Received: 4 March 2019 | Revised: 23 October 2019 | Accepted: 24 October 2019

DOI: 10.1002/jrsm.1384

REVIEW

WILEY Research Synthesis Methods

Individual participant data meta-analysis of intervention studies with time-to-event outcomes: A review of the methodology and an applied example

Valentijn M.T. de Jong[1] | **Karel G.M. Moons**[1,2] | **Richard D. Riley**[3] | **Catrin Tudur Smith**[4] | **Anthony G. Marson**[5] | **Marinus J.C. Eijkemans**[1] | **Thomas P.A. Debray**[1,2]

[1]Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, Utrecht, the Netherlands

[2]Cochrane Netherlands, Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, Utrecht, the Netherlands

[3]Centre for Prognosis Research, Research Institute for Primary Care and Health Sciences, Keele University, Staffordshire, UK

[4]Department of Biostatistics, University of Liverpool, Liverpool, UK

[5]Department of Molecular and Clinical Pharmacology, University of Liverpool, Liverpool, UK

Correspondence

Valentijn de Jong, Julius Center for Health Sciences and Primary Care, UMC Utrecht, Str. 6.131, P.O. Box 85500, 3508GA Utrecht, the Netherlands.
Email: valentijn.m.t.de.jong@gmail.com

Funding information

European Union, European Commission, Directorate-General for Research and Innovation, Horizon 2020 Framework Programme, Grant/Award Number: 825746; ZonMW, Netherlands Organization for Health Research and Development, Grant/Award Numbers: 91617050, 91810615

Many randomized trials evaluate an intervention effect on time-to-event outcomes. Individual participant data (IPD) from such trials can be obtained and combined in a so-called IPD meta-analysis (IPD-MA), to summarize the overall intervention effect. We performed a narrative literature review to provide an overview of methods for conducting an IPD-MA of randomized intervention studies with a time-to-event outcome. We focused on identifying good methodological practice for modeling frailty of trial participants across trials, modeling heterogeneity of intervention effects, choosing appropriate association measures, dealing with (trial differences in) censoring and follow-up times, and addressing time-varying intervention effects and effect modification (interactions).We discuss how to achieve this using parametric and semi-parametric methods, and describe how to implement these in a one-stage or two-stage IPD-MA framework. We recommend exploring heterogeneity of the effect(s) through interaction and non-linear effects. Random effects should be applied to account for residual heterogeneity of the intervention effect. We provide further recommendations, many of which specific to IPD-MA of time-to-event data from randomized trials examining an intervention effect.We illustrate several key methods in a real IPD-MA, where IPD of 1225 participants from 5 randomized clinical trials were combined to compare the effects of Carbamazepine and Valproate on the incidence of epileptic seizures.

KEYWORDS

heterogeneity, individual participant data, intervention, meta-analysis, time-to-event

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.
© 2019 The Authors. *Research Synthesis Methods* published by John Wiley & Sons Ltd.

1 | INTRODUCTION

Relative intervention effects (eg, hazard ratios) are most reliably evaluated in randomized clinical trials (RCT). However, multiple RCTs of the same intervention may provide inconclusive or conflicting evidence on efficacy or safety. Discrepancies between evidence from different RCTs may arise due to chance, or in particular due to heterogeneity in the true intervention effect. This heterogeneity is commonly caused by across-trial differences in, for example, study design (eg, recruitment strategy, length of follow-up, or analysis methods), case-mix of participants, definition of the studied outcome(s), the implementation (eg, dosage or intensity) of the intervention. This motivates the need to systematically integrate and summarize evidence across trials, to facilitate evidence-based-medicine.

This can be achieved using a systematic review with meta-analysis (MA). Whereas most meta-analyses are based on aggregated data (AD) from available literature, individual participant (or patient) data meta-analyses (IPD-MA) of multiple intervention studies are considered the gold standard. [1-3] IPD-MA offers several advantages, as the meta-analyst has full control of the data analysis and uses the data at the individual participant level. [4] Key advantages are the standardisation of outcome and follow-up definitions, checking of data and quality, proper modelling of time-to-event outcomes, and the exploration of intervention-covariate interactions at the participant level. [4,5] It may thus come to no surprise that IPD-MA are increasingly common. [6,7]

Extensive guidance has previously been provided for conducting an IPD-MA of intervention effects, for various types of outcome data, such as binary, [7-9] continuous, [6,7,10,11] ordinal [7] and count outcomes. [7] Yet, IPD-MA are especially useful when analyzing time-to-event outcomes in intervention studies, as censored outcomes can be reassessed for the meta-analysis, survival measures (eg, hazard ratios, median survival) can be calculated directly and independent to trial reporting, follow-up length can often be increased, time-varying hazard ratios can be examined, and effect modifiers (intervention-covariate interactions) can be assessed. [12,13]

Whereas a wealth of methods have been developed for analyzing and predicting time-to-event outcomes in single studies, [14-17] limited guidance exists on their application in IPD-MA settings. In this article, we aim to provide readers with this guidance, by means of our systematic search of databases, narrative review and explanation, and an applied example. Although we focus IPD-MA of trials, the methods we describe are also applicable to multi-center trials.

In the next section, we provide the principles as well as several major issues of time-to-event analyses, that are common in not only IPD-MA but also in single studies. In section 3 we provide details of our systematic literature search of methodology for IPD-MA of time-to-event outcomes, and then a narrative review thereof follows in section 4 where we discuss the one- and two-stage approaches to meta-analysis, and in section 5 where we discuss issues in more detail. Then, in section 6 we apply several key methods of the review to a real IPD meta-analysis of clinical trials. Finally, we give provide a discussion in section 7 and concluding remarks in section 8.

Highlights
What is known?

- Time-to-event (survival) data can be analyzed with Cox Proportional Hazards regression, but proportionality of hazards should be tested.
- Individual participant data (IPD) from multiple randomized trials can be summarized by meta-analyzing the trial-specific estimates of the individual trials (studies) or by analyzing the pooled data with a mixed-effects model that accounts for between-trial heterogeneity in intervention effect and frailty of participants.

What is new?

- We summarize published guidance, statistical methods and software for survival analysis using IPD from multiple randomized clinical trials.
- We discuss how between-trial heterogeneity of intervention effects may appear and how its sources can be investigated.
- We illustrate the methods on real epilepsy data and provide R code.

Potential impact for other fields

- Meta-analysis is not only relevant in medical research, but also in other research areas.
- The methods naturally extend to meta-analysis of non-randomized studies, where treatment effect estimates need to be adjusted for confounding.

