	0% (0)	1% (1)	0% (0)	3% (2)
	Not reported	97% (97)	94% (47)	100% (50)	98% (212)	95% (88)	100% (124)	97% (86)	100% (63)	97% (63)
Methods to obtain TTE data per review	HR and confidence intervals	64% (64)	66% (33)	62% (31)						
	Specified set of methods (e.g., Tierney 2008)	46% (46)	76% (38)	16% (8)						
	log(HR) and standard error	16% (16)	26% (13)	6% (3)						
	Survival curves	13% (13)	14% (7)	12% (6)						
	HR with other information (e.g., events)	10% (10)	16% (8)	4% (2)						
	P-value with other information (e.g., events)	8% (8)	8% (4)	8% (4)						
	IPD (recalculated or from publication)	4% (4)	4% (2)	4% (2)						
	Median survival times	4% (4)	8% (4)	0% (0)						
	Other (Formular by Parmar, not TTE specific)	1% (1)	0% (0)	2% (1)						
	Time point specific survival times	1% (1)	2% (1)	0% (0)						
	Risk ratio	1% (1)	0% (0)	2% (1)						
	Unclear	6% (6)	8% (4)	4% (2)						
	Not reported	16% (16)	0% (0)	32% (16)						
Recalculation of TTE data reported for an outcome	Yes	18% (18)	34% (17)	2% (1)						
Methods to obtain TTE data for an outcome	HR and confidence intervals	9% (9)	18% (9)	0% (0)	7% (16)	17% (16)	0% (0)	9% (8)	5% (3)	8% (5)
	P-value with other information (e.g., events)	4% (4)	8% (4)	0% (0)	5% (10)	11% (10)	0% (0)	4% (4)	5% (3)	5% (3)
	Survival curves	7% (7)	12% (6)	2% (1)	5% (10)	10% (9)	1% (1)	6% (5)	2% (1)	6% (4)
	HR with other information (e.g., events)	1% (1)	2% (1)	0% (0)	1% (3)	3% (3)	0% (0)	1% (1)	2% (1)	2% (1)
	RevMan calculator	1% (1)	2% (1)	0% (0)	1% (3)	3% (3)	0% (0)	1% (1)	2% (1)	2% (1)
	Time point specific survival times	1% (1)	2% (1)	0% (0)	1% (2)	2% (2)	0% (0)	1% (1)	0% (0)	2% (1)
	Trial exclusion due to inability to recalculate data	1% (1)	2% (1)	0% (0)	0% (1)	1% (1)	0% (0)	0% (0)	2% (1)	0% (0)
	IPD (recalculated or from publication)	1% (1)	2% (1)	0% (0)	0% (1)	1% (1)	0% (0)	1% (1)	0% (0)	0% (0)
	Median survival times	1% (1)	2% (1)	0% (0)	0% (1)	1% (1)	0% (0)	1% (1)	0% (0)	0% (0)
	Unclear	5% (5)	10% (5)	0% (0)	3% (6)	6% (6)	0% (0)	4% (4)	0% (0)	3% (2)

Abbreviations: CI = confidence intervals, HR = hazard ratio; KM = Kaplan-Meier, TTE = time-to-event

Appendix-Table A10: Handling of specific trial characteristics with relevance for time-to-event outcomes in the included reviews

Category	Item	Review			Review time-to-event outcomes		
		Overall (N = 100)	**Cochrane (n = 50)**	**Non-Cochrane (n = 50)**	**Overall (N = 217)**	**Cochrane (n = 93)**	**Non-Cochrane (n = 124)**
Heterogeneous outcome definitions of included trials							
Reviews mentioning heterogeneous TTE outcome definitions	In discussion	3% (3)	0% (0)	6% (3)			
	In results	3% (3)	6% (3)	0% (0)			
	In results and discussion	1% (1)	2% (1)	0% (0)			
	Not reported	93% (93)	92% (46)	94% (47)			
Heterogeneous outcome definitions discussed	Yes	4% (4)	4% (2)	4% (2)	2% (4)	2% (2)	2% (2)
Follow-up							
Reviews reporting a planned follow-up duration	Minimum duration of follow-up required	5% (5)	6% (3)	4% (2)			
	Longest follow-up	4% (4)	4% (2)	4% (2)			
	Maximum duration of follow-up specified	2% (2)	4% (2)	0% (0)			
	Time-specific (12 months, 2 year, 10 year, ...)	1% (1)	2% (1)	0% (0)			
	Not reported	88% (88)	84% (42)	92% (46)			
Follow-up time specification for TTE outcomes	Longest follow-up	8% (8)	8% (4)	8% (4)	13% (29)	9% (8)	17% (21)
	Minimum duration of follow-up required	2% (2)	4% (2)	0% (0)	3% (6)	6% (6)	0% (0)
	Maximum duration of follow-up specified	3% (3)	4% (2)	2% (1)	1% (3)	2% (2)	1% (1)
	"Time-specific (12 months, 2 year, 10 year, ...)"	2% (2)	4% (2)	0% (0)	1% (2)	2% (2)	0% (0)
	Not reported	85% (85)	80% (40)	90% (45)	82% (177)	81% (75)	82% (102)
Analyses - Varying follow-up between included trials							
Dealing with varying follow-up in reviews	Sensitivity analyses (e.g., shorter/longer follow-up)	8% (8)	14% (7)	2% (1)			
	Included in interpretation of heterogeneity	2% (2)	4% (2)	0% (0)			
	Included in meta-regression	2% (2)	0% (0)	4% (2)			
	Excluded studies with divergent follow-up time	1% (1)	0% (0)	2% (1)			
	Follow-up time for comparisons pre-determined	1% (1)	2% (1)	0% (0)			
	Mentioned as RoB criterion in methods	1% (1)	2% (1)	0% (0)			
	Pre-defined timing as inclusion criterion	1% (1)	0% (0)	2% (1)			
	Unclear	1% (1)	0% (0)	2% (1)			
	Not reported	83% (83)	78% (39)	88% (44)			
Dealing with varying follow-up for individual outcomes	Included in meta-regression	2% (2)	2% (1)	2% (1)	1% (2)	1% (1)	1% (1)
	Results reported for multiple time-points	1% (1)	2% (1)	0% (0)	0% (1)	1% (1)	0% (0)
	Not reported per outcome	97% (97)	96% (48)	98% (49)	99% (214)	98% (91)	99% (123)
Varying follow-up discussed	In discussions	10% (10)	10% (5)	10% (5)			
	In results	7% (7)	10% (5)	4% (2)			
	In results and discussion	5% (5)	8% (4)	2% (1)			
	Not reported	77% (77)	70% (35)	84% (42)			
	Not applicable	1% (1)	2% (1)	0% (0)			
Varying follow-up in discussion for individual outcomes	Yes	5% (5)	4% (2)	6% (3)	3% (7)	4% (4)	2% (3)
	No	94% (94)	94% (47)	94% (47)	96% (209)	95% (88)	98% (121)

Category	Item	Review			Review time-to-event outcomes		
		Overall (N = 100)	Cochrane (n = 50)	Non-Cochrane (n = 50)	Overall (N = 217)	Cochrane (n = 93)	Non-Cochrane (n = 124)
	Not applicable	1% (1)	2% (1)	0% (0)	0% (1)	1% (1)	0% (0)
Analyses - Missing outcome data							
Dealing with missing outcome data in reviews	Mentioned as RoB criterion in methods	71% (71)	94% (47)	48% (24)			
	Contact with authors	8% (8)	14% (7)	2% (1)			
	Sensitivity analyses (according to rate of MOD)	7% (7)	14% (7)	0% (0)			
	Single imputation	4% (4)	8% (4)	0% (0)			
	Not reported	26% (26)	0% (0)	52% (26)			
Dealing with missing data in individual outcomes	Single imputation	1% (1)	2% (1)	0% (0)	0% (1)	1% (1)	0% (0)
	Not reported per outcome	99% (99)	98% (49)	100% (50)	100% (216)	99% (92)	100% (124)
Missing outcome data discussed	In results	59% (59)	78% (39)	40% (20)			
	In results and discussion	6% (6)	12% (6)	0% (0)			
	Not reported	35% (35)	10% (5)	60% (30)			
Missing outcome data included in discussion for individual outcomes	Yes	2% (2)	4% (2)	0% (0)	1% (2)	2% (2)	0% (0)
	No	98% (98)	96% (48)	100% (50)	99% (215)	98% (91)	100% (124)
Analyses - Informative censoring							
Dealing with informative censoring in reviews	Mentioned as RoB criterion in methods	3% (3)	6% (3)	0% (0)			
	Not reported	97% (97)	94% (47)	100% (50)			
Dealing with informative censoring in individual outcomes	Not reported per outcome	100% (100)	100% (50)	100% (50)	0% (217)	100% (93)	100% (124)
Informative censoring discussed	In discussion	1% (1)	2% (1)	0% (0)			
	In results	1% (1)	2% (1)	0% (0)			
	Not reported	98% (98)	96% (48)	100% (50)			
Informative censoring in discussion for individual outcomes	No	100% (100)	100% (50)	100% (50)	0% (217)	100% (93)	100% (124)
Analyses - Competing events							
Dealing with (deaths as) *competing event in reviews*	Mentioned as RoB criterion in methods	1% (1)	0% (0)	2% (1)			
	Not reported	59% (59)	58% (29)	60% (30)			
	Not applicable	40% (40)	42% (21)	38% (19)			
Dealing with (deaths as) competing events in individual outcomes	Not reported per outcome	60% (60)	58% (29)	62% (31)	48% (105)	46% (43)	50% (62)
	Not applicable	40% (40)	42% (21)	38% (19)	52% (112)	54% (50)	50% (62)
(Deaths as) Competing events discussed	In results and in discussion	1% (1)	0% ()	2% (1)			
	Not reported	59% (59)	58% (29)	60% (30)			
	Not applicable	40% (40)	42% (21)	38% (19)			
(Deaths as) Competing events included in discussion for individual outcomes	Yes	1% (1)	2% (1)	0% (0)	0% (1)	1% (1)	0% (0)
	No	59% (59)	56% (28)	62% (31)	48% (104)	45% (42)	50% (62)
	Not applicable	40% (40)	42% (21)	38% (19)	52% (112)	54% (50)	50% (62)
Analyses - Treatment switching							
Dealing with treatment switching in reviews	Mentioned as RoB criterion in methods	4% (4)	4% (2)	4% (2)			
	Complies with review PICO	2% (2)	4% (2)	0% (0)			
	Presence reported for each trial	2% (2)	4% (2)	0% (0)			
	Sensitivity analysis (e.g., according to rate)	2% (2)	2% (1)	2% (1)			
	Sensitivity analysis (e.g., as treated trial data)	1% (1)	0% (0)	2% (1)			
	Not reported	91% (91)	88% (44)	94% (47)			

Category	Item	Review			Review time-to-event outcomes		
		Overall (N = 100)	**Cochrane (n = 50)**	**Non-Cochrane (n = 50)**	**Overall (N = 217)**	**Cochrane (n = 93)**	**Non-Cochrane (n = 124)**
Dealing with treatment switching in individual outcomes	Not reported per outcome	100% (100)	100% (50)	100% (50)	0% (217)	100% (93)	100% (124)
Treatment switching discussed	In discussions	5% (5)	6% (3)	4% (2)			
	In results	5% (5)	8% (4)	2% (1)			
	In results and discussion	2% (2)	2% (1)	2% (1)			
	Not reported	88% (88)	84% (42)	92% (46)			
Treatment switching included in discussion for individual outcomes	Yes	3% (3)	2% (1)	4% (2)	1% (3)	1% (1)	2% (2)
	No	97% (97)	98% (49)	96% (48)	99% (214)	99% (92)	98% (122)
Analyses - Proportional hazards							
Proportional hazards assessed in reviews	Not reported	100% (100)	100% (50)	100% (50)			
Proportional hazards assessed in individual outcomes	Not reported per outcome	100% (100)	100% (50)	100% (50)	0% (217)	100% (93)	100% (124)
Dealing with (non-)proportional hazards	Not applicable	100% (100)	100% (50)	100% (50)			
Test for proportionality for individual outcomes	Not applicable	100% (100)	100% (50)	100% (50)	0% (217)	100% (93)	100% (124)
Non-proportionality of hazards indicated	Not applicable	100% (100)	100% (50)	100% (50)	0% (217)	100% (93)	100% (124)
Dealing with (non-) proportional hazards	Not applicable	100% (100)	100% (50)	100% (50)	0% (217)	100% (93)	100% (124)
Risk of Bias							
Risk of bias tools specified	Risk of Bias 1 (study level)	55% (55)	58% (29)	52% (26)			
	Risk of Bias 1 (outcome level)	23% (23)	42% (21)	4% (2)			
	Other (e.g., CONSORT, MERGE)	6% (6)	0% (0)	12% (6)			
	Jadad scale	4% (4)	0% (0)	8% (4)			
	Risk of Bias 2.0	3% (3)	0% (0)	6% (3)			
	No assessment	9% (9)	0% (0)	18% (9)			
TTE specific risk of bias criteria used	Yes (e.g., "risk of bias related to censoring")	4% (4)	8% (4)	0% (0)			
	No	9% (9)	92% (46)	82% (41)			
	Not applicable	87% (87)	0% (0)	18% (9)			

Abbreviations: TTE = time to event, RoB = risk of bias

Appendix-Table A11: Absolute effects presented in reviews for time-to-event meta-analyses

Category	Item	Review			Review time-to-event outcomes		
		Overall (N = 100)	**Cochrane (n = 50)**	**Non-Cochrane (n = 50)**	**Overall (N = 217)**	**Cochrane (n = 93)**	**Non-Cochrane (n = 124)**
Absolute effects reported	Yes	44% (44)	80% (40)	8% (4)	40% (86)	80% (74)	10% (12)
	No	50% (50)	10% (5)	90% (45)	57% (123)	13% (12)	90% (111)
	Explicitly not calculated (e.g., " because TTE outcome")	6% (6)	10% (5)	2% (1)	4% (8)	8% (7)	1% (1)
Type of absolute effects	Natural frequencies	25% (25)	46% (23)	4% (2)	22% (48)	42% (39)	7% (9)
	Risk difference	13% (13)	24% (12)	2% (1)	13% (28)	29% (27)	1% (1)
	Median survival or difference in median survival	4% (4)	8% (4)	0% (0)	3% (7)	8% (7)	0% (0)
	Natural frequencies and risk difference	2% (2)	2% (1)	2% (1)	1% (3)	1% (1)	2% (2)
	Not applicable	56% (56)	20% (10)	92% (46)	60% (131)	20% (19)	90% (112)
Baseline risk applicable for events or absence of events	Event	26% (26)	48% (24)	4% (2)	24% (51)	52% (48)	2% (3)
	Absence of event	6% (6)	12% (6)	0% (0)	6% (12)	13% (12)	0% (0)
	Unclear	8% (8)	14% (7)	2% (1)	4% (9)	9% (8)	1% (1)
	Not applicable	60% (60)	26% (13)	94% (47)	67% (145)	27% (25)	97% (120)
Description of outcome adapted to direction of baseline risk	No (no changes)	32% (32)	60% (30)	4% (2)	26% (57)	59% (55)	2% (2)
	Yes (description changed with reasoning)	8% (8)	14% (7)	0% (0)	6% (12)	13% (12)	0% (0)
	Yes (description changed without reasoning)	2% (2)	2% (1)	2% (1)	2% (4)	2% (2)	2% (2)
	Unclear	1% (1)	2% (1)	0% (0)	1% (3)	3% (3)	0% (0)
	Not applicable	57% (57)	22% (11)	94% (47)	65% (141)	23% (21)	97% (120)
Absolute effects correctly calculated	Yes	22% (22)	40% (20)	4% (2)	19% (41)	41% (38)	2% (3)
	Correct calculation but wrong labeling	5% (5)	10% (5)	0% (0)	5% (11)	12% (11)	0% (0)
	No	6% (6)	12% (6)	0% (0)	6% (12)	13% (12)	0% (0)
	Unclear (e.g., applicability of baseline risk or HR unclear)	7% (7)	12% (6)	2% (1)	4% (8)	8% (7)	1% (1)
	Not applicable	60% (60)	26% (13)	94% (47)	67% (145)	27% (25)	97% (120)

Abbreviations: TTE = time to event

11.2. Paper 2: Characteristics, methods and reporting of trials included in meta-analyses of time-to-event outcomes

11.2.1. Work shares

Authors: Goldkuhle M, Hirsch C, Iannizzi C, Bora AM, Bender R, Van Dalen EC, Hemkens LG, Monsef I, Trivella, Kreuzberger N, Skoetz N

Contributions by the doctoral student:
This sub-project was funded through the previously described DFG grant, which was acquired under significant participation of the doctoral student. Similar to the previously described sub-project, this meta-epidemiological study was managed and administrated by the doctoral student. He was responsible for its conceptualization, design and the development of all methods of the study, as well as for the data extraction forms and the analysis plan. Similar to the previously described sub-project, the doctoral student was supported by other project participants who reviewed his initial drafts and provided important suggestions. The meta-epidemiological study was performed based on the same study protocol as the previously described sub-project, which was primarily developed by the doctoral student and published under his last authorship (osf.io/6825g/).
The screening of the results of the systematic searches for eligible literature, the extraction of the relevant data and the preparation of study data for analysis were carried out by him. Furthermore, the doctoral student analyzed and interpreted all study data. Due to projects quality standards, a duplicate check of relevant project steps was undertaken by independent project participants. Finally, the doctoral student was responsible for the presentation of the results, the drafts of the publication, their revision according to the comments of the project participants and their revision according to the received comments during peer review.

Co-author contributions:
This project was also supervised by Prof. Skoetz. She supported the doctoral student through her guidance in the design and development of the methodology for the project, as well as her advice on publication and her revision of all drafts. Nina Kreuzberger, Caroline Hirsch, Claire Iannizzi and Ana-Mihaela Bora extracted data independently and in duplicate, supported the analysis of the data and provided important recommendations on the drafts of the publication. Nina Kreuzberger was furthermore involved in screening of the literature searches and supported the conceptual design and development of the methodology. Prof. Ralf Bender, Dr. Elvira van Dalen, PD Dr. Lars Hemkens, and Dr. Marialena Trivella (Oxford, United Kingdom) provided suggestions on the underlying concepts of the sub-project and commented on later versions of the publication. The systematic database search is based on Ina Monsef's search strategy for the previously described work.

11.2.2. Publication appendix

Appendix

Appendix A1: Complete search strategy for non-Cochrane reviews

Medline on February 8th, 2021

#	Searches
1	"time-to-event".tw,kf.
2	"log rank".tw,kf.
3	survival.tw,kf.
4	hazard.tw,kf.
5	Kaplan-meier estimate/
6	kaplan-meier.tw,kf.
7	(method* adj1 (product* or limit*)).tw,kf.
8	(cumulative* adj1 incidence*).tw,kf.
9	outcome expectation.tw,kf.
10	(cox adj2 (model* or proportional*)).tw,kf.
11	proportional hazards models/
12	or/1-11
13	(randomi?ed or placebo or randomly).ab.
14	meta analysis.mp,pt.
15	12 and 13 and 14
16	limit 15 to dt=20170101-20200801

Appendix A2: List of extraction items

For a complete list of extraction items extracted for reviews and their assessed time-to-event outcomes, please see the appendix of Goldkuhle et al. 2023 (*submitted*).

Extraction items for trials

#	Item	Option	Description
1	Review ID		Individual ID (number) of review corresponding to the overall review sheet
2	Trial ID		Individual ID (number) of trial (following alphabetical order from review)
3	Review description of trial		Description of trial in review (if only a reference is used, please choose last name of fist author and publication data of that reference)

4	Other than primary publication	Yes;
		No;
		Not applicable (e.g. no primary publication defined)
			Do you use another than the primary publication in the review for the assessment on overall trial level (e.g. because the review included time-to-event data is only reported in another publication)?

5	Comment on publications of this trial		Field to comment on the publications of this trial in the review
6	First author		Last name of the first author of the trial publication at hand (primary publication or, if applicable, most recent referenced full text publication including relevant outcome data) Primary publication must include results data that was used in the review (If no result data for eligible outcomes included in publication that was labeled in review as "primary publication", please choose most current full-text publication with utilized result data)
7	Publication year		Date of trial publication at hand (primary publication or, if applicable, most recent referenced full text publication including relevant outcome data) Primary publication must include results data that was used in the review (If no result data for eligible outcomes included in publication that was labeled in review as "primary publication", please choose most current full-text publication with utilized result data)
8	Journal		Full title of the journal

9	Publication format	Journal publication (first full publication or not otherwise reported);
		Journal publication (updated analysis);
		Journal publication (final analysis);
		Abstract (e.g. conference presentation);
		Registry entry;
		Clinical trial report;
		Other	

10	Trial PICO		Please enter the PICO of the trial as complete as possible

11	Trial design	Superiority trial or not otherwise specified;
		Non-inferiority trial;
		Equivalency trial;
		Other
			Was this trial designed as a non-inferiority trial, equivalence trial or any other design except superiority?
			If not explicitly reported choose "no"

12	Experimental treatment type	Biologics/drug;
		Surgical procedure;
		Medical devices;
		Radiotherapy;
		Behavioral intervention;
		Exercise intervention;
		Screening;
		Other (please specify)
			Which type of experimental treatment did the trial participants receive?

13	Control treatment type	Placebo;
		No treatment;
		Observation;
		Usual or best-supportive care;
		Biologics/drug;
		Surgical procedure;
		Medical devices;
		Radiotherapy;
		Behavioral intervention;
		Exercise intervention;
		Screening;
		Other (please specify)
			Which type of control treatment did the trial participants receive?

14	Type of follow-up	Median;
		Mean;
			Which type of follow-up measure(s) where reported for the overall trial population.

#	Item	Option	Description
		Minimum follow-up; Maximum follow-up; IQR/ lower and upper range of IQR; 95% CI of median; 95% CI of mean; Standard deviation Fixed time-point of outcome measurement only; No indicator of follow-up reported; Follow-up reported for outcomes;	Irrespective of whether in total or per arm
15	Follow-up calculation		How was follow-up time calculated?
16	Overall follow-up reported		Was a measure of duration of follow-up for the entire analyzed population reported in the trial publication?
17	Median overall follow-up		Median overall trial follow-up time (in months) Empty field = "Not reported"
18	Follow-up reported per arm		Was a measure of follow-up for the population analyzed in each of the compared arms reported in the trial publication?
19	Median experimental follow-up		Median follow-up time in experimental group (in months) Empty field = "Not reported"
20	Median control follow-up		Median follow-up time in control group (in months) Empty field = "Not reported"
21	Field for commenting on the PICO or follow-up in the assessed trial		
22	Total number of TTE outcomes		What was the total number of outcomes compared in the trial as time-to-event outcomes? (quantitatively compared outcomes with relative effect measure only) If necessary type "Unclear" "Not reported"
23	List of TTE outcomes		List of outcomes examined as TTE in this trial according to the assessed trial publication

24	Safety data as TTE outcomes	Yes;
		No;
		Unclear;
		Not reported;
		Not applicable (no safety outcomes reported)
			Where any adverse events (safety data) assessed with time-to-event methodology?