2 | PRINCIPLES OF TIME-TO-EVENT ANALYSIS

The analysis of trials with a survival outcome (eg, death) typically involves statistical models that account for the time $T_{\text{surv},\ i}$ elapsed until subject i, $i = 1, .., n$ developed the event of interest. We here denote the probability for subject i to remain event-free for at least t time by the survival function $S(t) = Pr(T_{\text{surv},\ i} > t)$. A key challenge in time-to-event (TTE) data is that for many participants $T_{\text{surv},\ i}$ is censored to $T_{\text{cens},\ i}$, for instance due to dropout or the end of the study. This implies that for those participants $T_{\text{surv},\ i} > T_{\text{cens},\ i}$. Hence, the outcome for subject i is typically summarized by the observed event-free or survival time $T_i = \min(T_{\text{surv},\ i}, T_{\text{cens},\ i})$ and the event status D_i (where $D = 0$ when censored, and $D = 1$ when the event of interest was observed to have occured). We can compare the survival times of intervention groups and control, while accounting for censoring, with a variety of regression methods.

A commonly used method for analyzing right-censored TTE data is the Cox proportional hazards (PH) model. [18] In this semi-parametric model the effect of the covariates is modeled parametrically, whereas the baseline is left unspecified. It is typically assumed that the ratio of the hazards for any two individuals is constant, irrespective of t. The hazard $h(t|\mathtt{X})$ for an individual with covariate vector $\mathtt{X}' = (X_1,...,X_k)$ is given by Equation 1.1 (Table 1), where $^{\mathtt{T}} = (\beta_1,...,\beta_k)$ is a vector of regression parameters. The function $h_0(t)$ represents the baseline hazard, which is left unspecified. [14,15] The hazard ratio for two individuals $i = 1, 2$ is then given by $\exp\{'(\mathtt{X}_1 - \mathtt{X}_2)\}$. For the analysis of randomized trials, $\mathtt{X}$ typically just contains a single covariate representing the intervention indicator (eg, $X_i = 0$ for subjects in the control arm and $X_i = 1$ for subjects in the intervention arm) such that $\exp(\beta)$ can directly be interpreted as the relative intervention effect.

An important consideration is whether to include other (prognostic) covariates in the Cox PH model alongside treatment. In many time-to-event models, including the Cox PH model, the observed unadjusted intervention effect of a protective intervention may change over time due to covariates (ie, frailty), even if these covariates are perfectly balanced between the intervention groups. [19,20] Frail participants will have a higher incidence rate than less frail participants. If the intervention is protective, frail participants in the intervention group will have a lower incidence rate than frail participants in a control (or an ineffective intervention) group and participants that are not frail. Over time, the proportion in the control group that is still at risk will increasingly consist of participants that are not frail, whereas this will take longer for the intervention group, thereby resulting in an imbalance in frailty. For trials with a high event rate and most frailty distributions, the unadjusted intervention effect will attenuate towards the null (hazard ratio of 1) as time progresses, which violates the proportional hazards assumption. [21] The unadjusted intervention effect is then the marginal intervention effect, [22] i.e. the average intervention effect for the population as a whole, averaged across all time-points. Hence, it is dependent on the length of the follow-up.

If the intention is to measure a conditional intervention effect, that is, the intervention effect for a participant with given covariate values, the observed unadjusted intervention effect is often not valid. Instead, covariates should be included in the model, to obtain a conditional intervention effect. [23,24] Further, the adjustment for a prognostic covariate often increases the power for finding an intervention effect. [25] Alternatively, an AFT model could be used (sections 5.1 and 5.2), for which the effect of missing covariates is absorbed into the baseline parameters, leaving the unadjusted intervention effect unaffected. [21]

The Cox PH model has numerous appealing properties, in particular allowing the estimation of hazard ratios for included covariates without requiring the shape of the baseline hazard to be specified. However, its implementation is not always justified. For instance, difficulties may arise when hazards are non-proportional. Although effects to model non-PH can be included (eg, with splines, interactions or time-varying effects) in a Cox PH

TABLE 1 Models for two-stage time-to-event meta-analysis

Type	Model	Hazard function	Survival function	Ref.	No.
Proportional Hazards	General model[a]	$h_0(t)\exp('\mathtt{X})$	$S(t\|\mathtt{X}) = S_0(t)^{\exp('\mathtt{X})}$	12,14,162	1.1
	Exponential	$\lambda\exp('\mathtt{X})$	$S(t\|\mathtt{X}) = \exp(-\lambda t\exp('\mathtt{X}))$	14,16,162	1.2
	Weibull[b]	$\lambda\nu t^{\nu-1}\exp('\mathtt{X})$	$S(t\|\mathtt{X}) = \exp(-\lambda t^{\nu}\exp('\mathtt{X}))$	14,16,162,163	1.3
	Gompertz[c]	$\lambda\exp(\psi t)\exp('\mathtt{X})$	$S(t\|\mathtt{X}) = \exp\left(-\frac{\lambda}{\psi}(\exp(\psi t)-1)\ \exp('\mathtt{X})\right)$	14,16,164	1.4
Accelerated Failure Time	General model	$h_0(t\exp\{'\mathtt{X}\})\exp('\mathtt{X})$	$S(t\|\mathtt{X}) = S_0(t\exp('\mathtt{X}))$	14,16,89	1.5
	Weibull	$\lambda\nu t^{\nu-1}(\exp('\mathtt{X}))^{\nu}$	$S(t\|\mathtt{X}) = \exp(-\lambda t^{\nu}\exp(\nu'\mathtt{X}))$	14,16	1.6
	Log-logistic[d]	$\frac{\varphi}{t\{1+t^{-\varphi}exp(-'\mathtt{X})\}}$	$\log\frac{1-S(t\|\mathtt{X})}{S(t\|\mathtt{X})} = \varphi log(t) + '\mathtt{X}$	89,98,99	1.7

[a] In the Cox Proportional Hazards model, the baseline hazard $h_0(t)$ is left unspecified.

[b] ν is a shape parameter, λ is a scale parameter.

[c] The Gompertz distribution can be generalized to the Gompertz-Makeham distribution by adding a constant to the hazard function. [165].

[d] The log-logistic model is a proportional odds model, where the β parameters can be interpreted as log-odds ratios.

model, this usually complicates the interpretation of the estimated intervention effect. For these reasons it is often recommended to adopt a model where proportionality occurs on another scale when proportionality of hazards is violated, which is discussed in section 5.2. When absolute survival probabilities for individual participants are of primary interest, it can be useful to define a parametric function for $h_0(t)$, and thus to abandon Cox PH models altogether, [26,27] which is discussed in section 5.1. Indeed, even when the focus is mainly on an intervention effect, translation of its hazard ratio to the absolute risk scale is important, which requires the baseline survival to be modelled, either parametrically or non-parametrically. For a full overview of R packages on time-to-event analysis, see cran.r-project.org/web/views/Survival.html.