25	Experimental randomization ratio		Randomization ratio experimental arm (1:1 ="1"; 2:1 ="2", ...)
26	Control randomization ratio		Randomization ratio control arm (1:1 ="1"; 1:2 ="2", ...)
27	Randomized experimental participants		What is the total number of participants randomized to the experimental group? Empty field = "Not reported"
28	Randomized control participants		What is the total number of participants randomized to the control group? Empty field = "Not reported"
29	Total randomized participants		
30	Any missing outcome data per arm reported		Was any missing outcome data reported for this trial per arm? (Including that there was NO missing data or similar statements!)
31	Missing outcome data per arm reported	Yes; No; Explicitly reported complete follow-up; Reported for individual outcomes; Reported across arms only; Explicitly reported no LTFU	
32	Total missing outcome data in experimental arm		We extract individuals with missing outcome data (MOD) (individuals clearly have MOD based on RCT reporting), this includes but is not limited to: 1.a: Individuals reported as with "Explained/ Unexplained LTFU", "Outcome not assessable", "Data not available", etc. 1.b: All other reasons, if explicitly reported as not followed-up, excluded, withdrawn, explicitly imputed, etc. 1.c: Otherwise clear that outcome data collection (assessment of outcome events) was not possible for individuals for a given reason (CAVE: Censoring in TTE analysis allows to include individuals with MOD for some duration into the trial and thus the "denominator" (e.g. the first number of individuals at risk under a survival curve)) In case a number of individuals who discontinued treatment is reported (e.g. in Lancet flow-diagrams) - Please only extract numbers of participants for which it is clear that they could not contribute outcome data (e.g. reported as lost to follow-up) If necessary type:

#	Item	Option	Description
			"unclear" "not reported"
33	Total missing outcome data in control arm		We extract individuals with missing outcome data (MOD) (individuals clearly have MOD based on RCT reporting), this includes but is not limited to: 1.a: Individuals reported as with "Explained/ Unexplained LTFU", "Outcome not assessable", "Data not available", etc. 1.b: All other reasons, if explicitly reported as not followed-up, excluded, withdrawn, explicitly imputed, etc. 1.c: Otherwise clear that outcome data collection (assessment of outcome events) was not possible for individuals for a given reason (CAVE: Censoring in TTE analysis allows to include individuals with MOD for some duration into the trial and thus the "denominator" (e.g. the first number of individuals at risk under a survival curve)) In case a number of individuals who discontinued treatment is reported (e.g. in Lancet flow-diagrams) - Please only extract numbers of participants for which it is clear that they could not contribute outcome data (e.g. reported as lost to follow-up) If necessary type: "unclear" "not reported"
34	Number: Control treatment in experimental group		How many individuals in the experimental group received the treatment assigned to the control group? Add number If necessary type: "Unclear" "Not reported"
35	Number: Experimental treatment in control group		How many individuals in the control group received the treatment assigned to the experimental group? Add number If necessary type: "Unclear" "Not reported"

36	Comparator treatments protocol specified	Yes, reported as protocol specified;
		Yes, reported as protocol amendment (after trial start);
		Otherwise reported as anticipated;
		Reported as not planned or allowed
		Other (specify);
		Unclear;
		Not reported;
		Not applicable;
			Was it explicitly reported that the reception of comparator treatments was protocol specified?
			Should be clear from the publication (information only reported in the protocol (not in appendix or publication) should not be counted)

37	Comparator treatments in sample size	Yes;
		Unclear;
		Not reported;
		Not applicable
			Was it reported that the reception of comparator treatments was included in sample size calculations?

38	Comments on comparator treatments		Comments on the reception of comparator treatments in this trial

39	Reason for comparator treatments	Course of disease - related (e.g. disease progression);
		Pre-condition related (e.g. too obese, allergies);
		Intervention related (e.g. adverse events);
		Participant related (e.g. choose to switch);
		Administrative (e.g. interim analysis);
		Other (please specify);
		Unclear;
		Not reported;
		Not applicable
			For what reason did participants in the trial arms receive comparator treatments?

40	Field to comment on the number of participants who received comparator treatments		

41	TTE specific RoB items considered	Yes (please specify);
		No;
		Unclear (please specify);
		Not applicable
			Did the review authors consider any time-to-event specific or related items in their risk of bias assessment for this trial?

42	Comments on the risk of bias assessment		Comments on the risk of bias assessment of the authors on trial outcome level

Abbreviations: CI = confidence interval; IQR = interquartile range; LTFU = loss-to-follow-up; MOD = missing outcome data; PICO = people-intervention-comparator-outcome; RCT = randomized controlled trial; RoB = risk of bias; TTE = time-to-event

Extraction items for trial outcomes

#	Item	Option	Description
1	Outcome ID		Individual number of this outcome (for a hierarchy, see Excel sheet; e.g. OS = 1, PFS = 2, ...) Refers to the outcome in the review (e.g. if the review authors named their outcome progression-free survival, but included data on relapse-free survival from this trial in the very same meta-analysis, please use the outcome ID for PFS)
2	Trial ID		Individual ID (number) of trial (following alphabetical order from review)
3	Review ID		Individual ID (number) of review corresponding to the overall review sheet (assessable in Excel sheet)
4	Trial outcome		Trial outcome assessed in this column Please shorten: overall survival = OS; progression-free survival = PFS; disease free survival = DFS; etc
5	TTE data used	HR and confidence intervals; HR together with other information (e.g. events in each arm, total events, etc.); Observed and expected events or hazard rates on research and control arm; Observed - expected events together with log-rank V; P-value together with additional information (e.g. events, total events, etc.); Survival curves; Median survival times; Time-point specific survival times; IPD (recalculated or from publication); Reported, but method not clear; Other (specify); Not specified for this trial outcome; Unclear	What type of time-to-event data were used for this outcome from the trial to include it in the review publication (according to the review authors)?
6	Specification of recalculated TTE data		Specification on how review authors recalculated time-to-event data for this outcome from the trial to include it in the review publication
7	HR from review		HR of this trial from the review (e.g. in Forest Plot)
8	Lower 95% CI from review		Lower bound 95% CI of this trial from the review (e.g. in Forest Plot)
9	Upper 95% CI from review		Upper bound 95% CI of this trial from the review (e.g. in Forest Plot)

10	TTE specific RoB assessment?	Yes (please specify);
		No;
		Unclear (please specify)
			Did the review authors consider any TTE specific trial characteristics in their outcome-specific rob assessment?

11	Comments on the risk of bias assessment		Comments on the risk of bias assessment of the authors on trial outcome level
12	Trial level information: Please make sure that you extract the following data using a publication of this trial that includes the time-to-event data eligible for the review! If this publication is not the primary publication in the review it must be: -referenced in the review -include TTE outcome date applicable for the population and follow-up included in the review You might have to compare the data included in the review with the data in the referenced trial publications before data extraction. If you are unsure which publication to use, please discuss with the extraction team.		

13	Relevant TTE data in primary publication?	Yes;
		No;
		Only single publication referenced in review;
		No primary publication highlighted
			Is there time-to-event data available in the primary trial publication that is applicable for the review?
			Refers to methodological and result data. Available result data should refer to the follow-up time-point in the review
			If no eligible time-to-event data is available, the extraction for trial_overall and trial_outcome must be performed using the publication with applicable time-to-event data

14	Outcome data in referenced publications?	Yes; No	Result time-to-event (!) data for this outcome (e.g. HR, survival curves or any source used for recalculation in the review) available in the assessed trial publications referenced in the review? If no time-to-event result data for this outcome is available in the trial publications referenced in the review, the extraction STOPS for trial level outcome data
15	Comment when data is not available		Please make a comment when and why data is not available in the primary publication/ publication at hand
16	Was it clear where TTE data for this trial outcome was used from?	Yes, reported by review authors for this trial outcome; Yes, HR corresponds to HR directly available in trial publication (slight deviations e.g. of upper CI due to statistical software should be considered); Yes, because only single source of TTE data in cited publication(s); Unclear, HR was recalculated but source not reported in review; Unclear, because publication where TTE event data for this trial outcome could be reported could not be identified (e.g. among the cited publications);	

#	Item	Option	Description
		No extraction possible, no TTE data in cited publications; No extraction possible, full text or publication where TTE data is reported is not accessible; No extraction possible, data received from secondary source (e.g. contact with authors); No extraction possible, completely unclear which/ whether data was included in review	

17	Primary trial outcome	Yes;
		No;
		No primary/ secondary outcomes defined
			Was this outcome one of the primary outcomes of the trial?

18	Outcome definition		Please provide the complete outcome definition from the trial publication

19	Composite outcome	Yes;
		No;
		Unclear;
		Not reported
			Was the outcome a combined outcome including several events of interests
			
			e.g. progression-free survival - progressive disease, overall mortality

20	Composite outcomes described?	Yes;
		No;
		Not applicable ("Composite outcome" unclear or not reported)
			Were the outcome events composing this composite outcome described?

21	Outcome events consistent with review?	Yes;
		No;
		Unclear;
		Not applicable
			Were the outcome events in the definition used in the trial consistent with the outcome definition in the review?

22	Start of outcome assessment reported	Yes;
		No;
		Unclear;
		Not applicable (e.g. outcome not defined)
			Was the start time-point of outcome assessment for this outcome reported (e.g. in the outcome definition or in the statistical methods section)?

23	Outcome assessment start	Randomization;
		Enrollment;
		Allocated treatment;
		Previous treatment (e.g. surgery);
		Other (specify);
		Not applicable (e.g. start of follow-up not reported);
		Unclear
			What was the defined or otherwise reported start time-point of outcome assessment for this outcome?

24	Competing events possible?	Yes;
		No;
		Unclear
			Are competing events possible by definition of the outcome?
			
			Choose yes, e.g. if overall mortality was not part of the defined outcome

25	Censoring reasons reported?	Yes;
		Unclear;
		Not reported
			Were any reasons for censoring individuals for this outcome reported?
			
			Censoring reasons are sometimes reported together with the definition of the outcome and sometimes in the statistical analysis section.

26	Censoring reasons	Participants last known to be event-free;
		End of follow-up;
		Loss to follow-up;
		Inadequate outcome assessment;
		Participant withdrawal or consent withdrawal;
		Alternative treatment;
		Treatment discontinuation;
		Other (specify);
		Unclear;
		Not applicable (no details on censoring reported for this outcome)
			What were the reasons for censoring for this outcome, if reported?
			
			Please choose the most applicable.

27	Field for commenting on the outcome definition		Field for commenting on the outcome definition
28	"Which and what kind of time-to-event data was available in the trial publication?" Refers to all time-to-event data in trial publication		

29	Available time-to-event data	HR or log(HR);
		Observed and expected events (log-rank) or hazard rates;
		P-value (log-rank);
		Survival curves;
		Restricted Mean Survival Time;
		Median survival times (per arm);
		Time-point specific survival rates (per arm);
		Median cumulative incidence (per arm);
		Time-point specific cumulative incidence (per arm);
		Greys Test;
		Wilcoxon-Gehan test;
		Mean and SD per arm;
		Other (specify);
		Type of test unclear or not reported
			What types of time-to-event data were available in the assessed trial publications for this outcome (excl. time-point specific or median survival times)?

30	Methods for HRs	Cox model;
		Fine and Gray;
		Parametric model (specify);
		Log-rank;
		Other (specify);
		No HR calculated;
			If hazard ratios (HR, log(HR), etc.) were available, which methods were used to calculate them?

#	Item	Option	Description
		Unclear; Not reported	

31	Available types of analyses (e.g. ITT, PP)	ITT;
		Modified ITT;
		Per-protocol;
		As treated;
		Unclear; not reported
			Which types of analyses are available (e.g. ITT or PP) in the trial publications for this outcome?
			Please use description by trial authors (e.g. if an analysis was labeled "ITT" and there were post randomization exclusions still use "ITT" and not "mITT". Whether it was "a real ITT analysis" is assessed in the next item.)
			Please choose all available analyses in the publication at hand (the type of analysis that was used in the review will be specified in the following)

32	ITT analysis in complete population?	Yes;
		No;
		Unclear;
		Not applicable (no ITT analysis mentioned, e.g. only mITT)
			If an analysis according to the ITT principle was described by the authors, was this analysis performed in the complete allocated trial population or were there post-randomization exclusions (e.g. participants did not receive the intended treatment, were mistakenly enrolled, did withdraw consent, died or developed the outcome of interest before treatment)

33	(Un)adjusted/ (un)stratified HRs available?	Unadjusted (univariate including treatment variables only);
		Adjusted, baseline characteristics;
		Adjusted, post-baseline exposure;
		Adjusted, but factors unclear/ not reported;
		Stratified, but factors unclear;
		Stratified, randomization stratification factors;
		Stratified, baseline characteristics;
		Other (please specify);
		Unclear;
		Not reported;
		Not applicable (no HR directly reported)
			Were unadjusted, adjusted and/ or stratified HRs available in the trial publications for this outcome and if adjusted for which factors?
			Please choose all available HRs in the publication at hand (the type of analysis that was used in the review will be specified in the following)

34	(Un)adjusted/ (un)stratified P-values available?	Unadjusted (univariate including treatment variables only);
		Adjusted, baseline characteristics;
		Adjusted, post-baseline exposure;
		Adjusted, but factors unclear/ not reported;
		Stratified, but factors unclear;
		Stratified, randomization stratification factors;
		Stratified, baseline characteristics;
		Other (please specify);
		Unclear;
		Not reported;
		Not applicable (no log-rank P-value directly reported)	

35	Field to comment on methods (e.g. specify method to calculate relative effect measures)		
36	Methods - Limited to outcome analysis included in meta-analysis "What was reported in the trial publication for the data included in the meta-analysis?" Refers to the outcome analysis included in the review meta-analysis only		

37	Type of analysis included in meta-analysis	ITT;
		modified ITT;
		per-protocol;
		as treated;
		unclear;
		not reported;
			Please indicate the analysis producing the estimate (e.g. HR, log-rank results, survival curves) included in the review meta-analysis as labeled by the trial authors.
			Please use description by trial authors (e.g. if the analysis was labeled "ITT" and there were post randomization exclusions still use "ITT" and not "mITT". Whether it was "a real ITT analysis" is assessed in the next item.)

38	Selected analysis in complete population?	Yes;
		No;
		Unclear;
		Not reported;
		Not applicable (e.g. only subgroup analysis included in review)
			Was this analysis performed in the complete allocated trial population?

39	Population analyzed in allocated arm?	Yes;
		No;
		Unclear;
		Not reported
			Were individuals analyzed in the arm they were allocated too for this outcome analysis (except those who were excluded from the sample, e.g. because of mITT)?

40	Pooled estimate unadjusted, adjusted or stratified?	Unadjusted;
		Adjusted;
		Stratified;
		Unclear;
		Not reported
			Was the effect estimate that was pooled in the meta-analysis for this trial an unadjusted or adjusted estimate (applicable to any type of available effect measure, e.g. HR, Observed - expected, log-rank results, survival curves)?
			Survival curves and median/ time-point specific survival probabilities calculated with Kaplan-Meier or cumulative incidence are expected to be unadjusted - if explicitly reported otherwise, please indicate by comment

41	Survival plots presented?	Yes, Kaplan-Meier;
		Yes, cumulative incidence;
		Yes, other (please specify);
		No, no graphs were presented
			Were survival plots presented for the assessed analysis?

42	Individuals at risk reported?	Yes;	Was the number of individuals at risk over time reported

#	Item	Option	Description
		No; Not applicable	along the survival curve for the assessed analysis
43	Censored observations presented?	Yes, marked on the survival curve; Yes, reported together with the number of individuals at risk; No; Not applicable	Were censored observations presented for the assessed analysis?
44	Censoring balanced?	Yes; No; Unclear; Not applicable	Please make a judgement whether censoring was balanced between arms or pattern in the trial groups differed over time to a degree that is not corresponding to event rates No = More individuals censored in one trial arm compared to the other or pattern in groups differing over time to a degree that is not corresponding to event rates (e.g. early censoring in one group compared to late censoring in the other

45	Proportional hazards tested?	Yes, visual inspection of curves;
		Yes, statistical test (e.g. Log-log, Schoenfeld Residuals));
		No;
		Not applicable (e.g. no HRs calculated)
			Was the proportional hazards assumption tested for the assessed analysis by the trial conductors?

46	Outcome of proportional hazards assessment	Reasonably proportional;
		Non-proportional;
		Not applicable;
		Not reported
			What was the outcome of the authors assessment of proportional hazards for the assessed analysis?

47	If analyzed population differs: Experimental individuals		If the analyzed population differs from allocated population (e.g. "mITT", "PP or "as treated" analysis, separate adjusted analysis, exclusion for missing outcome data, ...): number of individuals analyzed in experimental arm If necessary type: "Unclear" "Not reported" "Not applicable"
48	If analyzed population differs: Control individuals		If the analyzed population differs from allocated population (e.g. "mITT", "PP or "as treated" analysis, separate adjusted analysis, exclusion for missing outcome data, ...): number of individuals analyzed in control arm If necessary type: "Unclear" "Not reported" "Not applicable"

49	MOD specifically reported?	Yes;
		No;
		Unclear;
		Complete follow-up/ no LTFU reported at trial level;
		Complete follow-up/ no LTFU visible on trial outcome level
			Missing outcome data specifically reported for this outcome analysis?

50	MOD differing from "MOD of allocated population"	Yes; No	Does missing outcome data for this analysis differ from "missing outcome data for allocated population" (e.g. because of mITT, PP or as treated analysis, separate adjusted analysis, ...): -Data not already excluded in analysis set -Irrespective of whether explicitly reported or not Use explicitly reported data before using data reported for "missing outcome data for allocated population" and subtracting the individuals excluded (e.g. mITT - individuals excluded before treatment)
51	Comparator interventions specifically reported?	Yes; No	Reception of comparator interventions specifically reported for this outcome?
52	Competing events in experimental group		How many patients experienced a competing event in the experimental group? If necessary type: "Unclear" "Not reported" "Not applicable"
53	Competing events in control group		How many patients experienced a competing event in the control group? If necessary type: "Unclear" "Not reported" "Not applicable"
54	Comments on sample size and numbers		
55	Events in experimental arm		Number of events in experimental arm If necessary type: "Not reported" "Unclear"
56	Events in control arm		Number of events in control arm If necessary type: "Not reported" "Unclear"
57	Final experimental number at risk		Final number at risk at last follow-up in experimental arm from curve that is applicable to analysis

#	Item	Option	Description
			If necessary type: "Unclear" "Not applicable" (if not curves or no number at risk for this analysis are reported)
58	Final control number at risk		Final number at risk at last follow-up in control arm from curve that is applicable to analysis If necessary type: "Unclear" "Not applicable" (if not curves or no number at risk for this analysis are reported)
59	Comments regarding the sample size		
60	Applicable HR		Hazard ratio applicable to meta-analysis as reported in the trial publication (if HR reported as effect measure) Empty field = "Not reported/ unclear" NA = "Not applicable"
61	Applicable lower 95% CI		Lower 95% CI for the assessed analysis (if HR reported as effect measure) Empty field = "Not reported/ unclear" NA = "Not applicable"
62	Applicable upper 95% CI		Upper 95% CI for the assessed analysis (if HR reported as effect measure) Empty field = "Not reported/ unclear" NA = "Not applicable"

63	HR <1 increased or decreased risk of event in experimental group	Decreased risk;
		Increased risk;
		Unclear;
		Not applicable (e.g. no HR calculated)
			Does a HR <1 indicate an increased or decreased risk of the outcome in the group that is assessed as experimental group in the trial publication?

64	HR for events or absence of events?	Event;
		Absence of event;
		Unclear;
		Not applicable (e.g. no HR calculated)
			Is the respective HR from this trial publication applicable for events (e.g. death, relapse) or absence of events (e.g. overall survival, all cause mortality)?

65	Applicable HR directly available	Was inverted;
		Was pooled from multiple study arms (as reported in review);
		Other (specify);
		Not applicable (e.g. no HR calculated);
		HR(s) from study differ from HR in MA (HR likely recalculated)
			How did the review authors include this HR into the meta-analysis?
			
			Is this HR directly available from the assessed trial publication or was it altered in any way by the review authors? (e.g. inverted, pooled from more than one experimental arm, etc.)

66	Standard error		Standard error (if HR not reported or not the appropriate effect measure) If necessary type: "Not reported" "Not applicable"
67	Variance		Variance (log-rank; if HR not reported or not the appropriate effect measure)
68	P-value		P-value of statistical test of group comparison (logrank if not otherwise (e.g. Mantel-Hanezel, Cox) reported in comment field; if HR not reported or not the appropriate effect measure) Empty field = "Not reported/ unclear" NA = "Not applicable"
69	Field for comments on effect measures		Field for comments on effect measures (e.g. type of test, where necessary: Log-rank observed minus-expected events, etc.)
70	Follow-up time specifically reported?	Yes; No	Information on duration of follow-up specifically reported for this outcome? "Specifically" could be outcome specific in primary publication/ publication at hand or in separate publication that includes data for this outcome
71	Comments on trial outcome follow-up		Comments on trial outcome follow-up if reported specifically

72	Missing data handling	Censored;
		Excluded from analysis;
		Single imputation;
		Multiple imputation;
		Sensitivity analyses;
		No missing data; other (please specify);
		Unclear;
		Not reported;
		Not applicable (e.g. no effect measure calculated);
		Complete follow-up/ no LTFU reported at trial level
			How was missing data handled?

73	Comment on MOD analyses		E.g. did the advanced methods to assess the robustness of results for this outcome towards MOD change interpretation?

74	Advanced methods competing events	Fine and Gray and cumulative incidence curves;
		Cumulative incidence curves;
		Other;
		No;
		Not applicable
			Were any advanced methods to assess the robustness of results for this outcome towards competing events used?

#	Item	Option	Description
75	Comment competing event methods		E.g. did the advanced methods to assess the robustness of results for this outcome towards competing events change interpretation?

76	Advanced methods informative censoring	Rank preserving structural failure time;
		Inverse probability (censoring) weighting;
		Iterative parameter estimation;
		Multiple;
		Other;
		No;
		Not applicable
			Where any advanced methods to assess the robustness of results for this outcome towards informative censoring (non-administrative censoring) used?

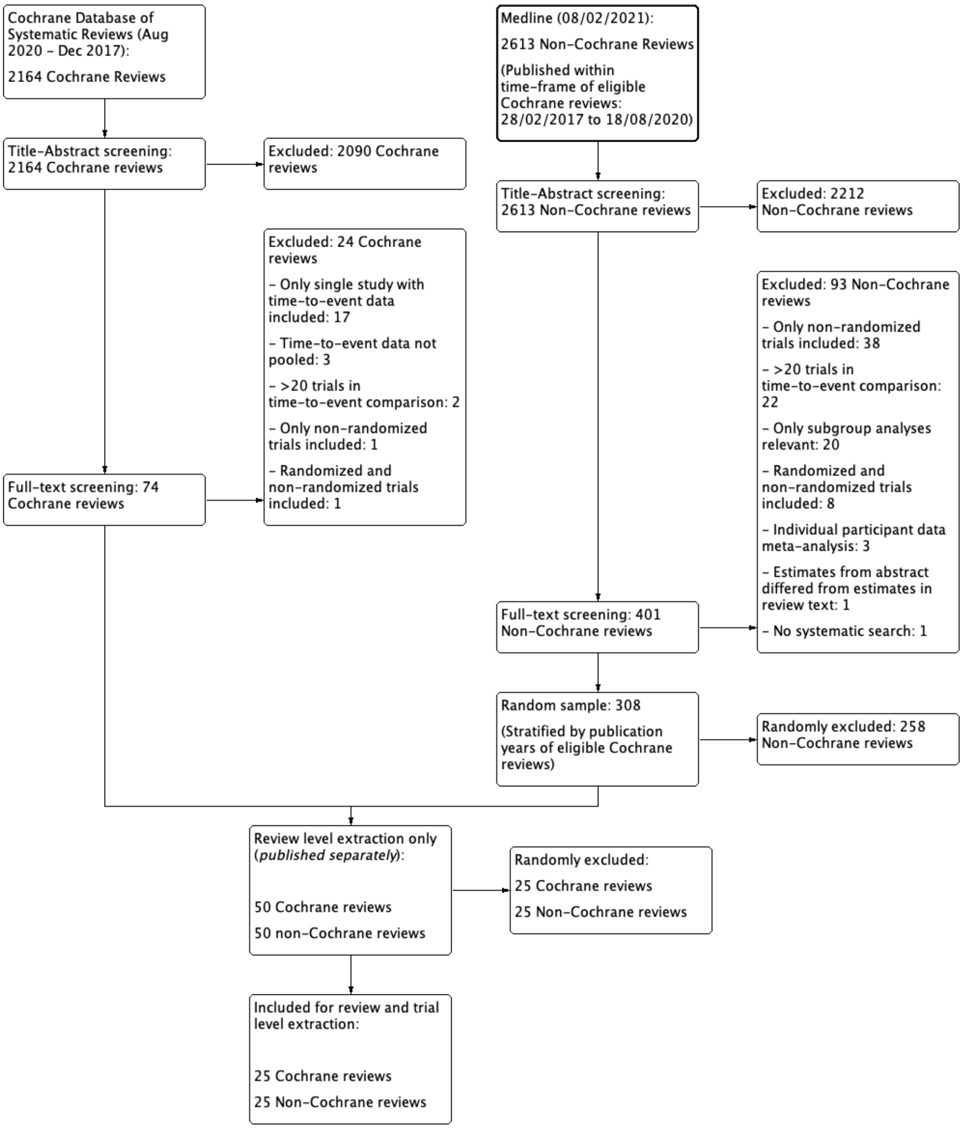
77	Comment informative censoring methods		E.g. did the advanced methods to assess the robustness of results for this outcome towards informative censoring (non-administrative censoring) change interpretation?

78	Advanced methods comparator treatments	Rank preserving structural failure time;
		Inverse probability (censoring) weighting;
		Iterative parameter estimation;
		Multiple;
		Other;
		No;
		Not applicable
			Were any advanced methods to assess the robustness of results towards the reception of comparator treatments in trial participants used?

79	Comment advanced methods comparator treatments		E.g. did the advanced methods to assess the robustness of results towards the reception of comparator treatments in trial participants change interpretation?
80	Comments on time-to-event specific methods or alternative TTE analytic methods		General comments on advanced time-to-event specific methods or alternative time-to-event analytic methods, that were included in the trial report (e.g. Inverse Probability (Censoring) Weighting applied to adjust for specific event, such as the reception of a relevant third intervention)

Abbreviations: CI = confidence interval; HR = hazard ratio; ITT = intention-to-treat; IQR = interquartile range; IPD = individual participant data; LTFU = loss-to-follow-up; mITT = modified intention-to-treat; MOD = missing outcome data; PP = per protocol; RoB = risk of bias; SD = standard deviation; TTE = time-to-event

Appendix A3: Flow-diagram


Cochrane Database of Systematic Reviews (Aug 2020 – Dec 2017):
2164 Cochrane Reviews
Title-Abstract screening: 2164 Cochrane reviews
Excluded: 2090 Cochrane reviews
Excluded: 24 Cochrane reviews
– Only single study with time-to-event data included: 17
– Time-to-event data not pooled: 3
– >20 trials in time-to-event comparison: 2
– Only non-randomized trials included: 1
– Randomized and non-randomized trials included: 1
Full-text screening: 74 Cochrane reviews
Medline (08/02/2021):
2613 Non-Cochrane Reviews
(Published within time-frame of eligible Cochrane reviews: 28/02/2017 to 18/08/2020)
Title-Abstract screening: 2613 Non-Cochrane reviews
Excluded: 2212 Non-Cochrane reviews
Excluded: 93 Non-Cochrane reviews
– Only non-randomized trials included: 38
– >20 trials in time-to-event comparison: 22
– Only subgroup analyses relevant: 20
– Randomized and non-randomized trials included: 8
– Individual participant data meta-analysis: 3
– Estimates from abstract differed from estimates in review text: 1
– No systematic search: 1
Full-text screening: 401 Non-Cochrane reviews
Random sample: 308
(Stratified by publication years of eligible Cochrane reviews)
Randomly excluded: 258 Non-Cochrane reviews
Review level extraction only (*published separately*):
50 Cochrane reviews
50 non-Cochrane reviews
Randomly excluded:
25 Cochrane reviews
25 Non-Cochrane reviews
Included for review and trial level extraction:
25 Cochrane reviews
25 Non-Cochrane reviews

Appendix A4: Characteristics of trials and trial outcomes that data could not be extracted for

Domain		Trial (N = 13)	Trial outcome (N = 18)	Review (N = 50)
Review type	Cochrane	69% (9)	67% (12)	12% (6)
	Non-Cochrane	31% (4)	33% (6)	8% (4)
Multiple publications of trial referenced in review	Yes	54% (7)		10% (5)
Other than primary trial publication used for extraction	Yes	8% (1)		2% (1)
	No	23% (3)		6% (3)
	Not applicable (e.g. no primary publication defined)	69% (9)		14% (7)
Relevant TTE data in primary publication for particular outcome	No		39% (7)	6% (3)
	No primary publication highlighted		17% (3)	4% (2)
	Only single publication referenced in review		44% (8)	12% (6)
Reasons why extraction was not possible	No TTE data in cited publications		56% (10)	10% (5)
	Completely unclear which/ whether data was included in review		22% (4)	6% (3)
	Full text or publication where TTE data is reported is not accessible		11% (2)	4% (2)
	Data received from secondary source (e.g. contact with authors)		11% (2)	2% (1)
Review outcomes adopted from trials	All-cause mortality/ Overall survival		67% (12)	18% (9)
	Disease-free survival		11% (2)	4% (2)
	Composite of all-cause death, myocardial infarction or stroke		6% (1)	2% (1)
	Progression-free survival		6% (1)	2% (1)
	Thrombolysis in myocardial infarction (TIMI) major bleeding		6% (1)	2% (1)
	Time to death from prostate cancer		6% (1)	2% (1)
Methods to recalculate TTE data reported by review authors	HR and confidence intervals		22% (4)	2% (1)
	Time-point specific survival times		11% (2)	2% (1)
	P-value together with additional information (e.g. events)		6% (1)	2% (1)
	Not specified for this trial outcome		61% (11)	16% (8)
Relative effect measures included in review	Favorable, statistically significant		11% (2)	4% (2)
	Favorable, statistically non-significant (CI crosses 1)		28% (5)	8% (4)
	Unfavorable, statistically significant		6% (1)	2% (1)
	Unfavorable, statistically significant non-significant (CI crosses 1)		50% (9)	10% (5)
	Direction of effect unclear (HR = 1)		6% (1)	2% (1)
TTE specific risk of bias rating at trial level	No		100% (18)	
Abbreviations: CI = confidence interval, HR = hazard ratio, TTE = time-to-event				

Appendix A5: Characteristics of included reviews

		Review			Review outcome		
Domain		Overall (N = 50)	Cochrane (n = 25)	Non-Cochrane (n = 25)	Overall (N = 70)	Cochrane (n = 33)	Non-Cochrane (n = 37)
Publication							
Publication year	2017	26% (13)	24% (6)	28% (7)			
	2018	26% (13)	24% (6)	28% (7)			
	2019	28% (14)	28% (7)	28% (7)			
	2020	20% (10)	24% (6)	16% (4)			
	Median (IQR)	2018 (2017.25 - 2019)	2019 (2018 - 2019)	2018 (2017 - 2019)			
	Mean (range)	2018.42 (2017 - 2020)	2018.52 (2017 - 2020)	2018.32 (2017 - 2020)			
Journal Impact Factor (2021)	Median (IQR)		12.008	4,501 (3.372 – 6.883)			
	Mean (range)		12.008	8,06 (1.817 – 35.855)			
Review updates		28% (14)	48% (12)	8% (2)			
Multiple review comparisons		30% (15)	44% (11)	16% (4)			
Population							
Medical field	Neoplasms	82% (41)	88% (22)	76% (19)			
	Diseases of the circulatory system	14% (7)	8% (2)	20% (5)			
	Diseases of the skin and subcutaneous tissue	6% (3)	12% (3)	0% (0)			
	Diseases of blood, blood-forming organs, immune mechanism	4% (2)	4% (1)	4% (1)			
	Other	4% (2)	4% (1)	4% (1)			
Medical condition	Breast cancer	14% (7)	12% (3)	16% (4)			
	Colorectal cancer	8% (4)	12% (3)	4% (1)			
	Prostate cancer	6% (3)	12% (3)	0% (0)			
	Biliary tract cancer	4% (2)	0% (0)	8% (2)			
	Gastric cancer	4% (2)	0% (0)	8% (2)			
	Non-ischaemic cardiomyopathy	4% (2)	0% (0)	8% (2)			
	Non-small cell lung cancer	4% (2)	4% (1)	4% (1)			
	Ovarian cancer	4% (2)	8% (2)	0% (0)			
	Other	52% (26)	52% (13)	52% (13)			
Clinical stage	Early/ First line	34% (17)	44% (11)	24% (6)			
	Advanced/ Second or third line	30% (15)	24% (6)	36% (9)			
	No restriction	20% (10)	24% (6)	16% (4)			
	Not reported	2% (1)	0% (0)	4% (1)			
	Not applicable	14% (7)	8% (2)	20% (5)			
Age group	Adults	92% (46)	96% (24)	88% (22)			
	Both	2% (1)	4% (1)	0% (0)			

Domain		Review			Review outcome		
		Overall (N = 50)	Cochrane (n = 25)	Non-Cochrane (n = 25)	Overall (N = 70)	Cochrane (n = 33)	Non-Cochrane (n = 37)
	Not reported	6% (3)	0% (0)	12% (3)			
Interventions							
Comparisons	Biologics/ drug vs. Biologic/ drug	32% (16)	20% (5)	44% (11)			
	Surgical procedure vs. Surgical procedure	8% (4)	12% (3)	4% (1)			
	Biologics/ drug vs. Observation	6% (3)	0% (0)	12% (3)			
	Biologics/ drug (schedule alteration)	4% (2)	4% (1)	4% (1)			
	Biologics/ drug vs. Placebo	4% (2)	0% (0)	8% (2)			
	Follow-up strategy vs. Follow-up strategy	4% (2)	8% (2)	0% (0)			
	Other	42% (21)	56% (14)	28% (7)			
Comparator treatment considered?	No	100% (50)	100% (25)	100% (25)			
Outcomes - Planned							
Planned outcome number	Median (IQR)	6 (4 - 8)	7 (6 - 8)	4 (2 - 5)			
	Mean (range)	5.89 (1 - 15)	6.72 (3 - 10)	4.95 (1 - 15)			
Planned TTE outcome number	Median (IQR)	2 (2 - 2)	2 (2 - 2)	2 (1.75 - 2)			
	Mean (range)	2.35 (1 - 12)	2.00 (1 - 3)	2.71 (1 - 12)			
Number of outcomes analyzed	Median (IQR)	5 (3 - 6)	5 (5 - 6)	4 (2 - 5)			
	Mean (range)	5.24 (1 - 18)	5.44 (1 - 10)	5.04 (1 - 18)			
Number of TTE outcomes analyzed	Median (IQR)	2 (1 - 2)	2 (1 - 2)	2 (1 - 2)			
	Mean (range)	2.22 (1 - 12)	1.80 (1 - 3)	2.64 (1 - 12)			
Outcomes - Definition							
Outcome reporting per review	Absence of event only	56% (28)	40% (10)	72% (18)			
	Event only	26% (13)	24% (6)	28% (7)			
	Both (with reasoning) only	6% (3)	12% (3)	0% (0)			
	Mixed	4% (2)	8% (2)	0% (0)			
	At least one unclear	8% (4)	16% (4)	0% (0)			
Outcome reporting per individual outcome	Absence of event				59% (41)	45% (15)	70% (26)
	Event				26% (18)	21% (7)	30% (11)
	Both (with reasoning)				10% (7)	21% (7)	0% (0)
	Unclear				6% (4)	12% (4)	0% (0)
Reviews including follow-up start in outcome definitions	Randomization	38% (19)	60% (15)	16% (4)			
	Allocated treatment	4% (2)	4% (1)	4% (1)			
	Enrollment	2% (1)	4% (1)	0% (0)			
	At least one not applicable	56% (28)	32% (8)	80% (20)			
Follow-up start included in outcome definition	Randomization				43% (30)	67% (22)	22% (8)
	Allocated treatment				4% (3)	3% (1)	5% (2)
	Enrollment				1% (1)	3% (1)	0% (0)
	Not reported				51% (36)	27% (9)	73% (27)
Reviews mentioning heterogeneous TTE outcome definitions	In discussion	4% (2)	0% (0)	8% (2)			
	In results	2% (1)	4% (1)	0% (0)			
	Not reported	94% (47)	96% (24)	92% (23)			
Heterogeneous outcome definitions discussed	Yes	2% (1)	0% (0)	4% (1)	1% (1)	0% (0)	3% (1)
	No	100% (50)	100% (25)	100% (25)	99% (69)	100% (33)	97% (36)
Follow-up							

Domain		Review			Review outcome		
		Overall (N = 50)	**Cochrane (n = 25)**	**Non-Cochrane (n = 25)**	**Overall (N = 70)**	**Cochrane (n = 33)**	**Non-Cochrane (n = 37)**
Reviews reporting a planned follow-up duration	Minimum duration of follow-up required	4% (2)	4% (1)	4% (1)			
	Longest follow-up	2% (1)	4% (1)	0% (0)			
	Maximum duration of follow-up specified	2% (1)	4% (1)	0% (0)			
	Not reported	92% (46)	88% (22)	96% (24)			
Follow-up time specification for TTE outcomes	Longest follow-up	6% (3)	8% (2)	4% (1)	6% (4)	6% (2)	5% (2)
	Minimum duration of follow-up required	2% (1)	4% (1)	0% (0)	3% (2)	6% (2)	0% (0)
	Maximum duration of follow-up specified	2% (1)	4% (1)	0% (0)	1% (1)	3% (1)	0% (0)
	Not reported	90% (45)	84% (21)	96% (24)	90% (63)	85% (28)	95% (35)
Sample size							
Number of included studies in reviews and meta-analyses	Median (IQR)	5 (4 - 8)	5 (4 - 9)	5 (4 - 7)	4 (2.25 - 5)	3 (2 - 5)	4 (4 - 5)
	Mean (range)	7 (2 - 21)	7 (2 - 21)	5 (2 - 13)	5 (2 - 15)	4 (2 - 15)	5 (2 - 12)
Total population in review or meta-analysis	Median (IQR)	1697 (957 - 3838)	1184 (505 - 4190)	1728 (1370 - 3252)	1521 (571 - 4580.5)	571 (351 - 1741)	1948 (1455 - 5093)
	Mean (range)	3621 (307 - 38723)	3229 (307 - 13216)	3962 (343 - 38723)	4133 (181 - 38723)	2369 (181 - 12528)	6117 (623 - 38723)
	Not reported	14% (7)	20% (5)	8% (2)	27% (19)	18% (6)	35% (13)
Analyses - Comparative effect measures							
HR type eligible in reviews	HR/ log(HR) not further specified	88% (44)	92% (23)	84% (21)			
	HR/ log(HR) from Cox model	2% (1)	4% (1)	0% (0)			
	Cox model HR/ log HR, log-rank and Kaplan Meier Curve	2% (1)	4% (1)	0% (0)			
	Not reported	8% (4)	0% (0)	16% (4)			
HR types eligible per outcome	HR/ log(HR) from Cox model	2% (1)	4% (1)	0% (0)	1% (1)	3% (1)	0% (0)
	Not reported	98% (49)	96% (24)	100% (25)	99% (69)	97% (32)	100% (37)
Methods to obtain TTE data per review	HR and confidence intervals	56% (28)	64% (16)	48% (12)			
	Specified set of methods (e.g. Tierney 2008 (4))	44% (22)	76% (19)	12% (3)			
	log(HR) and standard error	20% (10)	32% (8)	8% (2)			
	Survival curves	18% (9)	20% (5)	16% (4)			
	HR with other information (e.g. events)	8% (4)	16% (4)	0% (0)			
	IPD (recalculated or from publication)	4% (2)	4% (1)	4% (1)			
	P-value with additional information (e.g. events)	4% (2)	4% (1)	4% (1)			
	Median survival times	2% (1)	4% (1)	0% (0)			
	Reported, but method not clear	2% (1)	4% (1)	0% (0)			
	Risk ratio	2% (1)	0% (0)	4% (1)			
	Unclear	4% (2)	4% (1)	4% (1)			
	Not reported	20% (10)	0% (0)	40% (10)			
Recalculation of TTE data reported for an outcome	Yes	36% (18)	28% (7)	4% (1)			
	Not reported	164% (82)	132% (33)	196% (49)			
Methods to obtain TTE data for an outcome	HR and confidence intervals	6% (3)	12% (3)	0% (0)	7% (5)	15% (5)	0% (0)
	P-value with additional information (e.g. events)	4% (2)	8% (2)	0% (0)	3% (2)	6% (2)	0% (0)
	Survival curves	4% (2)	4% (1)	4% (1)	3% (2)	3% (1)	3% (1)
	IPD (recalculated or from publication)	2% (1)	4% (1)	0% (0)	1% (1)	3% (1)	0% (0)

Domain		Review			Review outcome		
		Overall (N = 50)	Cochrane (n = 25)	Non-Cochrane (n = 25)	Overall (N = 70)	Cochrane (n = 33)	Non-Cochrane (n = 37)
	Time point specific survival times	2% (1)	4% (1)	0% (0)	3% (2)	6% (2)	0% (0)
	Unclear	6% (3)	12% (3)	0% (0)	6% (4)	12% (4)	0% (0)
Analyses - ITT/ PP							
Types of analyses eligible in reviews	ITT	42% (21)	76% (19)	8% (2)			
	Not reported	58% (29)	24% (6)	92% (23)			
Types of analyses eligible for outcome analyses	Not reported	100% (50)	100% (25)	100% (25)	100% (70)	100% (33)	100% (37)
Types of analyses included in reviews	ITT	18% (9)	20% (5)	16% (4)			
	Included trial(s) did not report type of analysis	6% (3)	8% (2)	4% (1)			
	Not reported for all trials	16% (8)	24% (6)	8% (2)			
	Not reported for any trial	66% (33)	56% (14)	76% (19)			
Types of analyses included in outcome analyses	ITT	2% (1)	4% (1)	0% (0)	3% (2)	6% (2)	0% (0)
	Not reported for all trials	2% (1)	4% (1)	0% (0)	1% (1)	3% (1)	0% (0)
	Not reported	96% (48)	92% (23)	100% (25)	96% (67)	91% (30)	100% (37)
Unadjusted/ adjusted HRs eligible in reviews	Hierachical (adjusted before unadjusted)	4% (2)	8% (2)	0% (0)			
	Both	2% (1)	4% (1)	0% (0)			
	Unadjusted only	2% (1)	4% (1)	0% (0)			
	Adjusted only	2% (1)	4% (1)	0% (0)			
	Hierachical (unadjusted before adjusted)	2% (1)	4% (1)	0% (0)			
	Unclear	6% (3)	12% (3)	0% (0)			
	Not reported	82% (41)	64% (16)	100% (25)			
Dealing with unadjusted/ adjusted HRs	Unclear	8% (4)	16% (4)	0% (0)			
	Not reported	10% (5)	20% (5)	0% (0)			
	Not applicable	82% (41)	64% (16)	100% (25)			
Stratified HRs eligible in reviews	Yes	2% (1)	4% (1)	0% (0)			
	No	98% (49)	96% (24)	100% (25)			
Unadjusted/ adjusted HRs eligible in outcome analyses	Unadjusted only	2% (1)	4% (1)	0% (0)	1% (1)	3% (1)	0% (0)
	Adjusted only	2% (1)	4% (1)	0% (0)	1% (1)	3% (1)	0% (0)
	Not reported	96% (48)	92% (23)	100% (25)	97% (68)	94% (31)	100% (37)
Dealing with unadjusted/ adjusted HRs	Only adjusted HRs included, others likely excluded	2% (1)	4% (1)	0% (0)	1% (1)	3% (1)	0% (0)
	Unadjusted HRs recalculated	2% (1)	4% (1)	0% (0)	1% (1)	3% (1)	0% (0)
	Not applicable	96% (48)	92% (23)	100% (25)	97% (68)	94% (31)	100% (37)
Unadjusted/ adjusted HRs discussed in reviews	In results	2% (1)	4% (1)	0% (0)			
	Not reported	96% (48)	92% (23)	100% (25)			
Unadjusted/ adjusted analyzed discussed for individual outcomes	No	100% (50)	100% (25)	100% (25)	100% (70)	100% (33)	100% (37)
Analyses - Varying follow-up between included trials							
Dealing with varying follow-up in reviews	Sensitivity analyses (e.g. shorter/longer)	10% (5)	16% (4)	4% (1)			
	Included in meta-regression	4% (2)	0% (0)	8% (2)			
	Exclusion of studies with divergent follow-up time	2% (1)	0% (0)	4% (1)			
	Included in interpretation of heterogeneity	2% (1)	4% (1)	0% (0)			
	Mentioned as RoB criterion in methods	2% (1)	4% (1)	0% (0)			
	Unclear	2% (1)	0% (0)	4% (1)			
	Not reported	78% (39)	76% (19)	80% (20)			