3 | IPD META-ANALYSIS METHODS: REVIEW

Increasingly often, IPD from multiple studies are available for analysis. This introduces new challenges and allows for different approaches for analysis, which we set out to identify. We conducted a literature review to identify scientific articles concerning statistical methods for IPD-MA of time-to-event data.

3.1 | Methods

We systematically searched through Pubmed and Web of Science using the search filters supplied in Supporting Information 1, from conception until December 31st, 2018. In addition, we added suggestions and performed cross-reference checks of the obtained articles. Articles were considered eligible for inclusion if they described statistical methods for analyzing multiple or clustered individual participant data sets with a time-to-event outcome. Publications that met at least one of the following criteria were excluded from our review:

- Full text of the manuscript not available,
- Not published in English,
- Not a peer reviewed article,
- Application of methods without methodological focus,
- No focus on at least one of the following topics:
  - time-to-event outcomes,
  - IPD,
  - estimation of intervention effects,
  - meta-analysis or analysis of clustered data.

3.2 | Results

A total of 1887 unique records were identified through our search strategy, and were deemed eligible for title and abstract screening (Figure 1). Of these, 1713 were removed during screening because the titles did not have a methodological focus. The remaining 174 records were assessed on the full-text, of which 58 met the inclusion criteria and 116 did not. Further, a total of 159 unique records were assessed after being suggested or found through cross-referencing. Of these, 16 suggestions and 54 cross-references met the inclusion criteria and were included in the review. A total of 128 articles were included in the review, of which a complete list can be found in Supporting Information 3.

The core methods for analyzing TTE outcomes in IPD-MA are described in section 4. The structure of this section was defined independent of the review, yet the description of methods therein has resulted from the review. Further, extensions to these methods, such as relaxing the proportionality of hazards assumption, modeling multiple interventions or outcomes, and methods for

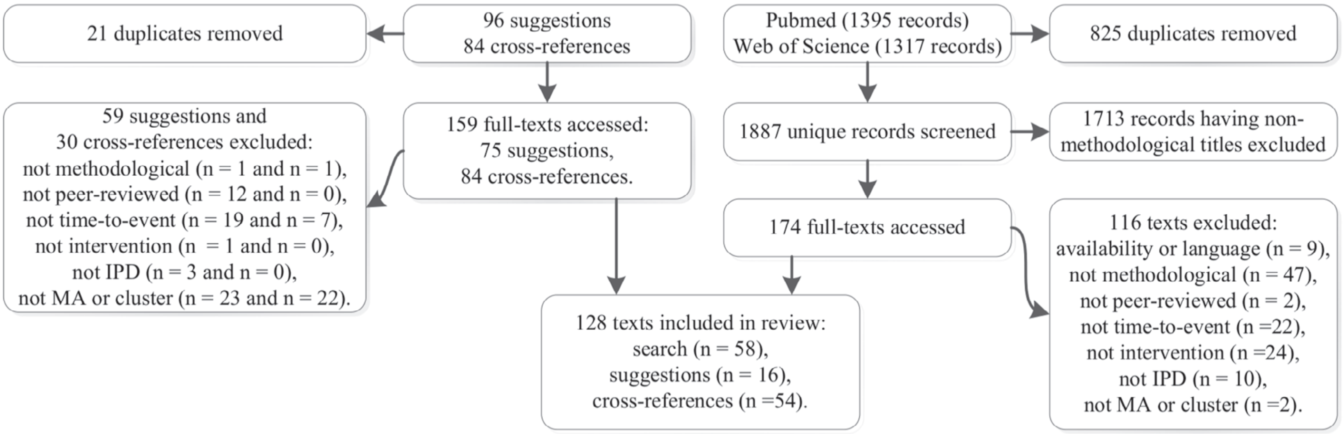
21 duplicates removed
96 suggestions
84 cross-references
Pubmed (1395 records)
Web of Science (1317 records)
825 duplicates removed
59 suggestions and
30 cross-references excluded:
not methodological (n = 1 and n = 1),
not peer-reviewed (n = 12 and n = 0),
not time-to-event (n = 19 and n = 7),
not intervention (n = 1 and n = 0),
not IPD (n = 3 and n = 0),
not MA or cluster (n = 23 and n = 22).
159 full-texts accessed:
75 suggestions,
84 cross-references.
1887 unique records screened
1713 records having non-methodological titles excluded
174 full-texts accessed
116 texts excluded:
availability or language (n = 9),
not methodological (n = 47),
not peer-reviewed (n = 2),
not time-to-event (n =22),
not intervention (n =24),
not IPD (n = 10),
not MA or cluster (n =2).
128 texts included in review:
search (n = 58),
suggestions (n = 16),
cross-references (n =54).

FIGURE 1 Flowchart of inclusion and exclusion of papers for review

TABLE 2 Ten Recommendations for the IPD-MA of TTE data from Randomized Trials Examining an Intervention Effect

Recommendation	Reference
The Cox model may be the default model of choice, but proportionality of hazards	46,166
should be tested, for example, with interaction or time-varying effects for the intervention.	
Consider non-PH models.	89,92,121,167,168
Account for clustering in one-stage models, preferably by stratification of the baseline.	19,29,65,67,73,74
Adjust for covariates measured before randomization.	25,163,169
Apply one-stage models if trials are very small or the outcome very rare.	16,28
In one-stage models, center covariates within trials.	88
Model participant-level interactions on the participant-level.	45
For the intervention effect (& its interaction effects), apply random effects & investigate heterogeneity.	5,12,44,78
If competing risks are present & absolute risks are of interest, apply competing risks models.	105,108,109,111,170
Multiple imputation of missing covariates must account for clustering & time-to-event,	136-138,140,143
using the event indicator and the Nelson-Aalen cumulative hazard.	

TABLE 3 Methods for Modeling Heterogeneity

Baseline	Coefficients	Modeled difference between trials
Common	Common	No difference, same for every trial
Frailty	Random Effects	Proportional differences, difference between trials follows distribution
Fixed[a]	Fixed[b]	Proportional differences, estimated per trial. Same shape between trials.
Stratified		Non-proportional differences. Estimated per trial, with different shapes.

[a] By adding trial indicators to the model.

[b] By adding trial indicators * variable interaction to the model.

missing data are described in section 5.1, which was grouped according to the topics identified in the review. The review has resulted in ten key recommendations backed by references, which are summarized in Table 2.