Domain		Review			Review outcome		
		Overall (N = 50)	Cochrane (n = 25)	Non-Cochrane (n = 25)	Overall (N = 70)	Cochrane (n = 33)	Non-Cochrane (n = 37)
Dealing with varying follow-up for individual outcomes	Results reported for multiple time-points	2% (1)	4% (1)	0% (0)	1% (1)	3% (1)	0% (0)
	Not reported per outcome	98% (49)	96% (24)	100% (25)	99% (69)	97% (32)	100% (37)
Varying follow-up discussed	In discussion	18% (9)	16% (4)	20% (5)			
	In results	8% (4)	8% (2)	8% (2)			
	In results and in discussion	6% (3)	8% (2)	4% (1)			
	Not reported	68% (34)	68% (17)	68% (17)			
Varying follow-up discussed for individual outcomes	Yes	6% (3)	4% (1)	8% (2)	4% (3)	3% (1)	5% (2)
	No	96% (48)	96% (24)	96% (24)	96% (67)	97% (32)	95% (35)
Analyses - Missing outcome data							
Dealing with missing outcome data in reviews	Mentioned as RoB criterion in methods	68% (34)	92% (23)	44% (11)			
	Contact with authors	40% (20)	76% (19)	4% (1)			
	Sensitivity analyses (rate of missing values)	8% (4)	16% (4)	0% (0)			
	Single imputation	4% (2)	8% (2)	0% (0)			
	Not reported	28% (14)	0% (0)	56% (14)			
Dealing with missing data in individual outcomes	Not reported per outcome	100% (50)	100% (25)	100% (25)	100% (70)	100% (33)	100% (37)
Missing outcome data discussed	In results	56% (28)	80% (20)	32% (8)			
	In results and discussion	8% (4)	16% (4)	0% (0)			
	Not reported	36% (18)	4% (1)	68% (17)			
Missing outcome data discussed for individual outcomes	No	100% (50)	100% (25)	100% (25)	100% (70)	100% (33)	100% (37)
Analyses - Informative censoring							
Dealing with informative censoring in reviews	Mentioned as RoB criterion in methods	2% (1)	4% (1)	0% (0)			
	Not reported	98% (49)	96% (24)	100% (25)			
Dealing with informative censoring for individual outcomes	Not reported per outcome	100% (50)	100% (25)	100% (25)	100% (70)	100% (33)	100% (37)
Informative censoring discussed	Not reported	100% (50)	100% (25)	100% (25)			
Informative censoring discussed for individual outcomes	No	100% (50)	100% (25)	100% (25)	100% (70)	100% (33)	100% (37)
Analyses - Competing events							
Dealing with deaths as competing event in reviews	Not reported	34% (17)	32% (8)	36% (9)			
	No outcomes with potential competing events	66% (33)	68% (17)	64% (16)			
Dealing with deaths as competing events in individual outcomes	Not reported per outcome	46% (23)	48% (12)	44% (11)	37% (26)	39% (13)	35% (13)
	Not applicable	70% (35)	64% (16)	76% (19)	63% (44)	61% (20)	65% (24)
Deaths as competing events discussed	Not reported	34% (17)	32% (8)	36% (9)			
	No outcomes with potential competing events	66% (33)	68% (17)	64% (16)			
Deaths as competing events discussed for individual outcomes	No	46% (23)	48% (12)	44% (11)	37% (26)	39% (13)	35% (13)
	Not applicable	70% (35)	64% (16)	76% (19)	63% (44)	61% (20)	65% (24)
Analyses - Treatment switching							
Dealing with treatment switching in reviews	RoB criterion in methods	2% (1)	4% (1)	0% (0)			
	Presence reported for each trial	2% (1)	4% (1)	0% (0)			
	Sensitivity analysis (e.g. rate of participants), RoB criterion	2% (1)	0% (0)	4% (1)			
	Not reported	94% (47)	92% (23)	96% (24)			
Dealing with treatment switching in individual outcomes	Not reported per outcome	100% (50)	100% (25)	100% (25)	100% (70)	100% (33)	100% (37)
Treatment switching discussed	In results	6% (3)	8% (2)	4% (1)			
	In discussion	4% (2)	8% (2)	0% (0)			

Domain		Review Overall (N = 50)	Review Cochrane (n = 25)	Review Non-Cochrane (n = 25)	Review outcome Overall (N = 70)	Review outcome Cochrane (n = 33)	Review outcome Non-Cochrane (n = 37)
	Not reported	90% (45)	84% (21)	96% (24)			
Treatment switching discussed for individual outcomes	No	100% (50)	100% (25)	100% (25)	100% (70)	100% (33)	100% (37)
Analyses - Proportional hazards							
Proportional hazards assessed in reviews	Not reported	100% (50)	100% (25)	100% (25)			
Proportional hazards assessed in individual outcomes	Not reported per outcome	100% (50)	100% (25)	100% (25)	100% (70)	100% (33)	100% (37)
Dealing with (non-)proportional hazards	Not applicable	100% (50)	100% (25)	100% (25)			
Test for proportionality for individual outcomes	Not applicable	100% (50)	100% (25)	100% (25)	100% (70)	100% (33)	100% (37)
Non-proportionality of hazards indicated	Not applicable	100% (50)	100% (25)	100% (25)	100% (70)	100% (33)	100% (37)
Dealing with (non-)proportional hazards	Not applicable	100% (50)	100% (25)	100% (25)	100% (70)	100% (33)	100% (37)
Results							
Interpretation of pooled HR	Favourable, statistically significant	44% (22)	20% (5)	68% (17)	37% (26)	18% (6)	54% (20)
	Favourable, statistically non-significant	46% (23)	48% (12)	44% (11)	39% (27)	42% (14)	35% (13)
	Unfavourable, statistically significant	4% (2)	4% (1)	4% (1)	3% (2)	3% (1)	3% (1)
	Unfavourable, statistically significant non-significant	26% (13)	44% (11)	8% (2)	20% (14)	33% (11)	8% (3)
	Direction of effect unclear (HR = 1)	4% (2)	4% (1)	4% (1)	1% (1)	3% (1)	0% (0)
HR for events or non events	Event	96% (48)	96% (24)	96% (24)	97% (68)	97% (32)	97% (36)
	Unclear	4% (2)	4% (1)	4% (1)	3% (2)	3% (1)	3% (1)
Interpretation of HR<1	Decreased risk	92% (46)	92% (23)	92% (23)	94% (66)	94% (31)	95% (35)
	Increased risk	4% (2)	4% (1)	4% (1)	3% (2)	3% (1)	3% (1)
	Unclear	4% (2)	4% (1)	4% (1)	3% (2)	3% (1)	3% (1)
Trial HRs inverted	Not reported	100% (50)	100% (25)	100% (25)	100% (70)	100% (33)	100% (37)
Risk of Bias							
Risk of bias tools specified	RoB 1, study level	58% (29)	64% (16)	52% (13)			
	RoB 1, outcome level	18% (9)	36% (9)	0% (0)			
	Other (e.g. CONSORT)	8% (4)	0% (0)	16% (4)			
	Jadad scale	4% (2)	0% (0)	8% (2)			
	RoB 2	4% (2)	0% (0)	8% (2)			
	No RoB assessment	8% (4)	0% (0)	16% (4)			
TTE specific risk of bias criteria used	Yes (e.g. "risk of bias related to censoring")	2% (1)	4% (1)	0% (0)			
	No	90% (45)	96% (24)	84% (21)			
	Not applicable	8% (4)	0% (0)	16% (4)			

Abbreviations: HR = hazard ratio; IPD = individual participant data, IQR = interquartile range; ITT = intention to treat; PP = per protocol; RoB = risk of bias; TTE = time-to-event

Figure A6: Frequency and combinations of available time-to-event summary data items per individual trial time-to-event outcome

	Which individual time-to-event data items (✓) were available for individual trial outcomes?											*Total*
Survival curves	✓	✓	✓	✓	✓	✓	✓	✓				83% (263)
P-value (log-rank)	✓	✓	✓	✓	✓	✓	✓			✓		76% (240)
HR or log(HR)	✓	✓	✓	✓					✓	✓		72% (226)
Time-point specific survival rates (per arm)	✓		✓		✓							46% (145)
Median survival times (per arm)		✓	✓				✓					40% (125)
*Other**											✓	33% (105)
Total	16% (50)	15% (47)	10% (33)	7% (22)	6% (18)	3% (10)	3% (9)	3% (8)	2% (7)	2% (7)	33% (104)	

*Appendix-Figure 1: Frequency (rows) and combinations (colomns) of available time-to-event summary data items per individual time-to-event outcome available in trial publications. Numbers in coloms represent the available individual items (e.g. surival curves), numbers in the bottom rows represent the frequency of available combinations (e.g. survival curves together with a P-value, HR/ log(HR) and time-point specific survival times). * Other included, e.g., median or time-point specific cumulative incidence per arm, observed and expected events, event times per participant and restricted mean survival time (RMST) (Abbreviations: HR = hazard ration).*

Table A7: *Extended time-to-event specific methodological characteristics of trials included in review time-to-event outcome meta-analyses*

		Trial (N = 235)	Trial outcome						Review		
Domain			Overall (N = 315)	Cochrane (n = 131)	Non-Cochrane (n = 184)	All-cause mortality/ Overall survival (n = 198)	Combined, including all-cause mortality (n = 77)	Not including all-cause mortality (n = 40)	Overall (N = 50)	Cochrane (n = 25)	Non-Cochrane (n = 25)
Analyses - Comparative effect measures											
Available TTE data	HR/log(HR), P-value (log-rank), Survival curves, Time-point specific survival rates (per arm)	15% (36)	16% (50)	20% (26)	13% (24)	14% (28)	25% (19)	8% (3)	38% (19)	32% (8)	44% (11)
	HR or log(HR), P-value (log-rank), Survival curves, Median survival times (per arm)	18% (42)	15% (47)	8% (10)	20% (37)	10% (20)	31% (24)	8% (3)	36% (18)	28% (7)	44% (11)
	HR or log(HR), P-value (log-rank), Survival curves, Median survival times (per arm), Time-point specific survival rates (per arm)	12% (29)	10% (33)	4% (5)	15% (28)	13% (25)	10% (8)	0% (0)	28% (14)	16% (4)	40% (10)
	HR or log(HR), P-value (log-rank), Survival curves	9% (21)	7% (22)	7% (9)	7% (13)	9% (18)	5% (4)	0% (0)	30% (15)	24% (6)	36% (9)
	P-value (log-rank), Survival curves, Time-point specific survival rates (per arm)	7% (16)	6% (18)	4% (5)	7% (13)	6% (11)	4% (3)	10% (4)	16% (8)	12% (3)	20% (5)
	P-value (log-rank), Survival curves	4% (10)	3% (10)	2% (3)	4% (7)	4% (7)	0% (0)	8% (3)	10% (5)	8% (2)	12% (3)
	P-value (log-rank), Survival curves, Median survival times (per arm)	4% (9)	3% (9)	2% (2)	4% (7)	4% (8)	0% (0)	3% (1)	10% (5)	8% (2)	12% (3)
	Survival curves	3% (6)	3% (8)	5% (7)	1% (1)	3% (5)	0% (0)	8% (3)	10% (5)	16% (4)	4% (1)
	HR or log(HR)	2% (5)	2% (7)	1% (1)	3% (6)	3% (5)	1% (1)	3% (1)	10% (5)	4% (1)	16% (4)
	HR or log(HR), P-value (log-rank)	3% (6)	2% (7)	4% (5)	1% (2)	2% (4)	3% (2)	3% (1)	12% (6)	16% (4)	8% (2)
	Other	41% (96)	33% (104)	44% (58)	25% (46)	34% (67)	21% (16)	53% (21)	172% (86)	192% (48)	152% (38)
TTE data	Survival curves	83% (195)	83% (263)	76% (100)	89% (163)	85% (168)	90% (69)	65% (26)	92% (46)	84% (21)	100% (25)
	P-value (log-rank)	78% (183)	76% (240)	71% (93)	80% (147)	75% (148)	87% (67)	63% (25)	94% (47)	92% (23)	96% (24)
	HR or log(HR)	71% (166)	72% (226)	64% (84)	77% (142)	68% (135)	95% (73)	45% (18)	90% (45)	80% (20)	100% (25)
	Time-point specific survival rates (per arm)	49% (115)	46% (145)	50% (66)	43% (79)	48% (95)	49% (38)	30% (12)	82% (41)	76% (19)	88% (22)
	Median survival times (per arm)	43% (100)	40% (125)	29% (38)	47% (87)	39% (78)	51% (39)	20% (8)	58% (29)	56% (14)	60% (15)
	Type of test unclear or not reported	8% (18)	6% (20)	7% (9)	6% (11)	6% (12)	5% (4)	10% (4)	26% (13)	20% (5)	32% (8)
	Median cumulative incidence (per arm)	2% (5)	2% (6)	2% (2)	2% (4)	1% (2)	4% (3)	3% (1)	8% (4)	8% (2)	8% (2)
	Mean and standard deviation per arm	2% (4)	1% (4)	2% (3)	1% (1)	2% (4)	0% (0)	0% (0)	8% (4)	12% (3)	4% (1)
	Observed and expected events (log-rank) or hazard rates	2% (4)	1% (4)	3% (4)	0% (0)	2% (4)	0% (0)	0% (0)	4% (2)	8% (2)	0% (0)
	Wilcoxon-Gehan test	1% (3)	1% (3)	2% (3)	0% (0)	2% (3)	0% (0)	0% (0)	6% (3)	12% (3)	0% (0)
	Time-point specific cumulative incidence	1% (2)	1% (2)	1% (1)	1% (1)	1% (2)	0% (0)	5% (2)	6% (3)	8% (2)	4% (1)
	Event times per participant	1% (3)	1% (3)	2% (3)	0% (0)	1% (2)	0% (0)	3% (1)	4% (2)	8% (2)	0% (0)
	Test results not numerically reported	1% (3)	1% (3)	0% (0)	2% (3)	2% (3)	0% (0)	0% (0)	4% (2)	0% (0)	8% (2)
	Cox model coefficients and/or P-values	1% (3)	1% (3)	1% (1)	1% (2)	1% (1)	0% (0)	5% (2)	4% (2)	4% (1)	4% (1)
	Restricted mean survival time (RMST)	1% (2)	1% (2)	1% (1)	1% (1)	1% (2)	0% (0)	0% (0)	4% (2)	4% (1)	4% (1)
	Greys Test	0% (1)	0% (1)	1% (1)	0% (0)	0% (0)	0% (0)	3% (1)	2% (1)	4% (1)	0% (0)
	Absolute risk reduction (Andersen and Altman methodology)	0% (1)	0% (1)	0% (0)	1% (1)	0% (0)	0% (0)	3% (1)	2% (1)	0% (0)	4% (1)

		Trial (N = 235)	Trial outcome						Review		
Domain			Overall (N = 315)	Cochrane (n = 131)	Non-Cochrane (n = 184)	All-cause mortality/ Overall survival (n = 198)	Combined, including all-cause mortality (n = 77)	Not including all-cause mortality (n = 40)	Overall (N = 50)	Cochrane (n = 25)	Non-Cochrane (n = 25)
HR calculation	Cox model	59% (138)	60% (188)	50% (66)	66% (122)	57% (113)	75% (58)	43% (17)	86% (43)	72% (18)	100% (25)
	Log rank	1% (2)	1% (2)	0% (0)	1% (2)	1% (2)	0% (0)	0% (0)	4% (2)	0% (0)	8% (2)
	Cox model and RPSFT model	0% (1)	0% (1)	0% (0)	1% (1)	1% (1)	0% (0)	0% (0)	2% (1)	0% (0)	4% (1)
	Cox model and Cox model with time dependent variable(s)	0% (1)	0% (1)	1% (1)	0% (0)	1% (1)	0% (0)	0% (0)	2% (1)	4% (1)	0% (0)
	Cox Markov model	0% (1)	0% (1)	1% (1)	0% (0)	1% (1)	0% (0)	0% (0)	2% (1)	4% (1)	0% (0)
	Cox model and Fine and Gray model	0% (1)	0% (1)	0% (0)	1% (1)	0% (0)	0% (0)	3% (1)	2% (1)	0% (0)	4% (1)
	Andersen-Gill regression model	0% (1)	0% (1)	1% (1)	0% (0)	1% (1)	0% (0)	0% (0)	2% (1)	4% (1)	0% (0)
	Cox model and Log rank method	0% (1)	0% (1)	0% (0)	1% (1)	0% (0)	1% (1)	0% (0)	2% (1)	0% (0)	4% (1)
	Fine and Gray	0% (1)	0% (1)	1% (1)	0% (0)	0% (0)	0% (0)	3% (1)	2% (1)	4% (1)	0% (0)
	Unclear	0% (1)	1% (2)	0% (0)	1% (2)	1% (1)	1% (1)	0% (0)	2% (1)	0% (0)	4% (1)
	Not reported	12% (29)	11% (34)	13% (17)	9% (17)	10% (19)	18% (14)	3% (1)	40% (20)	40% (10)	40% (10)
	No HR calculated	31% (72)	26% (82)	34% (44)	21% (38)	30% (59)	4% (3)	50% (20)	54% (27)	64% (16)	44% (11)
Survival plots											
Survival plots available	Kaplan-Meier	80% (188)	79% (249)	74% (97)	83% (152)	81% (161)	88% (68)	50% (20)	92% (46)	84% (21)	100% (25)
	Cumulative incidence	3% (6)	3% (8)	2% (2)	3% (6)	1% (2)	1% (1)	13% (5)	10% (5)	8% (2)	12% (3)
	Type of curve not reported	2% (4)	2% (5)	2% (2)	2% (3)	2% (3)	0% (0)	5% (2)	6% (3)	8% (2)	4% (1)
	Adjusted Kaplan-Meier	0% (1)	0% (1)	1% (1)	0% (0)	1% (1)	0% (0)	0% (0)	2% (1)	4% (1)	0% (0)
	No, but for other analyses of this outcome	3% (6)	2% (7)	4% (5)	1% (2)	3% (6)	0% (0)	0% (0)	8% (4)	12% (3)	4% (1)
	No, no graphs were presented	17% (40)	14% (45)	18% (24)	11% (21)	13% (25)	9% (7)	33% (13)	56% (28)	56% (14)	56% (14)
Number of individuals at risk reported	Yes	57% (134)	58% (184)	44% (58)	68% (126)	55% (108)	78% (60)	40% (16)	88% (44)	76% (19)	100% (25)
	No, but for other analyses of this outcome	2% (4)	2% (5)	2% (3)	1% (2)	2% (4)	1% (1)	0% (0)	6% (3)	8% (2)	4% (1)
	No	28% (66)	26% (81)	35% (46)	19% (35)	31% (61)	12% (9)	28% (11)	54% (27)	60% (15)	48% (12)
	Not applicable	17% (40)	14% (45)	18% (24)	11% (21)	13% (25)	9% (7)	33% (13)	56% (28)	56% (14)	56% (14)
Censoring events reported	On survival curve	38% (89)	38% (119)	24% (32)	47% (87)	37% (74)	49% (38)	18% (7)	68% (34)	64% (16)	72% (18)
	On survival curve and reported with individuals at risk	3% (8)	3% (11)	3% (4)	4% (7)	3% (5)	8% (6)	0% (0)	14% (7)	12% (3)	16% (4)
	No	43% (102)	43% (136)	51% (67)	38% (69)	45% (90)	34% (26)	50% (20)	80% (40)	80% (20)	80% (20)
	Not applicable	18% (43)	16% (49)	21% (28)	11% (21)	15% (29)	9% (7)	33% (13)	62% (31)	68% (17)	56% (14)
Censoring balanced	Yes	32% (75)	30% (96)	21% (28)	37% (68)	31% (61)	40% (31)	10% (4)	66% (33)	64% (16)	68% (17)
	No	9% (22)	8% (24)	5% (7)	9% (17)	6% (12)	14% (11)	3% (1)	28% (14)	20% (5)	36% (9)
	Unclear	4% (9)	3% (9)	1% (1)	4% (8)	3% (5)	3% (2)	5% (2)	14% (7)	4% (1)	24% (6)
	Not applicable	61% (143)	59% (186)	73% (95)	49% (91)	61% (121)	43% (33)	80% (32)	88% (44)	92% (23)	84% (21)
TTE data recalculation reported in revies per outcome											
TTE data recalculation	HR together with other infomation (e.g. events)	3% (8)	5% (15)	2% (2)	7% (13)	4% (8)	1% (1)	15% (6)	4% (2)	4% (1)	4% (1)
	P-value together with additional information (e.g. events)	6% (13)	5% (15)	11% (14)	1% (1)	6% (12)	3% (2)	3% (1)	8% (4)	12% (3)	4% (1)
	HR and confidence intervals	3% (8)	3% (10)	3% (4)	3% (6)	2% (4)	8% (6)	0% (0)	6% (3)	8% (2)	4% (1)

Domain		Trial (N = 235)	Trial outcome						Review		
			Overall (N = 315)	Cochrane (n = 131)	Non-Cochrane (n = 184)	All-cause mortality/ Overall survival (n = 198)	Combined, including all-cause mortality (n = 77)	Not including all-cause mortality (n = 40)	Overall (N = 50)	Cochrane (n = 25)	Non-Cochrane (n = 25)
	IPD (recalculated or from publication)	3% (6)	2% (6)	2% (3)	2% (3)	3% (5)	0% (0)	3% (1)	6% (3)	8% (2)	4% (1)
	Survival curves	2% (5)	2% (5)	2% (2)	2% (3)	2% (4)	0% (0)	3% (1)	8% (4)	8% (2)	8% (2)
	Only specified to be recalculated or obtained from authors	1% (2)	1% (2)	2% (2)	0% (0)	1% (2)	0% (0)	0% (0)	2% (1)	4% (1)	0% (0)
	Time-point specific survival times	0% (1)	1% (2)	2% (2)	0% (0)	1% (1)	0% (0)	3% (1)	2% (1)	4% (1)	0% (0)
	Not specified for this trial outcome	83% (196)	83% (260)	78% (102)	86% (158)	82% (162)	88% (68)	75% (30)	86% (43)	76% (19)	96% (24)

Abbreviations: HR = hazard ratio; IPD = individual participant data; RPSFT = Rank Preserving Structural Failure Time; TTE = time-to-event

Table A8: *General methodological characteristics of trials included in review time-to-event outcome meta-analyses*

Domain		Trial (N = 235)	Trial outcome						Review		
			Overall (N = 315)	Cochrane (n = 131)	Non-Cochrane (n = 184)	All-cause mortality/ Overall survival (n = 198)	Combined, including all-cause mortality (n = 77)	Not including all-cause mortality (n = 40)	Overall (N = 50)	Cochrane (n = 25)	Non-Cochrane (n = 25)
Analyses available in trial publication											
Available types of analyses	ITT	63% (147)	63% (198)	58% (76)	66% (122)	61% (121)	73% (56)	53% (21)	92% (46)	92% (23)	92% (23)
	ITT, Per-protocol	6% (14)	5% (15)	8% (10)	3% (5)	6% (11)	4% (3)	3% (1)	26% (13)	32% (8)	20% (5)
	Modified ITT	4% (9)	4% (12)	1% (1)	6% (11)	4% (8)	5% (4)	0% (0)	14% (7)	4% (1)	24% (6)
	Per-protocol	2% (5)	2% (6)	3% (4)	1% (2)	2% (4)	0% (0)	5% (2)	10% (5)	12% (3)	8% (2)
	ITT, As treated	2% (4)	2% (5)	0% (0)	3% (5)	2% (3)	1% (1)	3% (1)	8% (4)	0% (0)	16% (4)
	Modified ITT, Per-protocol	1% (2)	1% (3)	1% (1)	1% (2)	1% (1)	3% (2)	0% (0)	4% (2)	4% (1)	4% (1)
	ITT, Modified ITT	0% (1)	0% (1)	0% (0)	1% (1)	0% (0)	0% (0)	3% (1)	2% (1)	0% (0)	4% (1)
	As treated	0% (1)	0% (1)	0% (0)	1% (1)	0% (0)	0% (0)	3% (1)	2% (1)	0% (0)	4% (1)
	ITT, Per-protocol, As treated	0% (1)	0% (1)	0% (0)	1% (1)	0% (0)	1% (1)	0% (0)	2% (1)	0% (0)	4% (1)
	Unclear	0% (1)	0% (1)	1% (1)	0% (0)	1% (1)	0% (0)	0% (0)	2% (1)	4% (1)	0% (0)
	Not reported	26% (60)	23% (72)	29% (38)	18% (34)	25% (49)	13% (10)	33% (13)	58% (29)	68% (17)	48% (12)
ITT analysis available	ITT	69% (161)	70% (220)	66% (86)	73% (134)	68% (135)	79% (61)	60% (24)	96% (48)	96% (24)	96% (24)
	Modified ITT	5% (11)	5% (15)	2% (2)	7% (13)	5% (9)	8% (6)	0% (0)	18% (9)	8% (2)	28% (7)
	No	3% (6)	2% (7)	3% (4)	2% (3)	2% (4)	0% (0)	8% (3)	12% (6)	12% (3)	12% (3)
	Unclear	0% (1)	0% (1)	1% (1)	0% (0)	1% (1)	0% (0)	0% (0)	2% (1)	4% (1)	0% (0)
	Not reported	26% (60)	23% (72)	29% (38)	18% (34)	25% (49)	13% (10)	33% (13)	58% (29)	68% (17)	48% (12)
Available ITT analysis in complete population	Yes	43% (102)	44% (139)	34% (44)	52% (95)	43% (85)	49% (38)	40% (16)	92% (46)	88% (22)	96% (24)
	No	19% (45)	20% (62)	30% (39)	13% (23)	20% (39)	22% (17)	15% (6)	46% (23)	56% (14)	36% (9)
	Unclear	2% (4)	2% (5)	0% ()	3% (5)	1% (2)	3% (2)	3% (1)	8% (4)	0% ()	16% (4)
	Not applicable (no ITT, only mITT or subgroup))	38% (89)	35% (109)	37% (48)	33% (61)	36% (72)	26% (20)	43% (17)	78% (39)	80% (20)	76% (19)
Analyses included in review meta-analyses											
Type of analysis included	ITT	67% (158)	69% (216)	65% (85)	71% (131)	67% (133)	79% (61)	55% (22)	96% (48)	96% (24)	96% (24)
	mITT	5% (11)	5% (16)	2% (2)	8% (14)	5% (9)	6% (5)	5% (2)	18% (9)	8% (2)	28% (7)
	Per protocol	2% (5)	2% (7)	3% (4)	2% (3)	3% (5)	1% (1)	3% (1)	10% (5)	12% (3)	8% (2)
	As treated	1% (3)	0% (1)	0% (0)	1% (1)	0% (0)	0% (0)	3% (1)	6% (3)	8% (2)	4% (1)
	Unclear	0% (1)	1% (3)	2% (2)	1% (1)	1% (2)	0% (0)	3% (1)	2% (1)	0% (0)	4% (1)
	Not reported	26% (60)	23% (72)	29% (38)	18% (34)	25% (49)	13% (10)	33% (13)	58% (29)	68% (17)	48% (12)
Included analysis in complete population	Yes	55% (130)	55% (174)	45% (59)	63% (115)	56% (110)	55% (42)	55% (22)	96% (48)	96% (24)	96% (24)
	No	32% (75)	32% (100)	44% (57)	23% (43)	31% (62)	31% (24)	35% (14)	70% (35)	80% (20)	60% (15)
	Unclear	4% (9)	3% (10)	3% (4)	3% (6)	3% (5)	3% (2)	8% (3)	16% (8)	12% (3)	20% (5)
	Not reported	6% (15)	5% (17)	6% (8)	5% (9)	6% (12)	6% (5)	0% (0)	20% (10)	16% (4)	24% (6)
	Not applicable (e.g. subgroup included in review)	5% (12)	4% (14)	2% (3)	6% (11)	5% (9)	5% (4)	3% (1)	14% (7)	8% (2)	20% (5)

		Trial (N = 235)	Trial outcome						Review		
Domain			**Overall (N = 315)**	**Cochrane (n = 131)**	**Non-Cochrane (n = 184)**	**All-cause mortality/ Overall survival (n = 198)**	**Combined, including all-cause mortality (n = 77)**	**Not including all-cause mortality (n = 40)**	**Overall (N = 50)**	**Cochrane (n = 25)**	**Non-Cochrane (n = 25)**
Included analysis in allocated arm	Yes	87% (204)	88% (276)	79% (103)	94% (173)	88% (175)	91% (70)	78% (31)	98% (49)	96% (24)	100% (25)
	No	0% (1)	0% (1)	0% (0)	1% (1)	0% (0)	0% (0)	3% (1)	2% (1)	0% (0)	4% (1)
	Unclear	2% (5)	2% (5)	3% (4)	1% (1)	1% (2)	1% (1)	5% (2)	10% (5)	16% (4)	4% (1)
	Not reported	12% (28)	10% (33)	18% (24)	5% (9)	11% (21)	8% (6)	15% (6)	40% (20)	56% (14)	24% (6)
Covariate adjustment of available estimates in trial publication											
Available unadjusted and adjusted analyses	Adjusted, baseline characteristics	13% (30)	12% (39)	11% (15)	13% (24)	12% (24)	13% (10)	13% (5)	36% (18)	24% (6)	48% (12)
	Unadjusted; Adjusted, baseline characteristics	11% (26)	9% (29)	11% (14)	8% (15)	9% (17)	13% (10)	5% (2)	36% (18)	40% (10)	32% (8)
	Stratified, baseline characteristics	6% (15)	7% (23)	6% (8)	8% (15)	6% (12)	9% (7)	10% (4)	22% (11)	16% (4)	28% (7)
	Stratified, randomization stratification factors	5% (11)	5% (16)	2% (3)	7% (13)	6% (11)	6% (5)	0% (0)	14% (7)	8% (2)	20% (5)
	Unadjusted (univariate including treatment variables only)	5% (12)	5% (15)	5% (7)	4% (8)	5% (9)	1% (1)	13% (5)	22% (11)	16% (4)	28% (7)
	Stratified, factors unclear	3% (6)	3% (8)	1% (1)	4% (7)	2% (4)	5% (4)	0% (0)	10% (5)	4% (1)	16% (4)
	Unadjusted; Stratified, factors unclear	2% (5)	2% (7)	0% (0)	4% (7)	3% (5)	3% (2)	0% (0)	4% (2)	0% (0)	8% (2)
	Adjusted, factors unclear	3% (6)	2% (7)	4% (5)	1% (2)	3% (5)	1% (1)	3% (1)	10% (5)	16% (4)	4% (1)
	Adjusted, baseline characteristics; Not reported	1% (3)	2% (6)	3% (4)	1% (2)	2% (3)	4% (3)	0% (0)	4% (2)	4% (1)	4% (1)
	Unadjusted; Stratified, baseline characteristics	1% (2)	1% (2)	1% (1)	1% (1)	1% (2)	0% (0)	0% (0)	4% (2)	4% (1)	4% (1)
	Adjusted, factors unclear; Stratified, factors unclear	0% (1)	1% (2)	0% (0)	1% (2)	1% (1)	1% (1)	0% (0)	2% (1)	0% (0)	4% (1)
	Adjusted, baseline characteristics; Stratified, factors unclear	0% (1)	0% (1)	1% (1)	0% (0)	0% (0)	1% (1)	0% (0)	2% (1)	4% (1)	0% (0)
	Unadjusted; Stratified, randomization stratification factors	0% (1)	0% (1)	0% (0)	1% (1)	1% (1)	0% (0)	0% (0)	2% (1)	0% (0)	4% (1)
	Adjusted, factors unclear; Stratified, baseline characteristics	0% (1)	0% (1)	0% (0)	1% (1)	0% (0)	1% (1)	0% (0)	2% (1)	0% (0)	4% (1)
	Adjusted, baseline characteristics; Stratified, baseline characteristics	0% (1)	0% (1)	0% (0)	1% (1)	0% (0)	1% (1)	0% (0)	2% (1)	0% (0)	4% (1)
	Unclear	3% (8)	3% (11)	2% (3)	4% (8)	4% (7)	4% (3)	3% (1)	14% (7)	8% (2)	20% (5)
	Not reported	20% (48)	19% (59)	18% (24)	19% (35)	18% (35)	30% (23)	3% (1)	54% (27)	40% (10)	68% (17)
	Not applicable (no HR directly reported)	31% (73)	28% (87)	34% (45)	23% (42)	31% (62)	5% (4)	53% (21)	54% (27)	64% (16)	44% (11)
Unadjusted HR reported	Unadjusted (univariate including treatment variables only)	19% (44)	17% (54)	17% (22)	17% (32)	17% (34)	17% (13)	18% (7)	54% (27)	48% (12)	60% (15)
	No	51% (120)	52% (163)	47% (61)	55% (102)	48% (95)	74% (57)	28% (11)	78% (39)	56% (14)	100% (25)
	Unclear	3% (8)	3% (11)	2% (3)	4% (8)	4% (7)	4% (3)	3% (1)	14% (7)	8% (2)	20% (5)
	Not applicable (no HR directly reported)	31% (73)	28% (87)	34% (45)	23% (42)	31% (62)	5% (4)	53% (21)	54% (27)	64% (16)	44% (11)
Adjusted HR reported	Adjusted, baseline characteristics	26% (60)	24% (75)	26% (34)	22% (41)	22% (43)	32% (25)	18% (7)	58% (29)	52% (13)	64% (16)
	Adjusted, factors unclear	3% (8)	3% (10)	4% (5)	3% (5)	3% (6)	4% (3)	3% (1)	14% (7)	16% (4)	12% (3)
	No	42% (98)	42% (132)	34% (44)	48% (88)	40% (80)	55% (42)	25% (10)	70% (35)	52% (13)	88% (22)
	Unclear	3% (8)	3% (11)	2% (3)	4% (8)	4% (7)	4% (3)	3% (1)	14% (7)	8% (2)	20% (5)
	Not applicable (no HR directly reported)	31% (73)	28% (87)	34% (45)	23% (42)	31% (62)	5% (4)	53% (21)	54% (27)	64% (16)	44% (11)
Stratified HR	Stratified, baseline characteristics	8% (19)	9% (27)	7% (9)	10% (18)	7% (14)	12% (9)	10% (4)	24% (12)	16% (4)	32% (8)

Domain		Trial (N = 235)	Trial outcome						Review		
			Overall (N = 315)	Cochrane (n = 131)	Non-Cochrane (n = 184)	All-cause mortality/ Overall survival (n = 198)	Combined, including all-cause mortality (n = 77)	Not including all-cause mortality (n = 40)	Overall (N = 50)	Cochrane (n = 25)	Non-Cochrane (n = 25)
reported	Stratified, factors unclear	5% (11)	6% (18)	2% (2)	9% (16)	5% (10)	10% (8)	0% (0)	14% (7)	4% (1)	24% (6)
	Stratified, randomization stratification factors	5% (12)	5% (17)	2% (3)	8% (14)	6% (12)	6% (5)	0% (0)	14% (7)	8% (2)	20% (5)
	No	51% (119)	49% (155)	53% (69)	47% (86)	47% (93)	62% (48)	35% (14)	84% (42)	80% (20)	88% (22)
	Unclear	3% (8)	3% (11)	2% (3)	4% (8)	4% (7)	4% (3)	3% (1)	14% (7)	8% (2)	20% (5)
	Not applicable (no HR directly reported)	31% (73)	28% (87)	34% (45)	23% (42)	31% (62)	5% (4)	53% (21)	54% (27)	64% (16)	44% (11)
Available unadjusted log-rank P-values	Stratified, baseline characteristics	7% (16)	8% (24)	5% (6)	10% (18)	5% (9)	19% (15)	0% (0)	14% (7)	8% (2)	20% (5)
	Unadjusted (univariate including treatment variables only)	7% (17)	6% (19)	5% (7)	7% (12)	7% (14)	4% (3)	5% (2)	26% (13)	20% (5)	32% (8)
	Stratified, randomization stratification factors	5% (11)	5% (16)	4% (5)	6% (11)	6% (11)	6% (5)	0% (0)	12% (6)	12% (3)	12% (3)
	Stratified, factors unclear	5% (11)	4% (14)	0% (0)	8% (14)	5% (9)	6% (5)	0% (0)	18% (9)	0% (0)	36% (9)
	Unadjusted; Adjusted, baseline characteristics	3% (8)	3% (10)	5% (6)	2% (4)	3% (6)	4% (3)	3% (1)	14% (7)	16% (4)	12% (3)
	Adjusted, baseline characteristics	3% (7)	3% (8)	0% (0)	4% (8)	4% (7)	1% (1)	0% (0)	14% (7)	0% (0)	28% (7)
	Adjusted, baseline characteristics; Stratified, randomization stratification factors	0% (1)	1% (2)	0% (0)	1% (2)	1% (1)	1% (1)	0% (0)	2% (1)	0% (0)	4% (1)
	Adjusted, baseline characteristics; Not reported	0% (1)	1% (2)	0% (0)	1% (2)	1% (1)	1% (1)	0% (0)	2% (1)	0% (0)	4% (1)
	Adjusted, factors unclear; Stratified, factors unclear	0% (1)	1% (2)	0% (0)	1% (2)	1% (1)	1% (1)	0% (0)	2% (1)	0% (0)	4% (1)
	Unadjusted; Stratified, baseline characteristics	0% (1)	0% (1)	0% (0)	1% (1)	0% (0)	1% (1)	0% (0)	32% (16)	60% (15)	4% (1)
	Not reported	49% (115)	47% (148)	53% (69)	43% (79)	48% (96)	40% (31)	53% (21)	70% (35)	80% (20)	60% (15)
	Not applicable (no log-rank P-value directly reported)	23% (53)	22% (69)	29% (38)	17% (31)	22% (43)	13% (10)	40% (16)	38% (19)	0% (0)	76% (19)
Unadjusted log-rank P-value reported	Unadjusted (univariate including treatment variables only)	11% (26)	10% (30)	10% (13)	9% (17)	10% (20)	9% (7)	8% (3)	38% (19)	32% (8)	44% (11)
	No	69% (163)	69% (216)	61% (80)	74% (136)	68% (135)	78% (60)	53% (21)	88% (44)	80% (20)	96% (24)
	Not applicable (no log-rank P-value directly reported)	23% (53)	22% (69)	29% (38)	17% (31)	22% (43)	13% (10)	40% (16)	60% (30)	60% (15)	60% (15)
Adjusted log-rank P-value reported	Adjusted, baseline characteristics	7% (17)	7% (22)	5% (6)	9% (16)	8% (15)	8% (6)	3% (1)	26% (13)	16% (4)	36% (9)
	Adjusted, but factors unclear	0% (1)	1% (2)	0% (0)	1% (2)	1% (1)	1% (1)	0% (0)	2% (1)	0% (0)	4% (1)
	No	72% (170)	70% (222)	66% (87)	73% (135)	70% (139)	78% (60)	58% (23)	90% (45)	88% (22)	92% (23)
	Not applicable (no log-rank P-value directly reported)	23% (53)	22% (69)	29% (38)	17% (31)	22% (43)	13% (10)	40% (16)	58% (29)	60% (15)	56% (14)
Stratified log-rank P-value	Stratified, baseline characteristics	7% (17)	8% (25)	5% (6)	10% (19)	5% (9)	21% (16)	0% (0)	16% (8)	8% (2)	24% (6)
	Stratified, randomization stratification factors	5% (12)	6% (18)	4% (5)	7% (13)	6% (12)	8% (6)	0% (0)	14% (7)	12% (3)	16% (4)
	Stratified, but factors unclear	5% (12)	5% (16)	0% (0)	9% (16)	5% (10)	8% (6)	0% (0)	20% (10)	0% (0)	40% (10)
	No	63% (148)	59% (187)	63% (82)	57% (105)	63% (124)	51% (39)	60% (24)	90% (45)	92% (23)	88% (22)
	Not applicable (no log-rank P-value directly reported)	23% (53)	22% (69)	29% (38)	17% (31)	22% (43)	13% (10)	40% (16)	60% (30)	60% (15)	60% (15)
Covariate adjustment of estimate included in review meta-analysis											
Included analysis unadjusted, adjusted, stratified	Unadjusted	28% (66)	25% (80)	36% (47)	18% (33)	28% (56)	16% (12)	30% (12)	60% (30)	68% (17)	52% (13)
	Stratified	15% (36)	18% (56)	8% (11)	24% (45)	16% (31)	27% (21)	10% (4)	34% (17)	20% (5)	48% (12)
	Adjusted	14% (32)	13% (41)	8% (11)	16% (30)	12% (24)	16% (12)	13% (5)	44% (22)	28% (7)	60% (15)
	Unclear	7% (17)	6% (20)	7% (9)	6% (11)	7% (14)	4% (3)	8% (3)	26% (13)	24% (6)	28% (7)

Domain		Trial (N = 235)	Trial outcome						Review		
			Overall (N = 315)	**Cochrane (n = 131)**	**Non-Cochrane (n = 184)**	**All-cause mortality/ Overall survival (n = 198)**	**Combined, including all-cause mortality (n = 77)**	**Not including all-cause mortality (n = 40)**	**Overall (N = 50)**	**Cochrane (n = 25)**	**Non-Cochrane (n = 25)**
	Not reported	40% (95)	37% (118)	40% (53)	35% (65)	37% (73)	38% (29)	40% (16)	74% (37)	68% (17)	80% (20)

Abbreviations: HR = hazard ratio; ITT = intention to treat; TTE = time-to-event

Table A9: *Outcome results of trials included in review time-to-event outcome meta-analyses.*

Domain		Trial outcome				Review		
		Overall (N = 315)	**All-cause mortality/ Overall survival (n = 198)**	**Combined, including all-cause mortality (n = 77)**	**Not including all-cause mortality (n = 40)**	**Overall (N = 50)**	**Cochrane (n = 25)**	**Non-Cochrane (n = 25)**
Relative effect measures included in review according to point estimate and confidence interval								
*Effect in review**	Favourable, statistically significant	26% (81)	21% (42)	35% (27)	30% (12)	54% (27)	40% (10)	68% (17)
	Favourable, statistically non-sign.	51% (161)	57% (113)	49% (38)	25% (10)	78% (39)	68% (17)	88% (22)
	Unfavourable, statisitically significant	3% (10)	2% (4)	0% (0)	15% (6)	8% (4)	8% (2)	8% (2)
	Unfavourable, statisitically non-sign.	19% (60)	19% (38)	13% (10)	30% (12)	58% (29)	64% (16)	52% (13)
	Direction of effect unclear (HR = 1)	1% (3)	1% (1)	3% (2)	0% (0)	6% (3)	4% (1)	8% (2)
Relative effect measures from trial included in review								
*Effect of trial HR**	Favourable, statistically significant	18% (57)	15% (30)	29% (22)	13% (5)	44% (22)	24% (6)	64% (16)
	Favourable, statistically non-sign.	29% (91)	28% (55)	40% (31)	13% (5)	68% (34)	52% (13)	84% (21)
	Unfavourable, statistically significant	3% (8)	1% (1)	4% (3)	10% (4)	10% (5)	4% (1)	16% (4)
	Unfavourable, statistically non-sign.	12% (39)	14% (28)	12% (9)	5% (2)	46% (23)	44% (11)	48% (12)
	Confidence level unclear	3% (11)	3% (6)	5% (4)	3% (1)	18% (9)	16% (4)	20% (5)
	Direction of effect unclear	0% (1)	0% (0)	1% (1)	0% (0)	2% (1)	0% (0)	4% (1)
	Not applicable	34% (108)	39% (78)	9% (7)	58% (23)	64% (32)	76% (19)	52% (13)
HR <1 in experimental arm #	Decreased risk	57% (179)	54% (106)	77% (59)	35% (14)	84% (42)	68% (17)	100% (25)
	Increased risk	5% (16)	3% (6)	9% (7)	8% (3)	20% (10)	24% (6)	16% (4)
	Unclear	4% (13)	5% (9)	5% (4)	0% (0)	18% (9)	20% (5)	16% (4)
	Not applicable (e.g. no HR)	34% (107)	39% (77)	9% (7)	58% (23)	62% (31)	76% (19)	48% (12)
HR for events	Event	60% (190)	56% (110)	82% (63)	43% (17)	86% (43)	72% (18)	100% (25)
	Absence of event	2% (5)	1% (2)	4% (3)	0% (0)	6% (3)	4% (1)	8% (2)
	Unclear	4% (13)	5% (9)	5% (4)	0% (0)	18% (9)	20% (5)	16% (4)
	Not applicable (e.g. no HR)	34% (107)	39% (77)	9% (7)	58% (23)	62% (31)	76% (19)	48% (12)
HR directly available	Trial HR directly available	51% (162)	48% (96)	71% (55)	28% (11)	80% (40)	60% (15)	100% (25)
	Trial HR inverted	7% (23)	6% (11)	10% (8)	10% (4)	30% (15)	40% (10)	20% (5)
	Other $	10% (33)	11% (22)	12% (9)	5% (2)	30% (15)	40% (10)	20% (5)
	Not applicable (e.g. no HR calculated)	31% (97)	35% (69)	6% (5)	58% (23)	56% (28)	68% (17)	44% (11)

Abbreviations: CI = Confidence interval; HR = Hazard ratio; IQR = Interquartile range

**Favourable/ unfavourable corresponds to review intervention indicated, for example, in Summary of Findings tables or forest plots. The direction is based on the point estimate. #Decreased/increased risk of an event as HR < 1 is based on the review intervention and the review authors interpretation of the effect, i.e. when a HR<1 for overall survival that was interpreted as "beneficial for intervention", that HR clearly represented a decreased the risk of the event (death), irrespective of whether review authors named it as absence of the event (overall survival). $Other includes,* *e.g., difference between trial HR/ CI and HR/ CI in forest plot, unclear or different confidence levels (e.g. 99%, 80% or 97.5% CIs) or explicit reporting by review authors not to have included a given trial HR.*

Table A10: *Extended specific trial characteristics with relevance for time-to-event outcomes in trials included in review time-to-event outcome meta-analyses*