4 | DESCRIPTION OF METHODS

4.1 | Time-to-event analysis in individual participant data meta-analysis

When IPD from multiple trials are available, summary estimates for relative intervention effects can be obtained using the so-called one-stage or two-stage approaches. [28] In the conceptually simpler two-stage approach (section 4.2), the IPD from each trial is analyzed separately to produce trial-specific estimates of relative intervention effect (eg, hazard ratios), using the same methodology in each trial (eg, Cox regression). In the second stage, estimates of intervention effect are combined into a weighted average using traditional meta-analysis methods that ideally account for possible between-trial heterogeneity. In the one-stage approach (section 4.3), data from all studies are analyzed in one analysis, and a variety of methods can be used to account for clustering of participants within studies. [7,14,16,29,30] In both the one- and two-stage approaches, methods to account for heterogeneity in intervention effects across studies are available (Table 3). [1,7,30] In the one-stage approach, one must also decide how to model or account for heterogeneity in other parameters (such as adjustment factors or terms defining the baseline hazard). For a discussion on the choice between the one-stage and two-stage approaches see section 8.

4.2 | Two-stage approach

The two-stage approach is often considered the most convenient approach for IPD meta-analysis, as it does not necessarily require IPD to be exchanged. For instance, each trial can be analyzed separately, and only their summary statistics are combined. The approach is particularly appealing when not all trials provide IPD, as it allows reported intervention effects and their respective standard errors from non-IPD trials to be analyzed in the second stage, together with the estimates from the IPD trials.

In the first stage, common methods for TTE analysis can be used to obtain estimates of relative intervention effect for each trial (so-called aggregate data). For instance, when applying Cox regression (Equation 1.1), this yields the log hazard ratio estimates $\hat{\beta}_j$ and their corresponding error variance $V\left(\hat{\beta}_j\right)$, for trial $j = 1, ..., J$. Afterwards, the estimated intervention effects can be summarized by calculating a weighted average. For instance, in a so-called common (or fixed) effect meta-analysis it is assumed that all trials share a common intervention effect β_{IV}, which can be derived as follows:

$$\begin{aligned}\beta_{\mathrm{IV}} &= \frac{\sum_{j=1}^{J}\frac{\hat{\beta}_j}{V\left(\hat{\beta}_j\right)}}{\sum_{j=1}^{J}\frac{1}{V\hat{\beta}_j)}}\\ V(\beta_{\mathrm{IV}}) &= \frac{1}{\sum_{j=1}^{J}\frac{1}{V\left(\hat{\beta}_j\right)}}\end{aligned} \tag{1}$$

where V is the variance. Hereby, it is assumed that the within-trial variances $V\left(\hat{\beta}_j\right)$ are known (ie, estimated without uncertainty). The common effect meta-analysis model can also be formulated as follows:

$$\hat{\beta}_j\tilde{\mathcal{N}}\left(\beta_{\mathrm{IV}}, V\left(\hat{\beta}_j\right)\right) \tag{2}$$

If certain trials provide no IPD, but the intervention effect and its variance are available in the literature, these can be included in the second stage of the two-stage framework, [31] provided that the models in the first stage are specified the same. If a trial has a small sample size, the Maximum Likelihood estimator of the intervention effect can be affected by small sample bias. [32] Worse still, if considerable censoring is present, the likelihood may be monotone and the Maximum Likelihood may be inestimable, depending on the intervention and covariate distributions. [33] This can be resolved by applying Firth's correction to the likelihood in the first stage, [32,33] or by opting for a one-stage model instead.

The assumption that an intervention effect is common across trials is often unrealistic, as trials are often affected by between-trial heterogeneity. [34,35] This heterogeneity may, for instance, appear when participant-level covariates interact with the intervention effect (ie, effect modification), when small sample bias is present in some estimates of the intervention effect, or when aggregate data are based on invalid modeling assumptions (eg, in the presence of non-proportional hazards, non-PH). For time-to-event analysis, between-trial heterogeneity may also arise due to selection effects. In particular, participants who are more frail and therefore more susceptible to the outcome, are no longer at risk after having an event. Therefore, over time, the most frail participants are removed from the risk set, whereas the less frail participants remain at risk (see section 2). [14,19,23,36] This, in turn, may lead to different intervention effects across trials if the follow-up length differs across trials. For these reasons, in the two-stage approach it is generally recommended to adopt a random effects meta-analysis model, which is typically specified as:

$$\begin{aligned}&\hat{\beta}_j\ \tilde{\mathcal{N}}\left(\beta_j, V\left(\hat{\beta}_j\right)\right)\\ &\beta_j\ \tilde{\mathcal{N}}\left(\beta_{\mathrm{RE}}, \tau^2\right)\end{aligned} \tag{3}$$

In contrast to common effect models, random effects models allow for differences in $\hat{\beta}_j$ due to sampling error *within* studies (reflected by $V\left(\hat{\beta}_j\right)$) and due to heterogeneity in the true intervention effects β_j *across* studies (reflected by τ^2). Estimates for β_{RE} can thus be interpreted as the average intervention effect across studies. A confidence interval for $\hat{\beta}_{\mathrm{RE}}$ is traditionally constructed as $\hat{\beta}_{\mathrm{RE}} \pm z_{1-\alpha/2}\sqrt{V\left(\hat{\beta}_{\mathrm{RE}}\right)}$, where $z_{1-\alpha/2}$ is the upper $\alpha/2$ quantile of the standard normal distribution. [37] To account for the uncertainty in τ^2 and thereby improve the coverage of the interval, the Hartung-Knapp approach to confidence intervals is given by $\hat{\beta}_{\mathrm{RE}} \pm t_{J-1,1-\alpha/2}\sqrt{V_{\mathrm{HK}}\left(\hat{\beta}_{\mathrm{RE}}\right)}$, where $t_{J-1,\ 1-\alpha/2}$ is the upper $\alpha/2$ quantile of a t-distribution with J - 1° of freedom, and $V_{\mathrm{HK}}\left(\hat{\beta}_{\mathrm{RE}}\right)$ is a modified variance estimate. [38-42]

4.2.1 | Heterogeneity of the intervention effect in the two-stage approach

Statistical heterogeneity in the intervention effect can be recognized by $\hat{\tau} > 0$. The influence of heterogeneity on intervention effects may be explored by constructing a prediction interval, which estimates the interval of the likely intervention effect in a (new) individual trial and can be calculated approximately as follows [43,44]:

$$\hat{\beta}_{\mathrm{RE}} \pm t_{J-k,1-\alpha/2}\sqrt{\hat{\tau}^2 + V\left(\hat{\beta}_{\mathrm{RE}}\right)}, \tag{4}$$

where $\hat{\beta}_{\mathrm{RE}}$ is an estimate of β_{RE} and $V\left(\hat{\beta}_{\mathrm{RE}}\right)$ its variance. Typically the $t_{J-2,\ 1-\alpha/2}$ quantile is used here, although similar to the confidence interval there is no consensus on the distribution and its degrees of freedom. [43,44] When random effects models indicate the presence of important statistical heterogeneity (ie, $\hat{\tau} > 0$, or a wide prediction interval) of the intervention effect, the interpretation of the overall summary estimate, $\hat{\beta}_{\mathrm{RE}}$, may become difficult or meaningless. Therefore, it is often helpful to identify sources of heterogeneity in intervention effect (see Table 4). [12] This can, for instance, be achieved by assessing the relation between relevant trial-level covariates (eg, level of blinding, or dosage) and the trial effect estimates, also known as meta-regression. [45]