		Trial			Trial outcome	Review		
Domain		**Overall (N = 235)**	**Cochrane (n = 102)**	**Non-Cochrane (n = 133)**	**(N = 315)**	**Overall (N = 50)**	**Cochrane (n = 25)**	**Non-Cochrane (n = 25)**
Follow-up								
Follow-up measures available	Follow-up reported for trial	79% (185)	72% (73)	84% (112)		96% (48)	92% (23)	100% (25)
	Follow-up reported for outcomes	3% (6)	4% (4)	2% (2)		10% (5)	12% (3)	8% (2)
	No indicator of follow-up reported	19% (44)	25% (25)	14% (19)		44% (22)	56% (14)	32% (8)
Follow-up measures	Median	66% (154)	57% (58)	72% (96)		92% (46)	84% (21)	100% (25)
	Minimum follow-up	25% (59)	23% (23)	27% (36)		56% (28)	44% (11)	68% (17)
	Maximum follow-up	23% (53)	19% (19)	26% (34)		54% (27)	40% (10)	68% (17)
	IQR/ lower and upper range of IQR	18% (43)	11% (11)	24% (32)		56% (28)	40% (10)	72% (18)
	Mean	3% (8)	3% (3)	4% (5)		12% (6)	8% (2)	16% (4)
	Fixed time-point of outcome measurement only	3% (8)	5% (5)	2% (3)		12% (6)	16% (4)	8% (2)
	Standard deviation	2% (5)	1% (1)	3% (4)		6% (3)	4% (1)	8% (2)
	95% CI of median	1% (2)	1% (1)	1% (1)		4% (2)	4% (1)	4% (1)
	Follow-up reported per outcome	3% (6)	4% (4)	2% (2)		10% (5)	12% (3)	8% (2)
Follow-up calculation	Median follow-up, surviving patients only	8% (19)	9% (9)	8% (10)		26% (13)	32% (8)	20% (5)
	Median follow-up, all patients	5% (11)	5% (5)	5% (6)		16% (8)	16% (4)	16% (4)
	Reverse Kaplan-Meier	3% (7)	4% (4)	2% (3)		14% (7)	16% (4)	12% (3)
	Median follow-up, multiple (e.g. all patients and surviving only)	1% (2)	2% (2)	0% (0)		4% (2)	8% (2)	0% ()
	Median follow-up, excluding censored	1% (2)	1% (1)	1% (1)		4% (2)	4% (1)	4% (1)
	Mean follow-up, multiple (e.g. all patients and surviving only)	0% (1)	1% (1)	0% (0)		2% (1)	4% (1)	0% ()
	Unclear	1% (3)	1% (1)	2% (2)		6% (3)	4% (1)	8% (2)
	Not reported	61% (143)	50% (51)	69% (92)		86% (43)	72% (18)	100% (25)
	Not applicable	20% (47)	27% (28)	14% (19)		46% (23)	56% (14)	36% (9)
Overall follow-up measure reported	Yes	59% (138)	46% (47)	68% (91)		84% (42)	72% (18)	96% (24)
Median overall follow-up	Median (IQR)	45 (22.8 - 67.6)	62.3 (44.5 - 98)	31.44 (15 - 48)				
	Mean (range)	52.87 (5 - 229.2)	75.68 (5 - 167)	38.94 (5.1 - 229.2)				
	Not reported/ unclear	47% (111)	54% (55)	42% (56)				
Analyses - Missing outcome data handling in included trials								
Missing outcome data handling	Excluded from analysis	18% (42)	25% (26)	12% (16)	17% (52)	42% (21)	48% (12)	36% (9)
	Censored	11% (26)	13% (13)	10% (13)	11% (36)	34% (17)	36% (9)	32% (8)
	Complete follow-up/ no LTFU reported at trial	11% (25)	6% (6)	14% (19)	10% (31)	28% (14)	24% (6)	32% (8)

Domain		Trial			Trial outcome	Review		
		Overall (N = 235)	Cochrane (n = 102)	Non-Cochrane (n = 133)	(N = 315)	Overall (N = 50)	Cochrane (n = 25)	Non-Cochrane (n = 25)
	level							
	Single imputation	0% (1)	1% (1)	0% (0)	0% (1)	2% (1)	4% (1)	0% (0)
	Multiple imputation	0% (1)	0% (0)	1% (1)	0% (1)	2% (1)	0% (0)	4% (1)
	Unclear	1% (3)	3% (3)	0% (0)	1% (4)	2% (1)	4% (1)	0% (0)
	Not reported	58% (136)	52% (53)	62% (83)	57% (178)	92% (46)	84% (21)	100% (25)
	No missing data	3% (8)	4% (4)	3% (4)	4% (12)	12% (6)	12% (3)	12% (3)
Reported missing outcome data								
Missing outcome data reported	Yes	46% (108)	48% (49)	44% (59)		88% (44)	88% (22)	88% (22)
	Explicitly reported complete follow-up	6% (14)	4% (4)	8% (10)		22% (11)	16% (4)	28% (7)
	Explicitly reported no LTFU	5% (12)	3% (3)	7% (9)		14% (7)	12% (3)	16% (4)
	No	36% (84)	36% (37)	35% (47)		78% (39)	72% (18)	84% (21)
	Reported across arms only	5% (11)	9% (9)	2% (2)		18% (9)	28% (7)	8% (2)
	Reported for individual outcomes	3% (6)	0% (0)	5% (6)		10% (5)	0% (0)	20% (5)
Total missing outcome data in experimental arm	0	2% (4)	4% (4)	0% (0)		8% (4)	16% (4)	0% (0)
	<5%	26% (60)	23% (23)	28% (37)		64% (32)	48% (12)	80% (20)
	≥5%, <10%	6% (14)	5% (5)	7% (9)		26% (13)	20% (5)	32% (8)
	≥10%, <20%	6% (14)	9% (9)	4% (5)		24% (12)	32% (8)	16% (4)
	≥20%	6% (14)	7% (7)	5% (7)		22% (11)	24% (6)	20% (5)
	Not reported	1% (1)	2% (1)	0% (0)		2% (1)	4% (1)	0% (0)
	Not applicable	19% (25)	11% (6)	24% (19)		28% (14)	24% (6)	32% (8)
	Number randomly allocated not unclear	1% (2)	2% (1)	1% (1)		4% (2)	4% (1)	4% (1)
Total missing outcome data in control arm	0	4% (10)	4% (4)	5% (6)		12% (6)	8% (2)	16% (4)
	<5%	20% (47)	24% (24)	17% (23)		50% (25)	52% (13)	48% (12)
	≥5%, <10%	11% (25)	7% (7)	14% (18)		34% (17)	24% (6)	44% (11)
	≥10%, <20%	7% (16)	8% (8)	6% (8)		28% (14)	28% (7)	28% (7)
	≥20%	3% (7)	4% (4)	2% (3)		10% (5)	12% (3)	8% (2)
	Not reported	15% (20)	2% (1)	24% (19)		2% (1)	4% (1)	0% (0)
	Not applicable	19% (25)	11% (6)	24% (19)		28% (14)	24% (6)	32% (8)
	Number randomized not reported/ unclear	1% (2)	2% (1)	1% (1)		4% (2)	4% (1)	4% (1)
Outcome specific missing outcome data reported	Yes	4% (9)	1% (1)	6% (8)		12% (6)	4% (1)	20% (5)
	Complete follow-up/ no LTFU at trial level	11% (26)	7% (7)	14% (19)		28% (14)	24% (6)	32% (8)
	Complete follow-up/ no LTFU on trial outcome level	3% (8)	4% (4)	3% (4)		12% (6)	12% (3)	12% (3)
	No	84% (198)	90% (92)	80% (106)		100% (50)	100% (25)	100% (25)
Analysis - Censoring								
Advanced methods for censoring in trials	Sensitivity analysis (results not shown)	0% (1)	1% (1)	0% (0)		2% (1)	4% (1)	0% (0)
	No	100% (234)	99% (101)	100% (133)		100% (50)	100% (25)	100% (25)
Analyses - (Death as) competing event								

		Trial			Trial outcome	Review		
Domain		**Overall (N = 235)**	**Cochrane (n = 102)**	**Non-Cochrane (n = 133)**	**(N = 315)**	**Overall (N = 50)**	**Cochrane (n = 25)**	**Non-Cochrane (n = 25)**
Advanced methods for (death as) competing in trials	Cumulative incidence curves	2% (5)	1% (1)	3% (4)		10% (5)	4% (1)	16% (4)
	Fine and Gray and cumulative incidence curves	1% (2)	1% (1)	1% (1)		4% (2)	4% (1)	4% (1)
	No	25% (59)	25% (26)	25% (33)		54% (27)	52% (13)	56% (14)
	Not applicable	86% (202)	85% (87)	86% (115)		92% (46)	88% (22)	96% (24)
Number of (deaths) as competing event in experimental arm	<5%	1% (3)	0% (0)	2% (3)		4% (2)	0% (0)	8% (2)
	≥5%, <10%	1% (2)	0% (0)	2% (2)		2% (1)	0% (0)	4% (1)
	≥10%, <20%	0% (1)	0% (1)	0% (0)		2% (1)	4% (1)	0% (0)
	≥20%, <30%	0% (1)	1% (1)	0% (0)		0% (0)	0% (0)	0% (0)
	≥40%, <50%	0% (1)	1% (1)	0% (0)		2% (1)	4% (1)	0% (0)
	Unclear	3% (7)	2% (2)	4% (5)		8% (4)	8% (2)	8% (2)
	Not reported	11% (25)	13% (13)	9% (12)		20% (10)	20% (5)	20% (5)
	Not applicable	93% (218)	90% (92)	95% (126)		94% (47)	92% (23)	96% (24)
	Number randomized not reported/ unclear	0% (1)	0% (0)	1% (1)		2% (1)	0% (0)	4% (1)
Number of (deaths) as competing event in comparator arm	<5%	1% (3)	0% (0)	2% (3)		4% (2)	0% (0)	8% (2)
	≥5%, <10%	1% (3)	1% (1)	2% (2)		4% (2)	4% (1)	4% (1)
	≥30%, <40%	0% (1)	1% (1)	0% (0)		2% (1)	4% (1)	0% (0)
	≥40%, <50%	0% (0)	0% (0)	0% (0)		0% (0)	0% (0)	0% (0)
	Unclear	3% (7)	2% (2)	4% (5)		8% (4)	8% (2)	8% (2)
	Not reported	11% (25)	13% (13)	9% (12)		20% (10)	20% (5)	20% (5)
	Not applicable	93% (218)	90% (92)	95% (126)		94% (47)	92% (23)	96% (24)
	Number randomized not reported/ unclear	0% (1)	0% (0)	1% (1)		2% (1)	0% (0)	4% (1)
Analyses - Treatment switching								
Advanced methods for treatment switching in trials	Rank preserving structural failure time	0% (1)	0% (0)	1% (1)		2% (1)	0% (0)	4% (1)
	Sensitivity analysis (Cross-over as event)	0% (1)	0% (0)	1% (1)		2% (1)	0% (0)	4% (1)
	Sensitivity analysis (Excluding cross-overs)	0% (1)	0% (0)	1% (1)		2% (1)	0% (0)	4% (1)
	No	93% (219)	97% (99)	90% (120)		438% (219)	396% (99)	480% (120)
	Not applicable	7% (17)	3% (3)	11% (14)		34% (17)	12% (3)	56% (14)
Control treatment in experimental arm	0	7% (17)	7% (7)	8% (10)		20% (10)	20% (5)	20% (5)
	<5%	9% (21)	8% (8)	10% (13)		30% (15)	28% (7)	32% (8)
	≥5%, <10%	5% (12)	8% (8)	3% (4)		20% (10)	24% (6)	16% (4)
	≥10%, <20%	5% (11)	6% (6)	4% (5)		20% (10)	24% (6)	16% (4)
	≥20%, <30%	5% (12)	2% (2)	8% (10)		12% (6)	8% (2)	16% (4)
	≥30%, <40%	1% (3)	1% (1)	2% (2)		4% (2)	4% (1)	4% (1)
	Unclear	0% (1)	1% (1)	0% (0)		14% (7)	8% (2)	20% (5)
	Not reported	64% (151)	65% (66)	64% (85)		90% (45)	84% (21)	96% (24)
	Not applicable	0% (0)	0% (0)	0% (0)		0% (0)	0% (0)	0% (0)
	Number randomized not reported/ unclear	1% (3)	3% (3)	0% (0)		6% (3)	12% (3)	0% (0)
Experimental treatment in	0	9% (20)	8% (8)	9% (12)		26% (13)	24% (6)	28% (7)

Domain		Trial			Trial outcome	Review		
		Overall (N = 235)	Cochrane (n = 102)	Non-Cochrane (n = 133)	(N = 315)	Overall (N = 50)	Cochrane (n = 25)	Non-Cochrane (n = 25)
control arm	<5%	11% (27)	10% (10)	13% (17)		32% (16)	28% (7)	36% (9)
	≥5%, <10%	4% (10)	7% (7)	2% (3)		18% (9)	24% (6)	12% (3)
	≥10%, <20%	3% (7)	2% (2)	4% (5)		12% (6)	8% (2)	16% (4)
	≥20%, <30%	1% (2)	1% (1)	1% (1)		4% (2)	4% (1)	4% (1)
	≥40%, <50%	0% (1)	0% (0)	1% (1)		2% (1)	0% (0)	4% (1)
	≥50%, <60%	0% (1)	0% (0)	1% (1)		2% (1)	0% (0)	4% (1)
	≥70%, <80%	0% (1)	0% (0)	1% (1)		2% (1)	0% (0)	4% (1)
	≥80%, <90%	1% (2)	0% (0)	2% (2)		2% (1)	0% (0)	4% (1)
	Unclear	4% (10)	3% (3)	5% (7)		6% (3)	4% (1)	8% (2)
	Not reported	64% (151)	67% (68)	62% (83)		88% (44)	84% (21)	92% (23)
	Not applicable	0% (0)	0% (0)	0% (0)		0% (0)	0% (0)	0% (0)
	Number randomized not reported/ unclear	1% (3)	3% (3)	0% (0)		6% (3)	12% (3)	0% (0)
Treatment switching reported per outcome	No	100% (235)	100% (102)	100% (133)		100% (50)	100% (25)	100% (25)
Treatment switching pre-specified	Reported as protocol specified	0% (1)	0% (0)	1% (1)		2% (1)	0% (0)	4% (1)
	Reported as not planned or allowed	4% (10)	0% (0)	8% (10)		10% (5)	0% (0)	20% (5)
	Otherwise reported as anticipated	2% (5)	2% (2)	2% (3)		6% (3)	8% (2)	4% (1)
	Unclear	1% (2)	0% (0)	2% (2)		4% (2)	0% (0)	8% (2)
	Not reported	92% (216)	98% (100)	87% (116)		94% (47)	96% (24)	92% (23)
	Not applicable	0% (1)	0% (0)	1% (1)		2% (1)	0% (0)	4% (1)
Treatment switching reported as protocol specified	Yes	1% (2)	2% (2)	0% (0)		4% (2)	8% (2)	0% (0)
	Not reported	77% (181)	66% (67)	86% (114)		98% (49)	96% (24)	100% (25)
	Not applicable	22% (52)	32% (33)	14% (19)		54% (27)	64% (16)	44% (11)
Treatment switching reasons	Disease-related (e.g. disease progression)	12% (29)	8% (8)	16% (21)		32% (16)	24% (6)	40% (10)
	Participant related (e.g. choose to switch)	9% (20)	14% (14)	5% (7)		20% (10)	28% (7)	12% (3)
	Administrative (e.g. interim analysis)	4% (9)	3% (3)	5% (6)		8% (4)	8% (2)	8% (2)
	Pre-condition related (e.g. too obese, allergies)	4% (9)	7% (7)	2% (2)		16% (8)	24% (6)	8% (2)
	Intervention related (e.g. adverse events)	2% (5)	3% (3)	2% (2)		8% (4)	8% (2)	8% (2)
	Investigator/ physician related (e.g. treating physicians decision)	1% (3)	1% (1)	2% (2)		6% (3)	4% (1)	8% (2)
	Not reported	13% (30)	10% (10)	15% (20)		38% (19)	24% (6)	52% (13)
	Not applicable	64% (151)	67% (68)	62% (83)		88% (44)	88% (22)	88% (22)
Analysis - Proportional hazards								
Proportional hazards assumption tested	Yes, statistical test (e.g. Log-log, Schoenfeld Residuals)	8% (19)	8% (8)	8% (11)	7% (23)	32% (16)	32% (8)	32% (8)
	Yes, visual inspection of curves	1% (2)	1% (1)	1% (1)	1% (2)	4% (2)	4% (1)	4% (1)
	No, but for other analyses of this outcome	0% (1)	1% (1)	0% (0)	0% (1)	2% (1)	4% (1)	0% (0)
	No	52% (123)	53% (54)	52% (69)	65% (206)	88% (44)	76% (19)	100% (25)
	Not applicable (e.g. no HRs calculated)	29% (69)	38% (39)	23% (30)	26% (83)	52% (26)	64% (16)	40% (10)

Domain		Trial			Trial outcome	Review		
		Overall (N = 235)	**Cochrane (n = 102)**	**Non-Cochrane (n = 133)**	**(N = 315)**	**Overall (N = 50)**	**Cochrane (n = 25)**	**Non-Cochrane (n = 25)**
Results of proportional hazards tests	Non-proportional	1% (3)	2% (2)	1% (1)	1% (3)	6% (3)	8% (2)	4% (1)
	Reasonably proportional	1% (2)	1% (1)	1% (1)	1% (2)	4% (2)	4% (1)	4% (1)
	Not reported for this analysis, but reasonably for other analysis of this outcome	0% (1)	0% (0)	1% (1)	0% (1)	2% (1)	0% (0)	4% (1)
	Not reported	6% (15)	6% (6)	7% (9)	6% (19)	26% (13)	24% (6)	28% (7)
	Not applicable	92% (216)	92% (94)	92% (122)	92% (290)	100% (50)	100% (25)	100% (25)

Abbreviations: HR = hazard ratio; LTFU = Loss to follow-up; RoB = risk of bias; TTE = time-to-event

11.3. Paper 3: Presentation of results of meta-analyses of time-to-event outcomes in form of absolute effects in systematic reviews

11.3.1. Work shares

Authors: Skoetz N*, **Goldkuhle M***, Weigl A, Dwan K, Labonté V, Dahm P, Meerpohl JJ, Djulbegovic B, Van Dalen EC (**contributed equally*)

Contributions by the doctoral student:
The article of this sub-project is published in shared first authorship of the doctoral student with Prof. Skoetz. After initiation of the meta-epidemiological study, the doctoral student designed the required data extraction schemes and respective forms together with Prof. Skoetz. He screened the literature and selected eligible evidence syntheses. Besides extracting relevant study data, the doctoral student coordinated the data extraction amongst all project participants and performed the analysis of the data. During the preparation of the publication, the doctoral student was responsible for the presentation of the data. In addition, he was involved in the first draft, extensively revised the later versions of the article and prepared it for publication. Finally, he was centrally involved in the acquisition of project participants and in contact with the co-authors.

Co-author contributions:
The project was initiated by Prof. Skoetz. Together with her, the doctoral student developed the tools for data analysis, selected relevant literature and extracted data. In addition, she too prepared the first draft of the publication and revised the article at later points. Other project participants with substantial impact on the project were the last author, Dr. Elvira van Dalen, who was also involved in the development of the data collection tools and supported the literature selection as well as the data extraction and analysis. She also revised the article. Aaron Weigl (Cologne, Germany), also selected the literature and extracted data. In addition, he supported the data analysis and the revision of the publication. The other co-authors (Prof. Philipp Dahm (Minnesota, USA), Prof. Benjamin Djulbegovic (Duarte, USA), Dr. Kerry Dwan (York, UK), Prof. Jörg Meerpohl (Freiburg, Germany), and Dr. Valerie Labonté (Freiburg, Germany) supported the data extraction and revision of the article.

11.3.2. Publication appendix

Online Appendix

Appendix-table 1: Baseline characteristics of included Cochrane Reviews

	Cochrane reviews (n=96)
Year last updated (Number of SR (%))	
2011	1 (1)
2012	9 (9)
2013	10 (10)
2014	9 (9)
2015	15 (16)
2016	26 (27)
2017	26 (27)
Cochrane Group (Number of SR (%))	
Anaesthesia, Critical and Emergency Care	1 (1)
Breast Cancer	12 (13)
Childhood Cancer	3 (3)
Cochrane Pain, Palliative and Supportive Care Group	1 (1)
Colorectal Cancer	8 (8)
Gynaecological, Neuro-oncology and Orphan Cancer	18 (19)
Haematological Malignancies	23 (24)
Hepato-Biliary	4 (4)
Lung Cancer	4 (4)
Oral Health	2 (2)
Skin	2 (2)
Upper GI and Pancreatic Diseases	14 (15)
Urology	4 (4)
Region (Number of SR (%))	
Africa	1 (1)
Asia	10 (10)
Australia/ New Zealand	11 (11)
Europe	62 (65)
North America	7 (7)
South America	5 (5)
Review type (Number of SR (%))	
Study-level data	87 (91)
IPD	3 (3)
Both	6 (6)
Network Meta-Analysis (Number of SR (%))	1 (1)
Study type (Number of SR (%))	
RCT	87 (91)
Non-RCT	2 (2)
Both	7 (7)
Disease (Number of SR (%))	
Bladder	2 (2)
Brain	5 (5)
Breast	12 (13)
Cancer in general	4 (4)
Cervical	2 (2)
Colorectal	9 (9)
Endometrial	3 (3)
Oesophagus	4 (4)
Gastric	4 (4)
Head and neck	2 (2)
Haematological	21 (22)
Liver	2 (2)

	Cochrane reviews (n=96)
Lung	4 (4)
Melanoma	2 (2)
Mixed	3 (3)
Other	6 (6)
Ovarian	7 (7)
Pancreas	2 (2)
Prostate	1 (1)
Renal	1 (1)
Intervention (Number of SR (%))	
Chemotherapy	28 (29)
Hormone	2 (2)
Mixed	5 (5)
Multiple	2 (2)
New drug	10 (10)
Radiotherapy	10 (10)
Supportive	23 (24)
Surgery	15 (16)
Thermal	1 (1)
Population (Number of SR (%))	
Adult	89 (93)
Paediatric	4 (4)
Both	3 (3)
Included studies (Median (IQR))	7 (3,25-12)
Included patients (Median (IQR))	1999 (551-4023)

SR= systematic review, IPD= Individual Patient Data,
RCT=randomised controlled trial, IQR= interquartile range	

References of included reviews

1. Archampong D, Borowski D, Wille-Jorgensen P, et al. Workload and surgeon's specialty for outcome after colorectal cancer surgery. *The Cochrane database of systematic reviews* 2012(3):Cd005391. doi: 10.1002/14651858.CD005391.pub3 [published Online First: 2012/03/16]

2. Balduzzi S, Mantarro S, Guarneri V, et al. Trastuzumab-containing regimens for metastatic breast cancer. *The Cochrane database of systematic reviews* 2014(6):Cd006242. doi: 10.1002/14651858.CD006242.pub2 [published Online First: 2014/06/13]

3. Bauer K, Rancea M, Roloff V, et al. Rituximab, ofatumumab and other monoclonal anti-CD20 antibodies for chronic lymphocytic leukaemia. *The Cochrane database of systematic reviews* 2012;11:Cd008079. doi: 10.1002/14651858.CD008079.pub2 [published Online First: 2012/11/16]

4. Bergner N, Monsef I, Illerhaus G, et al. Role of chemotherapy additional to high-dose methotrexate for primary central nervous system lymphoma (PCNSL). *The Cochrane database of systematic reviews* 2012;11:Cd009355. doi: 10.1002/14651858.CD009355.pub2 [published Online First: 2012/11/16]

5. Best LM, Mughal M, Gurusamy KS. Non-surgical versus surgical treatment for oesophageal cancer. *The Cochrane database of systematic reviews* 2016;3:Cd011498. doi: 10.1002/14651858.CD011498.pub2 [published Online First: 2016/03/30]

6. Best LMJ, Mughal M, Gurusamy KS. Laparoscopic versus open gastrectomy for gastric cancer. *Cochrane Database of Systematic Reviews* 2016(3) doi: 10.1002/14651858.CD011389.pub2