TABLE 4 Potential sources of Heterogeneity in Time-to-event Meta-Analysis

Source	Solutions	Reference
Non PH + Differences in follow-up time	Interaction terms	45,88,171
	Model effect(s) as time-varying, use splines	83,91
	Use a different model (eg, AFT)	27,83,92,103,167,168
Difference in case-mix	Include covariates/ prognostic factors	36,163
	AFT model	92,163,168
Selective dropout or competing risk	Model dropout or competing risk	108,111,170,172
Small sample bias in some studies	Bias correction	32
	One-stage MA	7,28,172,173
	Arcsine transform (for two-stage MA)	172

PH: Proportional Hazards; AFT: Accelerated Failure Time; MA: Meta-Analysis. Heterogeneity can be diagnosed by applying frailty and/or random effects terms. [13,29,78] If heterogeneity remains, for example, due to differences in study protocols, stratification of baseline hazard/frailty and/or random effects terms must be applied. [24,36].

When patient-level associations with treatment effect are of interest, it is better to model interactions between participant-level characteristics (eg, participant age) on the participant level. In the two-stage approach, the statistical interaction between the relevant covariate and intervention are first estimated separately in each trial, and then the resulting coefficients are meta-analyzed using traditional meta-analysis models. [34,46] When the intervention effect changes over time, differences in follow-up time between trials will lead to heterogeneous estimates of intervention effect across trials, if unaccounted for. This heterogeneity of intervention effects can be quantified with random-effects meta-analysis, but would preferably be modeled directly (section 5.2).

4.2.2 | Estimation

A commonly used approach to estimate the heterogeneity from the random effects model (Equation 3), is to use the method of moments by DerSimonian and Laird (DL). [47] This estimator is biased downwards when the true heterogeneity is moderate or high and sample sizes are low, as the variance estimates are assumed to be known and fixed, [48] leading many researchers to suggest alternatives, the most important of which are mentioned here. The two-step Paule-Mandel method is similar to DL, but iteratively estimates the study weights, and has reduced bias for high values of τ. Another alternative is the Maximum Likelihood (ML) estimator. Although the MSE of the ML estimator for τ is small, it is very biased when τ is large and the included studies are small. [49] The Restricted Maximum Likelihood (REML) estimator yields less biased estimates of τ and has relatively low MSE. [50-52] Therefore, REML and the two-step Paule-Mandel method are the recommended estimators for τ. [48,52]

As there may be considerable uncertainty in the heterogeneity estimate regardless of which estimator is used, [52] it is recommended to report a confidence interval for the heterogeneity as well. [53] This may be estimated with the Q-profile method [41,54] or the generalised Cochran between-study variance method. [49] Further, it should be noted that when fewer than 10 trials are included in the meta-analysis, or when trials are small or the outcome rare, no currently available method can reliably estimate the heterogeneity. [52]

Even though estimates for heterogeneity in meta-analysis tend to be biased in many situations, this barely biases the summary effect estimate, unless there are very few events. [52] The confidence intervals of the summary effect can be constructed by applying the Hartung-Knapp-Sidik-Jonkman HKSJ method for confidence intervals, [55,56] which had good coverage in simulations for a minimum of two studies, unless the number of events was very low. [52,57] This may be corrected by applying a modification that ensures that the confidence intervals are at least as wide as a fixed-effects meta-analysis confidence interval. [52] Hence, it is currently recommended to apply a random effects model estimated with REML or two-step Paule-Mandel, and to use the HKSJ method for confidence intervals. [52] Alternatively, Bayesian random-effects models may be used. However, in the simulation studies discussed here either aggregate data or non time-to-event IPD were generated, which is a concern considering that it has been suggested that the performance of the estimators may be related to the type of outcome. [49] For a comprehensive overview of meta-analysis estimators see [49,58,59], for a comparison of their performance see [48,52], for an overview of software see [49] as well as the two recent packages admetan and ipdmetan, [60] and for an up-to-date overview of R packages see cran.r-project.org/web/views/MetaAnalysis.html.

4.3 | One-stage approach

4.3.1 | Accounting for clustering

When applying the one-stage approach, within-trial and between-trial relationships are estimated simultaneously, which can give a more complete understanding of the data. [13] As is the case for two-stage meta-analysis, a one-stage meta-analysis must account for clustering (Table 3). [29,30] Participants in different studies may differ on unmeasured covariates, which will lead to a biased estimate of the conditional (ie, for a participant with given covariate values) intervention effect regardless of balance of these covariates between intervention groups, if not adjusted for (section 2). [20] Whereas the two-stage approach naturally deals with this by estimating separate baseline hazards for the different studies, in the one-stage approach we can use stratification (section 4.3.2), frailty models (section 4.3.3) or marginal models (section 4.3.4).

4.3.2 | Stratified models

A commonly used approach is to apply a Cox model with stratified baseline hazards but a common intervention effect (Equation 5.1, Table 5). [12,16,61,62] This allows the shapes of the baseline hazards to vary between trials, whereas the hazards of the different intervention groups are assumed to be proportional within trials, and gives a single estimate of overall intervention effect. When the sample sizes per trial are very small and many trials are included, the stratification of baselines is less efficient than the use of frailty terms, [16] though it also requires fewer assumptions as it fully accounts for any differences in baselines between trials. For the meta-analysis of trials that are each powered to detect a clinically significant intervention effect this should not be an issue, thereby making the stratification of the baseline the preferred model specification.

4.3.3 | Frailty models

Rather than stratifying the baseline hazard across the trials, it is possible to model their distribution through frailty terms. A frailty term is a random parameter (ie, random intercept) within the baseline hazard function that is assumed to follow a specified distribution and thereby allows for differences in baseline rate between (groups of) participants that are a result of unmeasured covariates. Shared frailty models (Equation 5.2, Table 5) are designed to account for these differences in unmeasured covariates between trials. Therefore, the assumption in a frailty model is that the baseline hazards in each study have the same shape but a different magnitude. The estimated intervention effect is then to be interpreted relative to other participants in the same trial with the same frailty and covariates. If the baseline hazard of this model is left unspecified, this leads to the Cox PH model with random trial intercept. [12,14] When data from multiple multi-center studies are combined, nested frailty models can be applied. [63]

It is common to assume a gamma distribution for the frailty, for mathematical or computational reasons,

TABLE 5 Models for one-stage time-to-event meta-analysis

Type	Model	Hazard function	Survival function	Ref.	No.
Proportional Hazards	Stratified baseline	$h_{0j}(t)\exp('\mathtt{X}_j)$	$S_j\left(t\vert\mathtt{X}_j\right)=S_{0j}(t)^{exp\left('\mathtt{X}_j\right)}$	12,14,16,18,174	5.1
	Shared frailty	$h_0(t)\eta_j\exp\left('\mathtt{X}_j\right)$ where $\eta_j\tilde{\text{Gamma}}(\theta)$ or $log\left(\eta_j\right)\tilde{\text{Normal}}(0,\tau^2)$	$S_j\left(t\vert\mathtt{X}_j\right)=S_0(t)^{\eta_j\exp\left('\mathtt{X}_j\right)}$	12,14,16,19,174	5.2
	Random effects	$h_0(t)\exp\left('\mathtt{X}_j+\mathtt{b}_j'\mathtt{Z}_j\right)$ where $\mathtt{b}_j\text{M}\tilde{\text{V}}\text{N}(\mathtt{0},\mathbf{\Sigma})$	$S_j\left(t\vert\mathtt{X}_j\right)=S_0(t)^{\exp\left('\mathtt{X}_j+\mathtt{b}_j'\mathtt{Z}_j\right)}$	13,29,79,80,83,85,174	5.3
Accelerated Failure Time	Stratified baseline	$h_{0j}(t\exp\{'\mathtt{X}\})\exp('\mathtt{X})$	$S_j\left(t\vert\mathtt{X}_j\right)=S_{0j}(t\exp('\mathtt{X}))$	16	5.4
	Shared frailty	$h_0\left(t\eta_j\exp\{'\mathtt{X}\}\right)\eta_j\exp('\mathtt{X})$ where η_j~Gamma(θ) or $log(\eta_j)$~Normal$(0,\tau^2)$	$S_j\left(t\vert\mathtt{X}_j\right)=S_0\left(t\eta_j\exp('\mathtt{X})\right)$	16,81,83,168	5.5
	Random effects	$h_0\left(t\exp\left\{'\mathtt{X}_j+\mathtt{b}_j'\mathtt{Z}_j\right\}\right)\exp\left('\mathtt{X}_j+\mathtt{b}_j'\mathtt{Z}_j\right)$ where $\mathtt{b}_j\text{M}\tilde{\text{V}}\text{N}(\mathtt{0},\mathbf{\Sigma})$	$S_j\left(t\vert\mathtt{X}_j,\mathtt{Z}_j\right)=S_0\left(t\exp\left('\mathtt{X}_j+\mathtt{b}_j'\mathtt{Z}_j\right)\right)$	16,81,83,168	5.6

In the Cox Proportional Hazards model, the baseline hazard $h_0(t)$ is left unspecified. For the baseline hazard of the parametric models, see Table 1.

[14,16,24,81,83,168] or a normal distribution for the log-frailty, as this bears similarity to the generalized linear mixed effects model, [14,64] though many other distributions including the inverse Gaussian, positive stable, and compound Poisson are possible. [14,16,17] Previous studies have demonstrated that the gamma frailty model appears to be fairly robust against misspecification of the frailty distribution, [65,66] that it describes the frailty of survivors for a large class of hazard models, [24] and that it can have more power than a stratified model. [16,66,67] Therefore, frailty models are generally recommended when the number of participants per trial is very low. Yet, when the number of participants per trial is large, as is often the case in meta-analysis when individual trials are designed to have sufficient power to test for an intervention effect, the frailty and stratification approaches will usually yield similar results, given that the assumptions are met.

When a frailty is applied to the baseline hazard, the median hazard ratio (MHR) can be used to evaluate the meaning of this frailty in the context of the different studies. [68-70] The MHR is the median relative difference in the hazard of the occurrence of the outcome when comparing identical participants from two randomly selected studies ordered by hazard. When a log-normal distribution is assumed for the frailty, the Median Hazard Ratio (MHR) can be computed as $\exp\left\{\sqrt{2\sigma^2}\Phi^{-1}(0.75)\right\}$, where Φ^{-1} is the inverse of the standard normal distribution. [69,70]

4.3.4 | Marginal models

In the analysis of clustered data, such as IPD from different studies, where the interest lies in the average intervention effect for the target population as a whole, we may use marginal models. In such models the dependence between participants from the same trial is not modeled explicitly but standard errors are adjusted for it. [71,72] Intervention effects are interpreted as relative to participants drawn randomly from the entire target population from which the participants are considered to be sampled. [73] When the interest lies in the intervention effect of participants in the individual studies or in the causes of heterogeneity of intervention effects across studies or subgroups, as in an IPD-MA often is the case, conditional models are needed. [74]

4.3.5 | Estimation

Maximum Likelihood (ML) estimates of the mixed effects Cox model may be obtained with a Newton–Raphson procedure, [75] with penalization methods by constraining the frailty terms with a penalty, [76-78] by expectation-maximisation, [79] or by expectation-maximisation and penalization. [80]

Further, residual maximum likelihood (REML) estimates of the mixed effects Cox model can be obtained with a Newton–Raphson procedure, [12,13,75] or with penalization methods by constraining the frailty terms with a penalty. [76,77] As the penalized method does not take uncertainty of τ^2 into account, it has been suggested that it produces less precise estimates of the intervention effect. [62] However, comparative evidence is currently lacking.

Alternatively, the mixed effects Cox model can be estimated with a poisson model, [81] where the time-scale is split into intervals defined by event times. [82] Mixed effects parametric models can be estimated with Maximum Likelihood by adaptive Gauss-Hermite quadrature. [83] Mixed effects Weibull models can also be estimated with REML. [81]

The Bayesian framework allows for the estimation of a wide range of time-to-event models. For instance, the Cox random effects model can be estimated using Bayesian methods. [16,84,85] A random trial effect and an intervention by trial interaction may be evaluated simultaneously in a Bayesian Cox PH model. [86] For a discussion of commensurate priors for incorporating between-trial variability in a Bayesian meta-analysis, see [87]. Finally, an overview of software for the estimation of one-stage time-to-event models is given in Table 6.