7. Blank O, von Tresckow B, Monsef I, et al. Chemotherapy alone versus chemotherapy plus radiotherapy for adults with early stage Hodgkin lymphoma. *Cochrane Database of Systematic Reviews* 2017(4) doi: 10.1002/14651858.CD007110.pub3

8. Bromham N, Schmidt-Hansen M, Astin M, et al. Axillary treatment for operable primary breast cancer. *Cochrane Database of Systematic Reviews* 2017(1) doi: 10.1002/14651858.CD004561.pub3

9. Cakmakkaya OS, Kolodzie K, Apfel CC, et al. Anaesthetic techniques for risk of malignant tumour recurrence. *The Cochrane database of systematic reviews* 2014(11):Cd008877. doi: 10.1002/14651858.CD008877.pub2 [published Online First: 2014/11/08]

10. Chan DLH, Segelov E, Wong RSH, et al. Epidermal growth factor receptor (EGFR) inhibitors for metastatic colorectal cancer. *Cochrane Database of Systematic Reviews* 2017(6) doi: 10.1002/14651858.CD007047.pub2

11. Chan KK, Glenny AM, Weldon JC, et al. Interventions for the treatment of oral and oropharyngeal cancers: targeted therapy and immunotherapy. *The Cochrane database of systematic reviews* 2015(12):Cd010341. doi: 10.1002/14651858.CD010341.pub2 [published Online First: 2015/12/02]

12. Cheuk DK, Chiang AK, Ha SY, et al. Interventions for prophylaxis of hepatic veno-occlusive disease in people undergoing haematopoietic stem cell transplantation. *The Cochrane database of*

systematic reviews 2015(5):Cd009311. doi: 10.1002/14651858.CD009311.pub2 [published Online First: 2015/05/29]

13. Chionh F, Lau D, Yeung Y, et al. Oral versus intravenous fluoropyrimidines for colorectal cancer. *Cochrane Database of Systematic Reviews* 2017(7) doi: 10.1002/14651858.CD008398.pub2
14. Deva S, Jameson M. Histamine type 2 receptor antagonists as adjuvant treatment for resected colorectal cancer. *The Cochrane database of systematic reviews* 2012(8):Cd007814. doi: 10.1002/14651858.CD007814.pub2 [published Online First: 2012/08/17]
15. Diaz-Nieto R, Orti-Rodriguez R, Winslet M. Post-surgical chemotherapy versus surgery alone for resectable gastric cancer. *The Cochrane database of systematic reviews* 2013(9):Cd008415. doi: 10.1002/14651858.CD008415.pub2 [published Online First: 2013/09/04]
16. Egger SJ, Willson ML, Morgan J, et al. Platinum-containing regimens for metastatic breast cancer. *Cochrane Database of Systematic Reviews* 2017(6) doi: 10.1002/14651858.CD003374.pub4
17. Estcourt LJ, Stanworth SJ, Doree C, et al. Comparison of different platelet count thresholds to guide administration of prophylactic platelet transfusion for preventing bleeding in people with haematological disorders after myelosuppressive chemotherapy or stem cell transplantation. *The Cochrane database of systematic reviews* 2015(11):Cd010983. doi: 10.1002/14651858.CD010983.pub2 [published Online First: 2015/11/19]
18. Falcetta FS, Medeiros LRF, Edelweiss MI, et al. Adjuvant platinum-based chemotherapy for early stage cervical cancer. *Cochrane Database of Systematic Reviews* 2016(11) doi: 10.1002/14651858.CD005342.pub4
19. Franklin J, Eichenauer DA, Becker I, et al. Optimisation of chemotherapy and radiotherapy for untreated Hodgkin lymphoma patients with respect to second malignant neoplasms, overall and progression-free survival: individual participant data analysis. *Cochrane Database of Systematic Reviews* 2017(9) doi: 10.1002/14651858.CD008814.pub2
20. Frost JA, Webster KE, Bryant A, et al. Lymphadenectomy for the management of endometrial cancer. *Cochrane Database of Systematic Reviews* 2017(10) doi: 10.1002/14651858.CD007585.pub4

21. Furness S, Glenny AM, Worthington HV, et al. Interventions for the treatment of oral cavity and oropharyngeal cancer: chemotherapy. *The Cochrane database of systematic reviews* 2011(4):Cd006386. doi: 10.1002/14651858.CD006386.pub3 [published Online First: 2011/04/15]

22. Galaal K, Al Moundhri M, Bryant A, et al. Adjuvant chemotherapy for advanced endometrial cancer. *The Cochrane database of systematic reviews* 2014(5):Cd010681. doi: 10.1002/14651858.CD010681.pub2 [published Online First: 2014/05/17]

23. Greenhalgh J, Dwan K, Boland A, et al. First-line treatment of advanced epidermal growth factor receptor (EGFR) mutation positive non-squamous non-small cell lung cancer. *The Cochrane database of systematic reviews* 2016(5):Cd010383. doi: 10.1002/14651858.CD010383.pub2 [published Online First: 2016/05/26]

24. Gurusamy KS, Kumar S, Davidson BR. Prophylactic gastrojejunostomy for unresectable periampullary carcinoma. *The Cochrane database of systematic reviews* 2013(2):Cd008533. doi: 10.1002/14651858.CD008533.pub3 [published Online First: 2013/03/02]

25. Gurusamy KS, Kumar S, Davidson BR, et al. Resection versus other treatments for locally advanced pancreatic cancer. *The Cochrane database of systematic reviews* 2014(2):Cd010244. doi: 10.1002/14651858.CD010244.pub2 [published Online First: 2014/03/01]

26. Gurusamy KS, Pallari E, Midya S, et al. Laparoscopic versus open transhiatal oesophagectomy for oesophageal cancer. *The Cochrane database of systematic reviews* 2016;3:Cd011390. doi: 10.1002/14651858.CD011390.pub2 [published Online First: 2016/04/01]

27. Haun MW, Estel S, Rücker G, et al. Early palliative care for adults with advanced cancer. *Cochrane Database of Systematic Reviews* 2017(6) doi: 10.1002/14651858.CD011129.pub2

28. Hickey BE, James ML, Lehman M, et al. Fraction size in radiation therapy for breast conservation in early breast cancer. *The Cochrane database of systematic reviews* 2016;7:Cd003860. doi: 10.1002/14651858.CD003860.pub4 [published Online First: 2016/07/19]

29. Hickey BE, Lehman M, Francis DP, et al. Partial breast irradiation for early breast cancer. *The Cochrane database of systematic reviews* 2016;7:Cd007077. doi: 10.1002/14651858.CD007077.pub3 [published Online First: 2016/07/19]

30. Holtick U, Albrecht M, Chemnitz JM, et al. Bone marrow versus peripheral blood allogeneic haematopoietic stem cell transplantation for haematological malignancies in adults. *The Cochrane database of systematic reviews* 2014(4):Cd010189. doi: 10.1002/14651858.CD010189.pub2 [published Online First: 2014/04/22]

31. Hu X, Fang Y, Hui X, et al. Radiotherapy for diffuse brainstem glioma in children and young adults. *The Cochrane database of systematic reviews* 2016(6):Cd010439. doi: 10.1002/14651858.CD010439.pub2 [published Online First: 2016/07/06]

32. Hutzschenreuter F, Monsef I, Kreuzer KA, et al. Granulocyte and granulocyte-macrophage colony stimulating factors for newly diagnosed patients with myelodysplastic syndromes. *The Cochrane database of systematic reviews* 2016;2:Cd009310. doi: 10.1002/14651858.CD009310.pub2 [published Online First: 2016/02/18]

33. Itchaki G, Gafter-Gvili A, Lahav M, et al. Anthracycline-containing regimens for treatment of follicular lymphoma in adults. *The Cochrane database of systematic reviews* 2013(7):Cd008909. doi: 10.1002/14651858.CD008909.pub2 [published Online First: 2013/07/09]

34. Jaaback K, Johnson N, Lawrie TA. Intraperitoneal chemotherapy for the initial management of primary epithelial ovarian cancer. *The Cochrane database of systematic reviews* 2016(1):Cd005340. doi: 10.1002/14651858.CD005340.pub4 [published Online First: 2016/01/13]

35. Janmaat VT, Steyerberg EW, van der Gaast A, et al. Palliative chemotherapy and targeted therapies for esophageal and gastroesophageal junction cancer. *Cochrane Database of Systematic Reviews* 2017(11) doi: 10.1002/14651858.CD004063.pub4

36. Jeffery M, Hickey BE, Hider PN, et al. Follow-up strategies for patients treated for non-metastatic colorectal cancer. *The Cochrane database of systematic reviews* 2016;11:Cd002200. doi: 10.1002/14651858.CD002200.pub3 [published Online First: 2016/11/25]

37. Khan L, Soliman H, Sahgal A, et al. External beam radiation dose escalation for high grade glioma. *The Cochrane database of systematic reviews* 2016(8):Cd011475. doi: 10.1002/14651858.CD011475.pub2 [published Online First: 2016/08/20]

38. Kharfan-Dabaja M, Mhaskar R, Reljic T, et al. Mycophenolate mofetil versus methotrexate for prevention of graft-versus-host disease in people receiving allogeneic hematopoietic stem cell transplantation. *The Cochrane database of systematic reviews* 2014(7):Cd010280. doi: 10.1002/14651858.CD010280.pub2 [published Online First: 2014/07/26]

39. Kidane B, Coughlin S, Vogt K, et al. Preoperative chemotherapy for resectable thoracic esophageal cancer. *The Cochrane database of systematic reviews* 2015(5):Cd001556. doi: 10.1002/14651858.CD001556.pub3 [published Online First: 2015/05/20]

40. Kindts I, Laenen A, Depuydt T, et al. Tumour bed boost radiotherapy for women after breast-conserving surgery. *Cochrane Database of Systematic Reviews* 2017(11) doi: 10.1002/14651858.CD011987.pub2

41. Kokka F, Bryant A, Brockbank E, et al. Hysterectomy with radiotherapy or chemotherapy or both for women with locally advanced cervical cancer. *The Cochrane database of systematic reviews* 2015(4):Cd010260. doi: 10.1002/14651858.CD010260.pub2 [published Online First: 2015/04/08]

42. Kong A, Johnson N, Kitchener HC, et al. Adjuvant radiotherapy for stage I endometrial cancer. *Cochrane Database of Systematic Reviews* 2012(4) doi: 10.1002/14651858.CD003916.pub4

43. Kunath F, Grobe HR, Rucker G, et al. Non-steroidal antiandrogen monotherapy compared with luteinising hormone-releasing hormone agonists or surgical castration monotherapy for advanced prostate cancer. *The Cochrane database of systematic reviews* 2014(6):Cd009266. doi: 10.1002/14651858.CD009266.pub2 [published Online First: 2014/07/01]

44. Kunath F, Schmidt S, Krabbe L-M, et al. Partial nephrectomy versus radical nephrectomy for clinical localised renal masses. *Cochrane Database of Systematic Reviews* 2017(5) doi: 10.1002/14651858.CD012045.pub2

45. Kyrgidis A, Tzellos T, Mocellin S, et al. Sentinel lymph node biopsy followed by lymph node dissection for localised primary cutaneous melanoma. *The Cochrane database of systematic reviews* 2015(5):Cd010307. doi: 10.1002/14651858.CD010307.pub2 [published Online First: 2015/05/17]

46. Lawal AO, Musekiwa A, Grobler L. Interferon after surgery for women with advanced (Stage II-IV) epithelial ovarian cancer. *The Cochrane database of systematic reviews* 2013(6):Cd009620. doi: 10.1002/14651858.CD009620.pub2 [published Online First: 2013/06/07]

47. Lawrie TA, Bryant A, Cameron A, et al. Pegylated liposomal doxorubicin for relapsed epithelial ovarian cancer. *The Cochrane database of systematic reviews* 2013(7):Cd006910. doi: 10.1002/14651858.CD006910.pub2 [published Online First: 2013/07/10]

48. Lawrie TA, Rabbie R, Thoma C, et al. Pegylated liposomal doxorubicin for first-line treatment of epithelial ovarian cancer. *The Cochrane database of systematic reviews* 2013(10):Cd010482. doi: 10.1002/14651858.CD010482.pub2 [published Online First: 2013/10/22]

49. Lawrie TA, Winter-Roach BA, Heus P, et al. Adjuvant (post-surgery) chemotherapy for early stage epithelial ovarian cancer. *The Cochrane database of systematic reviews* 2015(12):Cd004706. doi: 10.1002/14651858.CD004706.pub5 [published Online First: 2015/12/18]

50. Lee CI, Goodwin A, Wilcken N. Fulvestrant for hormone-sensitive metastatic breast cancer. *Cochrane Database of Systematic Reviews* 2017(1) doi: 10.1002/14651858.CD011093.pub2

51. Li X, Xu S, Tan Y, et al. The effects of idarubicin versus other anthracyclines for induction therapy of patients with newly diagnosed leukaemia. *The Cochrane database of systematic reviews* 2015(6):Cd010432. doi: 10.1002/14651858.CD010432.pub2 [published Online First: 2015/06/04]

52. Majumdar A, Roccarina D, Thorburn D, et al. Management of people with early- or very early-stage hepatocellular carcinoma. *Cochrane Database of Systematic Reviews* 2017(3) doi: 10.1002/14651858.CD011650.pub2

53. Mao C, Fu XH, Yuan JQ, et al. Interleukin-2 as maintenance therapy for children and adults with acute myeloid leukaemia in first complete remission. *The Cochrane database of systematic reviews* 2015(11):Cd010248. doi: 10.1002/14651858.CD010248.pub2 [published Online First: 2015/11/07]

54. Mhaskar R, Clark OA, Lyman G, et al. Colony-stimulating factors for chemotherapy-induced febrile neutropenia. *The Cochrane database of systematic reviews* 2014(10):Cd003039. doi: 10.1002/14651858.CD003039.pub2 [published Online First: 2014/10/31]

55. Mhaskar R, Kumar A, Miladinovic B, et al. Bisphosphonates in multiple myeloma: an updated network meta-analysis. *Cochrane Database of Systematic Reviews* 2017(12) doi: 10.1002/14651858.CD003188.pub4

56. Mhaskar R, Wao H, Miladinovic B, et al. The role of iron in the management of chemotherapy-induced anemia in cancer patients receiving erythropoiesis-stimulating agents. *The Cochrane database of systematic reviews* 2016;2:Cd009624. doi: 10.1002/14651858.CD009624.pub2 [published Online First: 2016/02/05]

57. Mocellin S, Baretta Z, Roqué i Figuls M, et al. Second-line systemic therapy for metastatic colorectal cancer. *Cochrane Database of Systematic Reviews* 2017(1) doi: 10.1002/14651858.CD006875.pub3

58. Mocellin S, Lens MB, Pasquali S, et al. Interferon alpha for the adjuvant treatment of cutaneous melanoma. *The Cochrane database of systematic reviews* 2013(6):Cd008955. doi: 10.1002/14651858.CD008955.pub2 [published Online First: 2013/06/19]

59. Mocellin S, McCulloch P, Kazi H, et al. Extent of lymph node dissection for adenocarcinoma of the stomach. *The Cochrane database of systematic reviews* 2015(8):Cd001964. doi: 10.1002/14651858.CD001964.pub4 [published Online First: 2015/08/13]

60. Moja L, Tagliabue L, Balduzzi S, et al. Trastuzumab containing regimens for early breast cancer. *The Cochrane database of systematic reviews* 2012(4):Cd006243. doi: 10.1002/14651858.CD006243.pub2 [published Online First: 2012/04/20]

61. Morgan J, Wyld L, Collins KA, et al. Surgery versus primary endocrine therapy for operable primary breast cancer in elderly women (70 years plus). *Cochrane Database of Systematic Reviews* 2014(5) doi: 10.1002/14651858.CD004272.pub3

62. Moschetti I, Cinquini M, Lambertini M, et al. Follow-up strategies for women treated for early breast cancer. *The Cochrane database of systematic reviews* 2016(5):Cd001768. doi: 10.1002/14651858.CD001768.pub3 [published Online First: 2016/05/28]

63. Mota BS, Riera R, Ricci MD, et al. Nipple- and areola-sparing mastectomy for the treatment of breast cancer. *Cochrane Database of Systematic Reviews* 2016(11) doi: 10.1002/14651858.CD008932.pub3

64. Muchtar E, Vidal L, Ram R, et al. The role of maintenance therapy in acute promyelocytic leukemia in the first complete remission. *The Cochrane database of systematic reviews* 2013(3):Cd009594. doi: 10.1002/14651858.CD009594.pub2 [published Online First: 2013/04/02]

65. O'Carrigan B, Wong MHF, Willson ML, et al. Bisphosphonates and other bone agents for breast cancer. *Cochrane Database of Systematic Reviews* 2017(10) doi: 10.1002/14651858.CD003474.pub4

66. Patil CG, Pricola K, Sarmiento JM, et al. Whole brain radiation therapy (WBRT) alone versus WBRT and radiosurgery for the treatment of brain metastases. *Cochrane Database of Systematic Reviews* 2017(9) doi: 10.1002/14651858.CD006121.pub4

67. Peinemann F, Bartel C, Grouven U. First-line allogeneic hematopoietic stem cell transplantation of HLA-matched sibling donors compared with first-line ciclosporin and/or antithymocyte or antilymphocyte globulin for acquired severe aplastic anemia. *The Cochrane database of systematic reviews* 2013(7):Cd006407. doi: 10.1002/14651858.CD006407.pub2 [published Online First: 2013/07/25]

68. Peinemann F, Enk H, Smith LA. Autologous hematopoietic stem cell transplantation following high-dose chemotherapy for nonrhabdomyosarcoma soft tissue sarcomas. *Cochrane Database of Systematic Reviews* 2017(4) doi: 10.1002/14651858.CD008216.pub5

69. Peinemann F, Tushabe DA, van Dalen EC, et al. Rapid COJEC versus standard induction therapies for high-risk neuroblastoma. *The Cochrane database of systematic reviews* 2015(5):Cd010774. doi: 10.1002/14651858.CD010774.pub2 [published Online First: 2015/05/20]

70. Peinemann F, van Dalen EC, Enk H, et al. Retinoic acid postconsolidation therapy for high-risk neuroblastoma patients treated with autologous haematopoietic stem cell transplantation. *Cochrane Database of Systematic Reviews* 2017(8) doi: 10.1002/14651858.CD010685.pub3

71. Rancea M, Monsef I, von Tresckow B, et al. High-dose chemotherapy followed by autologous stem cell transplantation for patients with relapsed/refractory Hodgkin lymphoma. *The Cochrane database of systematic reviews* 2013(6):Cd009411. doi: 10.1002/14651858.CD009411.pub2 [published Online First: 2013/06/21]

72. Resende HM, Jacob LFP, Quinellato LV, et al. Combination chemotherapy versus single-agent chemotherapy during preoperative chemoradiation for resectable rectal cancer. *Cochrane Database of Systematic Reviews* 2015(10) doi: 10.1002/14651858.CD008531.pub2

73. Riviere D, Gurusamy KS, Kooby DA, et al. Laparoscopic versus open distal pancreatectomy for pancreatic cancer. *The Cochrane database of systematic reviews* 2016;4:Cd011391. doi: 10.1002/14651858.CD011391.pub2 [published Online First: 2016/04/05]

74. Roccarina D, Majumdar A, Thorburn D, et al. Management of people with intermediate-stage hepatocellular carcinoma. *Cochrane Database of Systematic Reviews* 2017(3) doi: 10.1002/14651858.CD011649.pub2

75. Ronellenfitsch U, Schwarzbach M, Hofheinz R, et al. Perioperative chemo(radio)therapy versus primary surgery for resectable adenocarcinoma of the stomach, gastroesophageal junction, and lower esophagus. *The Cochrane database of systematic reviews* 2013(5):Cd008107. doi: 10.1002/14651858.CD008107.pub2 [published Online First: 2013/06/04]

76. Santos FN, de Castria TB, Cruz MR, et al. Chemotherapy for advanced non-small cell lung cancer in the elderly population. *The Cochrane database of systematic reviews* 2015(10):Cd010463. doi: 10.1002/14651858.CD010463.pub2 [published Online First: 2015/10/21]

77. Sarmiento JM, Venteicher AS, Patil CG. Early versus delayed postoperative radiotherapy for treatment of low-grade gliomas. *The Cochrane database of systematic reviews* 2015(6):Cd009229. doi: 10.1002/14651858.CD009229.pub2 [published Online First: 2015/06/30]

78. Shepherd ARH, Shepherd E, Brook NR. Intravesical Bacillus Calmette-Guérin with interferon-alpha versus intravesical Bacillus Calmette-Guérin for treating non-muscle-invasive bladder cancer. *Cochrane Database of Systematic Reviews* 2017(3) doi: 10.1002/14651858.CD012112.pub2

79. Sickinger MT, von Tresckow B, Kobe C, et al. Positron emission tomography-adapted therapy for first-line treatment in individuals with Hodgkin lymphoma. *The Cochrane database of systematic reviews* 2015;1:Cd010533. doi: 10.1002/14651858.CD010533.pub2 [published Online First: 2015/01/13]

80. Skoetz N, Bauer K, Elter T, et al. Alemtuzumab for patients with chronic lymphocytic leukaemia. *The Cochrane database of systematic reviews* 2012(2):Cd008078. doi: 10.1002/14651858.CD008078.pub2 [published Online First: 2012/02/18]

81. Skoetz N, Will A, Monsef I, et al. Comparison of first-line chemotherapy including escalated BEACOPP versus chemotherapy including ABVD for people with early unfavourable or advanced stage Hodgkin lymphoma. *Cochrane Database of Systematic Reviews* 2017(5) doi: 10.1002/14651858.CD007941.pub3

82. Song H, Zhu J, Lu D. Molecular-targeted first-line therapy for advanced gastric cancer. *The Cochrane database of systematic reviews* 2016;7:Cd011461. doi: 10.1002/14651858.CD011461.pub2 [published Online First: 2016/07/20]

83. Soon YY, Tham IW, Lim KH, et al. Surgery or radiosurgery plus whole brain radiotherapy versus surgery or radiosurgery alone for brain metastases. *The Cochrane database of systematic*

reviews 2014(3):Cd009454. doi: 10.1002/14651858.CD009454.pub2 [published Online First: 2014/03/04]

84. Sultan S, Coles B, Dahm P. Alvimopan for recovery of bowel function after radical cystectomy. *Cochrane Database of Systematic Reviews* 2017(5) doi: 10.1002/14651858.CD012111.pub2

85. Theurich S, Fischmann H, Shimabukuro-Vornhagen A, et al. Polyclonal anti-thymocyte globulins for the prophylaxis of graft-versus-host disease after allogeneic stem cell or bone marrow transplantation in adults. *The Cochrane database of systematic reviews* 2012(9):Cd009159. doi: 10.1002/14651858.CD009159.pub2 [published Online First: 2012/09/14]

86. Tonia T, Mettler A, Robert N, et al. Erythropoietin or darbepoetin for patients with cancer. *Cochrane Database of Systematic Reviews* 2012(12) doi: 10.1002/14651858.CD003407.pub5