4.3.6 | Heterogeneity of the intervention effect in the one-stage approach

Similar to the two-stage approach, we may expect heterogeneity of the intervention effect in the one-stage approach, which makes the common effects assumption untenable. As such, it is also recommended for one-stage models to assume random effects (Equation 5.3, Table 5), [79] and to investigate the causes of this heterogeneity, if present. [12] One possible cause of heterogeneity of the intervention effect is effect modification (ie, interaction) at the individual level, which can be investigated by adding an interaction term in the one-stage model. [29] Crucially, when including such an interaction term (eg, an intervention-covariate interaction) in the one-stage approach, special care must be taken to avoid the amalgamation of within- and across-trial information, as this may lead to ecological bias. This can be achieved by centering the covariates by their mean values within trials, such that the interaction estimate is then only based on within-trial information. [88] To improve the estimation of between-study variance and the coverage of confidence intervals, the

TABLE 6 Software for One-stage Time-to-event Models

Program	Package/method	Description	Code in	Mentioned in
R, S-Plus	-	Random effects Cox model	80	
	survival	Cox and parametric time-to-event models.	66, 77, 175, 176	
		Stratified, frailty and marginal specifications		
	coxme	Mixed effects Cox models		
	frailtypack	Cox and parametric random effects and stratified models.		63, 78, 111
		Correlated random effects. Competing events. Joint nested frailty models.		
	SemiCompRisks	Bayesian and frequentist random effects parametric and		111
		semi-parametric models for competing events.		
	parfm	Parametric frailty models		
	PenCoxFrail	Regularized Cox frailty models		
	mexhaz	Flexible (excess) hazard regression models,		
		non-proportional effects, and random effects		
	dynfrail	Semiparametric dynamic frailty models		
	frailtyEM	Frailty models with semi-parametric baseline hazard, recurrent events		
	joineR	Joint random effects models of repeated measurements & time-to-event		
	joint.Cox	Joint frailty-copula models with smoothing splines		
	JointModel	Joint model for longitudinal and time-to-event outcomes		
	joineRML	Joint time-to-event and multiple continuous longitudinal outcomes		
	rstanarm	Joint model for hierarchical longitudinal and time-to-event data	131	
	surrosurv	Time-to-event surrogate endpoints models	177	
SAS	PHREG	Cox models, including stratification or frailty	66, 175	178
	NLMIXED	Mixed effects parametric survival models		179
		Joint model for recurrent events and semi-competing risk	112	
	GENMOD	Poisson regression, marginal models		178
Stata	stcox	Cox model, stratified and frailty specifications.		
	stmixed	Flexible parametric time-to-event models with mixed effects		[7,83]
	xtmepoisson	Mixed effects Poisson regression	82	
WinBUGS, OpenBUGS, JAGS	-	Bayesian mixed effects models,	67, 82, 175	
	-	IPD network meta-analysis	118, 121	
MLwiN	-	Mixed effects time-to-event models		7, 180
The Survival Kit	-	Bayesian mixed effect time-to-event models		86

intervention variable can be centered within studies as well. To further prevent the borrowing of information across studies that may affect the estimate of the intervention effect in the one-stage approach, a covariate by trial indicator interaction can be included. This stratifies the covariates effects as it allows covariate effects to be estimated separately for each study (see Table 3).

When there are differences in follow-up time between trials and the intervention effect changes over time, the estimated intervention effects (as quantified by random effects) will be different per trial. If this is unaccounted for, this will lead to heterogeneity of the intervention effect. This can then be investigated by modeling the effect as time-dependent (section 5.2).

In the two-stage approach the influence of trial-level characteristics on the intervention effect can be estimated with meta-regression in the second stage. In the one-stage approach it is possible to simultaneously estimate the heterogeneity of baseline rate of the participants within different studies, the heterogeneity of intervention effects and their correlation. [78]

5 | EXTENSIONS

5.1 | Modeling the baseline hazard function

Whereas the Cox PH model leaves the baseline hazard unspecified, we may apply a parametric model by specifying a baseline hazard (Table 1), either in the first stage of the two-stage approach, or within the one-stage approach. To allow for flexible shapes of the baseline hazard, we can apply spline functions. Particularly the approach of Royston and Parmar is useful, where the baseline cumulative hazard is modelled using restricted cubic splines, [89] and which has been extended to allow for random effects. [83]

Parametric models are especially suitable when absolute (rather than relative) risks for individual subjects (rather than for subpopulations) are of primary interest. It leads to smooth predicted survival curves and is well suited to deal with non-proportionality of hazards. For instance, researchers increasingly often aim to develop prediction models that can assess individual intervention benefits (or harms). [90] Most simply, one can specify an exponential (Equation 1.2) or a Weibull (Equation 1.3) distribution within the proportional hazards framework. The exponential distribution assumes a constant rate over time, whereas the Weibull distribution (a generalization of the exponential distribution) allows for accelerated failure times (AFT). [14] Other (but less common) generalizations of the exponential distribution that can be used for modeling the baseline hazard are the Gompertz, gamma, and piecewise constant distributions. [14,83] Further, the log-logistic, log-normal and generalized gamma distributions may be used. [83,89] Unlike PH models, the estimate of an intervention effect in AFT models is unaffected by unmeasured prognostic covariates. [21] Also in one-stage models a wide range of distributions for parametric PH and AFT models is available. [83]

5.2 | Modeling non-proportional hazards

For short trials with a low event rate the proportionality of hazards across time may be reasonable (ie, the hazard ratio for the intervention effect may be assumed constant over time), but as the number of events in different intervention groups diverges a selection of participants remains in the trial for whom proportionality in the unadjusted intervention effect is not realistic. [91,92] If an intervention is protective, frail participants in the intervention group will be better protected against the outcome than frail participants in the control group. Hence, the proportion of frail participants at risk will decrease more quickly in the control group than in the intervention group. To account for this issue within studies we can include covariates in the model, whereas we can use a frailty model to account for this issue between studies.

Non-proportionality of hazards may also be present due to the intervention effect truly being dependent on time. For instance, an intervention (such as surgery or chemotherapy) may cause an increased risk of a negative outcome at first, but have a protective effect in the long run. This can be modeled by an interaction effect between the intervention (or a covariate) and time [18] in the one-stage approach or in the first stage of the two-stage approach. To allow for flexible shapes of this time-dependent effect, fractional polynomials or splines can be applied. [93-95]

Two methods have been developed for combining fractional polynomials or splines in the two-stage approach. The meta curve method directly meta-analyzes the curves estimated in the first stage. Though, this requires setting a reference level which may have an impact on the results. Alternatively, by using multivariate meta-analysis (section 5.3) the coefficients can be combined. This method only works when the same polynomials or splines have been fitted in each study, but that is not an issue when IPD are available. [96]

Alternatively, non-PH can sometimes be handled more naturally with models that assume proportionality on another scale. [89,92] For instance, an intervention might temporarily reduce the hazards, but as time progresses and the effect wears off, hazards converge and thereby violate the proportional hazards assumption. This can be modeled with a proportional odds regression model such as the log-logistic (Equation 1.7, Table 1), which assumes that covariates have a constant additive effect on the log odds of survival. [27,97-99] In this model, the modeled hazard ratio naturally approaches 1 over time, whereas the odds remain proportional. [98]

As the implementation of TTE models with non-proportional hazards (eg, with splines) may complicate the interpretation of regression parameters, alternate effect measures have been proposed to summarize intervention effects (Table 7). For instance, the restricted mean survival time (RMST, Equation 7.3) until time t^* represents the area under the survival curve until time t^*. [100-102] The RMST can thus be calculated for different intervention groups, and subsequently be subtracted to assess the intervention effect. This difference represents the expected gain (or loss) in survival until time t^* for the intervention group, as compared to the control group. An advantage is that it provides a clinically meaningful summary of the survival differences between intervention groups.