87. Vellayappan BA, Soon YY, Ku GY, et al. Chemoradiotherapy versus chemoradiotherapy plus surgery for esophageal cancer. *Cochrane Database of Systematic Reviews* 2017(8) doi: 10.1002/14651858.CD010511.pub2

88. Wagner AD, Syn NLX, Moehler M, et al. Chemotherapy for advanced gastric cancer. *Cochrane Database of Systematic Reviews* 2017(8) doi: 10.1002/14651858.CD004064.pub4

89. Weis S, Franke A, Berg T, et al. Percutaneous ethanol injection or percutaneous acetic acid injection for early hepatocellular carcinoma. *The Cochrane database of systematic reviews* 2015;1:Cd006745. doi: 10.1002/14651858.CD006745.pub3 [published Online First: 2015/01/27]

90. Weis S, Franke A, Mossner J, et al. Radiofrequency (thermal) ablation versus no intervention or other interventions for hepatocellular carcinoma. *The Cochrane database of systematic reviews* 2013(12):Cd003046. doi: 10.1002/14651858.CD003046.pub3 [published Online First: 2013/12/21]

91. Wiggans AJ, Cass GK, Bryant A, et al. Poly(ADP-ribose) polymerase (PARP) inhibitors for the treatment of ovarian cancer. *The Cochrane database of systematic reviews* 2015(5):Cd007929. doi: 10.1002/14651858.CD007929.pub3 [published Online First: 2015/05/21]

92. Wulaningsih W, Wardhana A, Watkins J, et al. Irinotecan chemotherapy combined with fluoropyrimidines versus irinotecan alone for overall survival and progression-free survival in patients with advanced and/or metastatic colorectal cancer. *The Cochrane database of systematic reviews* 2016;2:Cd008593. doi: 10.1002/14651858.CD008593.pub3 [published Online First: 2016/02/13]

93. Wuntakal R, Seshadri S, Montes A, et al. Luteinising hormone releasing hormone (LHRH) agonists for the treatment of relapsed epithelial ovarian cancer. *The Cochrane database of systematic reviews* 2016(6):Cd011322. doi: 10.1002/14651858.CD011322.pub2 [published Online First: 2016/06/30]

94. Yang ZY, Liu L, Mao C, et al. Chemotherapy with cetuximab versus chemotherapy alone for chemotherapy-naive advanced non-small cell lung cancer. *The Cochrane database of systematic reviews* 2014(11):Cd009948. doi: 10.1002/14651858.CD009948.pub2 [published Online First: 2014/11/18]

95. Zacher J, Kasenda B, Engert A, et al. The role of additional radiotherapy for primary central nervous system lymphoma. *The Cochrane database of systematic reviews* 2014(6):Cd009211. doi: 10.1002/14651858.CD009211.pub2 [published Online First: 2014/06/17]

96. Zhu J, Li R, Tiselius E, et al. Immunotherapy (excluding checkpoint inhibitors) for stage I to III non-small cell lung cancer treated with surgery or radiotherapy with curative intent. *Cochrane Database of Systematic Reviews* 2017(12) doi: 10.1002/14651858.CD011300.pub2

11.4. Paper 4: Guideline for handling informative censoring as a study-limitation in evidence syntheses

11.4.1. Work shares

Authors: Goldkuhle M, Bender R, Akl EA, van Dalen EC, Nevitt S, Mustafa RA, Guyatt GH, Trivella M, Djulbegovic B, Schünemann H, Cinquini M, Kreuzberger N, Skoetz N

Contributions by the doctoral student:
This sub-project was initiated and conceptualized by the doctoral student under supervision of Prof. Skoetz. He established the background of the problem through an extensive review of the literature as well as theoretical evaluations. Subsequently, he presented the problem to the GRADE Working Group and confirmed with them the necessity of a GRADE guideline article. In addition to the general project administration, the doctoral student recruited the involved experts, including, but not limited to the co-authors of the article, and managed the communication amongst them.
During the development process of the guideline, the doctoral student chaired all expert meetings, presented examples for structured discussions, evaluated the discussion results and translated them into the guidance article. Together with his co-authors Nina Kreuzberger and Prof. Ralf Bender, he performed the imputation, which is part of the guideline article, and interpreted and presented its results. Finally, the doctoral student drafted the publication of the guideline and incorporated all comments from co-authors during several rounds of systematic revision.

The GRADE Working Group policy requires that articles submitted under the name of the working group must be consented by all members. Respectively, any article in question must be approved by at least 80% of the participants in a face-to-face vote at an official meeting of the working group. The doctoral student presented and successfully defended the proposed guideline in Hamilton, Canada, to approximately 100 attendees. Finally, he revised the guideline publication in accordance with the peer reviewer comments and coordinated all additional notes from co-authors.

Co-author contributions:
The execution of the work was supervised by Prof. Skoetz. She, and also the other co-authors of the article, Prof. Ralf Bender (Cologne, Germany), Prof. Elie A Akl (Beirut, Lebanon), Dr. Elvira van Dalen, Dr. Sarah Nevitt (Liverpool, UK), Prof. Reem A. Mustafa (Kansas City, USA), Prof. Gordon Guyatt (Hamilton, Canada), Dr. Marialena Trivella (Oxford, UK), Prof. Benjamin Djulbegovic, Prof. Holger Schünemann (Hamilton, Canada), Dr. Michela Cinquini (Milan, Italy) and Nina Kreuzberger, participated in the structured discussions of the examples presented through the doctoral student and contributed important methodological expertise. They commented on and consented to the first version of the publication and supported its revision.

11.4.2. Publication appendix

Appendix

Appendix A1: Independent and non-informative censoring

Non-informative censoring, as described by Lagakos (15), requires *"that the time-point of a censoring event holds no information about an individual's likelihood to experience the event of interest* (its survival time)". This means that the true distribution of the survival time, where no individual is lost from observation and individuals are observed until the event occurs, and the true censoring distribution, where the study ends before all subjects experience the event and censored individuals do not experience the event prior to the end of study, provide no information for each other. Informative censoring is sometimes referred to as a type of selection bias under the reasoning that loss to follow-up or withdrawal in randomized trials leads to selection after randomization, when certain participants due to certain measured or unmeasured characteristics or conditions may be less likely or more likely to be censored and as well less likely or more likely to experience the event of interest. In other words, the association of the risk of being censored and the risk of experiencing the event results from a common source of both risks (40). The definition of independent censoring is not equivalent to non-informative censoring and Lagakos (15) shows that dependent censoring is a special form of informative censoring, however, in most situations where the assumption of independent censoring is violated, the assumption of non-informative censoring is too (13).

Appendix A2: Reporting requirements for survival analysis that allow to assess the risk of bias due to censoring

In order to assess the suitability of the independent censoring assumption by users, including systematic review and guideline authors, the methods in a primary study report should ideally provide detailed definitions of the assessed outcomes including the event(s) of interest, the time of origin and all conditions leading to censoring despite end-of observation (e.g. absence of the event at study closure, loss to follow-up or withdrawal due to competing events) (17, 18). Standardized outcome definitions would here be highly preferable (19). With regard to the applied analysis methods we would demand that it is explicitly reported why the assumption of dependent censoring is feasible. When outcomes which include competing risks are assessed, we would require the application and reporting of appropriate methods, which will be outlined in a future guidance. The result section should hold the total events of interest and number of censored individuals in each of the study arms and the number of participants censored separately of those before the end the observational period including the individual reasons (17, 18). It is highly desirable that Kaplan-Meier curves, if feasible, are given for each of the assessed outcomes. In the curves, the time-points of censored events should be indicated as well as the number at risk below the curves for appropriate time-points (22). The number of censored individuals for certain time-points with an indication of censoring reasons is an option to enhance transparency. Lastly, the duration of follow-up for each study arm should be given and the calculation method should be clearly stated (55).

Appendix A3. Reconstruction of survival data to illustrate the impact of early dependent censoring

To illustrate the impact of early dependent censoring on comparisons, we reconstructed individual participant data from the survival curves published for the analysis of overall survival in the article by Denis et al. (32). The study shows an unbalanced number of censored participants particularly during early follow-up, with more censored participants in the intervention arm compared to the control arm. Given the clear reporting in the survival curves, we were able to reconstruct outcome event and censoring time points for the individuals in each of the compared groups. We verified our proceeding with the algorithm presented Guyot et al. (33) that allows to reconstruct individual participant level data from published survival curves. The algorithm attributes a constant rate of censoring to intervals in between outcome events and time-points for which a number of individuals at risk is reported. It therefore works optimal assuming independent censoring. Under the objective of our illustration, we decided not to directly use the dataset resulting from the algorithm proposed by Guyot and colleagues but to work with individual patient data that we reconstructed directly from the published survival curve. Nevertheless, we used the data set produced under application of the algorithm to confirm the consistency of our manually extracted data by comparing the data points retrieved through both approaches.

We extracted data with the software DigitizeIt (www.digitizeit.de), which allows to assign each point on the survival curve a corresponding time-point on the x-axis. We marked all declines of the curve as outcome event and all crosses as censoring time-points. The reported curve for the experimental arm was unclear for two censoring events in the first interval (0 to 5 months) and the last interval (over 15 months) respectively, which were not directly identifiable on the curve, but must have occurred in these intervals as indicated by the number of individuals at risk. Similarly, for the curve representing survival in the control arm, two censoring events were not identifiable within the first interval (0 to 5 months). For all scenarios we assumed the missing censoring events to have happened on the last possible time-point of this interval (4.99 and 18.99 months). In the so retrieved dataset, we modified the survival data of participants censored within the first seven months of follow-up to illustrate the impact of early dependent censoring. We present a hypothetical scenario where all participants censored prior to seven months of follow-up experience the outcome event one month after the original censoring. Subsequently, we calculated hazard ratios with the Cox proportional hazards model and present Kaplan-Meier survival curves. All statistical analyses were performed using the software R (56). We want to point out that our imputation does not claim to compare a difference in treatment effects, but to illustrate the loss of certainty that is introduced to survival analyses through a high degree of censoring particularly during the early period of follow-up.

References

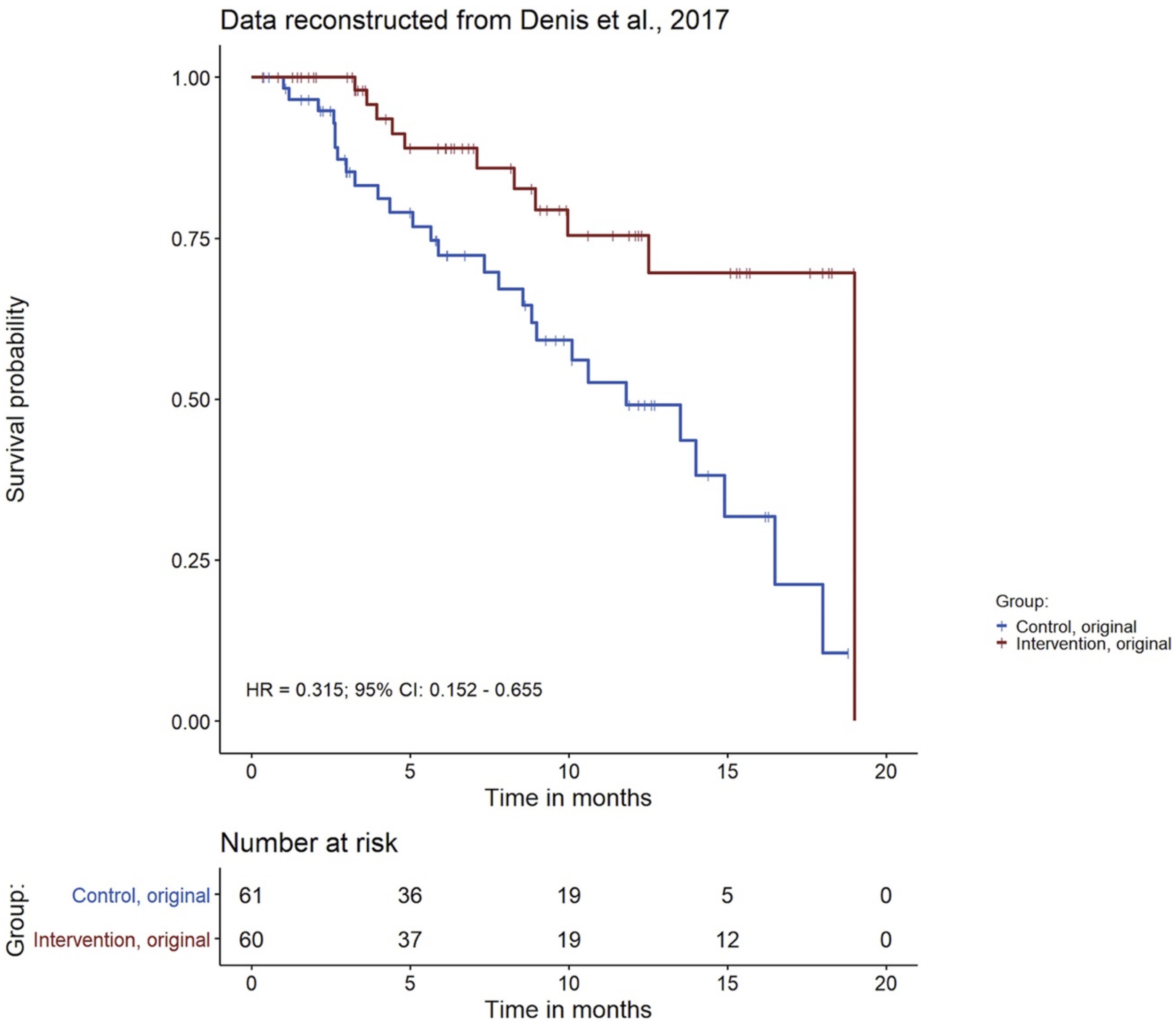
[55] Schemper M., Smith T.L. A note on quantifying follow-up in studies of failure time. Control Clin Trials 1996;17:343–346.

[56] R Development Core Team. R. A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2008.

Appendix A4. Reconstructed survival curves

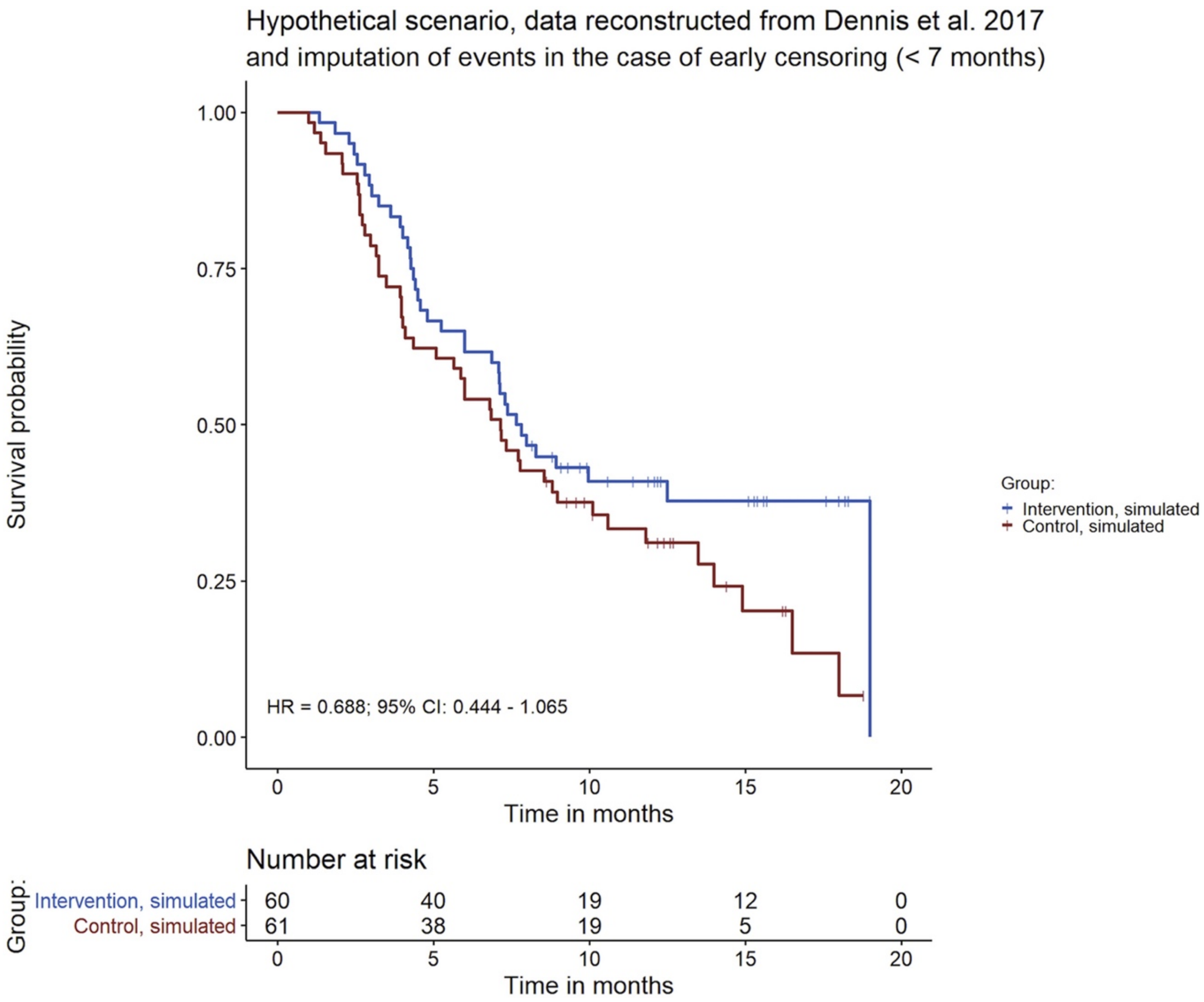
Appendix-figure 1: Kaplan-Meier survival curves calculated from the individual participant level data reconstructed from the analysis of overall survival in Denis et al. (32).

Appendix-figure 2: Kaplan-Meier survival curve calculated from the individual participant level data reconstructed from the analysis of overall survival in Denis et al. (32). Participants who were censored prior to seven months of follow-up in both study arms were set to experience the outcome event one month after original censoring.



Group:	0	5	10	15	20
Control, original	61	36	19	5	0
Intervention, original	60	37	19	12	0

Appendix figure 1



Appendix figure 2

11.5. Paper 5: Guideline for presenting the results of meta-analyses of time-to-event outcomes in form of absolute effects

11.5.1. Work shares

Authors: Skoetz N, **Goldkuhle M**, van Dalen EC, Akl EA, Trivella M, Mustafa RA, Nowak A, Dahm P, Schünemann H, Bender R

Contributions by the doctoral student:
The doctoral student was involved in the initial conceptualization of the guideline and administrated the entire project. He performed the underlying literature searches and the retrieval and review of necessary background literature. He formulated the introduction and conclusions of the guideline, and reviewed and extended all sections of the original draft of the manuscript in depths. He coordinated the involvement of the participating experts, whose contributions and revisions he collated, and implemented all expert contributions in the guideline article. During the preparation phase of the guideline publication, he was responsible for all reviews and their editing.

Co-author contributions:
Prof. Skoetz initiated and supervised the development of the guideline. She developed the original draft of the manuscript and reviewed and edited all later drafts of the article. Prof. Ralf Bender was also involved in the conceptualization of the guideline and provided central statistical expertise, he was involved in the original draft of the article and reviewed later versions. Also involved in the conceptualization of the article was Dr. Elvira C. van Dalen, she was also reviewed the original draft and subsequent versions of the guideline publication. Prof. Elie A. Akl., Dr. Marialena Trivella, Prof. Reem A. Mustafa, Prof. Philipp Dahm and Prof. Holger Schünemann provided methodological expertise and revised the original draft of the article as well as its later versions after review. Finally, Artur Nowak provided important expertise on underlying statistical software packages and also revised the versions of the publications.

11.6. Eidesstattliche Erklärung

Ich versichere, dass ich die von mir vorgelegte Dissertation selbstständig angefertigt, die benutzten Quellen und Hilfsmittel vollständig angegeben und die Stellen der Arbeit - einschließlich Tabellen, Karten und Abbildungen -, die anderen Werken im Wortlaut oder dem Sinn nach entnommen sind, in jedem Einzelfall als Entlehnung kenntlich gemacht habe; dass diese Dissertation noch keiner anderen Fakultät oder Universität zur Prüfung vorgelegen hat; dass sie - abgesehen von unten angegebenen Teilpublikationen - noch nicht veröffentlicht worden ist sowie, dass ich eine solche Veröffentlichung vor Abschluss des Promotionsverfahrens nicht ohne Genehmigung der Dekanin / dem Dekan vornehmen werde. Die Bestimmungen dieser Ordnung sind mir bekannt. Die von mir vorgelegte Dissertation ist von Frau Prof. Dr. med. Nicole Skoetz betreut worden.

Veröffentlichte Publikationen:

1. Skoetz N*, Goldkuhle M*, Weigl A, Dwan K, Labonté V, Dahm P, Meerpohl JJ, Djulbegovic B, van Dalen EC. Methodological review showed correct absolute effect size estimates for time-to-event outcomes in less than one-third of cancer-related systematic reviews. Journal of Clinical Epidemiology, 2019 Apr; 108, 1-9. doi: https://doi.org/10.1016/j.jclinepi.2018.12.006 (*contributed equally)

2. Skoetz N, Goldkuhle M, van Dalen EC, Akl EA, Trivella M, Mustafa RA, Nowak A, Dahm P, Schünemann H, Bender R. GRADE Working Group. GRADE guidelines 27: how to calculate absolute effects for time-to-event outcomes in summary of findings tables and Evidence Profiles. Journal of Clinical Epidemiology. 2020 Feb;118:124-131. doi: 10.1016/j.jclinepi.2019.10.015.

3. Goldkuhle M, Bender R, Akl EA, van Dalen EC, Nevitt S, Mustafa RA, Guyatt GH, Trivella M, Djulbegovic B, Schünemann H, Cinquini M, Kreuzberger N, Skoetz N. GRADE Guidelines: 29. Rating the certainty in time-to-event outcomes-Study limitations due to censoring of participants with missing data in intervention studies. Journal of Clinical Epidemiology. 2021 Jan 129:126-137. doi: 10.1016/j.jclinepi.2020.09.017.

4. Goldkuhle M, Hirsch C, Iannizzi C, Bora AM, Bender R, van Dalen EC, Hemkens LG, Trivella M, Monsef I, Kreuzberger N, Skoetz N. Meta-epidemiological review identified variable reporting and handling of time-to-event analyses in publications of trials included in meta-analyses of systematic reviews. Journal of Clinical Epidemiology. 2023 Jul;159:174-189. doi: 10.1016/j.jclinepi.2023.05.023.

5. Goldkuhle M, Hirsch C, Iannizzi C, Zorger AM, Bender R, van Dalen EC, et al. Exploring the characteristics, methods and reporting of systematic reviews with meta-analyses of time-to-event outcomes: a meta-epidemiological study. BMC Medical Research Methodology. 2024;24(1):291.

Ich versichere, dass ich alle Angaben wahrheitsgemäß nach bestem Wissen und Gewissen gemacht habe und verpflichte mich, jedmögliche, die obigen Angaben betreffenden Veränderungen, dem Promotionsausschuss unverzüglich mitzuteilen.

Datum Unterschrift