The percentile ratio, an effect measure alternative to the more common hazard ratio, was suggested by to make the interpretation of survival models more straightforward. [103] Briefly, the percentile ratio for an intervention is defined as the expected ratio for the time at which a certain fraction (given as 'k') of the participants will have an event in the intervention group as compared to the control group (Equation 7.4). The percentile ratio is easiest to interpret for AFT models, as the percentile ratio does not depend on the percentile chosen in such models and always equals the acceleration factor. Two-stage MA methods for the percentile-ratio have also been developed. [104]

5.3 | Modeling multiple outcomes

Throughout this manuscript, we have assumed that each patient in each trail is at risk of having a single type of event (ie, the outcome of interest, for example, all-cause mortality), until censoring takes place. Alternatively, patients may be at risk for different events, where one event (eg, death) prevents the patient from having another event (eg, liver failure or stroke). Unlike the survival function, relative intervention effects can then still be assessed by modeling cause-specific hazards, which involves the modeling of the time to each type of event in a separate model, where all alternative types of event are coded as censoring. [105,106] It is vital to do this for every type of event, to gain a full understanding of the relative intervention effect with respect to competing events. [105] Whereas for all-cause-mortality there is a direct relation between the hazard and the survival curve, when modeling cause-specific hazards this is not the case, [107] meaning that this approach does not have a direct interpretation in terms of absolute survival probabilities for the outcome of interest. [108] Only when independence of the event of interest and the competing event can be assumed, the survival function can be estimated by recoding the competing outcome as censoring, though this assumption is often not realistic. [109]

Therefore, when prediction of the average time-to-event per intervention group is wanted, competing events must be modeled using more complex survival models (for an introduction see [105,110]). In the two-stage approach, this can be analyzed with competing risk models in the first stage, whereas Bayesian hierarchical competing risk models have been developed for the one-stage approach, [111] which may also model recurrent events jointly with the competing risk. [112] Further, multi-state models can be used to model transitions to intermediate events. [113]

When multiple outcomes that do not compete are available across trials, these can be assessed jointly in the two-stage framework to improve the efficiency of the analyses. [114,115] For instance, outcomes may have been assessed at multiple follow-up times, or be defined for multiple endpoints. In the first stage, estimates of the intervention effects and variances are obtained for each outcome in each trial. Bootstrapping is used to obtain the covariance between intervention effects for each pair of outcomes in the same trial. [115] In the second stage, the vectors of estimates (and matrices of variances and covariances) are synthesised using a multivariate meta-

TABLE 7 Effect Measures for Time-to-Event Analysis

Measure	Definition		Ref.	No.
Hazard ratio	$\frac{\lambda(t\|X_1)}{\lambda(t\|X_0)}$,	$\lambda(t\|X_k) = -\frac{d\ln(S(t\|X_k))}{dt} = \frac{f(t\|X_k)}{S(t\|X_k)}$	14,15	7.1
Odds ratio	$\frac{O(t\|X_1)}{O(t\|X_0)}$,	$O(t\|X_k) = \frac{1-S(t\|X_k)}{S(t\|X_k)}$	27	7.2
RMSTD(t^*)	$\text{RMST}_1(t^*) - \text{RMST}_0(t^*)$,	$\text{RMST}(t^*) = \int_0^{t^*} S_k(t)dt$	100,102	7.3
Percentile Ratio	$q_k = \frac{k^{\text{th}}\text{percentileofdistforgroupA}}{k^{\text{th}}\text{percentileofdistforgroupB}}$		103	7.4

RMST = Restricted Mean Survival Time, D = Difference.

analysis model in the second stage. Hence, multivariate meta-analysis is particularly relevant to address outcomes or time-points in the IPD from some trials.

5.4 | Modeling multiple interventions

The concepts of multivariate meta-analysis can also be used to compare more than two interventions. In a so-called network meta-analysis (NMA), direct and indirect evidence about the difference in effect of two or more treatments is combined across trials, to summarize the relative effects of all available interventions. This may improve precision of the intervention estimates and allows for comparison of interventions that have not been compared head-to-head. This method uses direct evidence (intervention effects estimated within trials) and indirect evidence (intervention effects estimated across trials), by assuming that both sources of evidence are exchangeable. [116,117]. When direct and indirect evidence disagree, the network is said to be inconsistent and may be prone to bias or may cause heterogeneity of the estimated intervention effects. Such inconsistency can be caused by effect modification, which can be addressed by modelling interactions between the intervention and patient-level covariates. [118]

In the two-stage approach, an appropriate (eg, Cox) survival model is first estimated in each trial, possibly adjusting for relevant prognostic factors and effect modifiers. Corresponding effect estimates (eg, log hazard ratios) can then be pooled using traditional NMA methods. [116] In the one-stage approach, time-to-event NMA models can be estimated using Bayesian hierarchical models. [119,120] Also, Bayesian one-stage IPD-NMA Royston-Parmar models have been implemented. [121]

5.5 | Surrogate endpoints

Trials for measuring intervention efficacy tend to be expensive and require a lengthy follow-up to observe the clinical outcome. The cost and duration of a trial may be reduced if a more readily available outcome can be used. Validated surrogate endpoints can be used instead when the surrogate is well known or likely to predict clinical outcome. [122] These surrogate endpoints are to be validated on the trial and the participant level, where IPD form multiple trials are preferred. [123,124] When response to intervention is used to predict survival, response must be modeled as a time-dependent covariate or a landmarking method must be used. [125] Alternatively, a joint model with the survival outcome and a continuous surrogate or a dichotomous surrogate can be used. [126,127] For an overview and comparison of the performance of measures of surrogacy, see [128,129]. When few trials are available, the trial level surrogacy cannot reliably be estimated using AD alone. However, surrogacy can sometimes be estimated on the center level by splitting multicenter data by center. [124,130] This requires IPD when center specific parameter estimates are not available. For a recent overview of methods for estimating surrogacy, see [130]. To include a surrogate directly in the modeling of the outcome, a joint model can be used. [126,127] For the one-stage approach, joint models with up to three levels have also been developed. [131]

5.6 | Missing data

In a meta-analysis of survival data, several types of missing data may occur. It is possible, for instance, that not all studies provide IPD and thus that only AD are available for some of the studies. In such cases, it is recommended to combine the available IPD and AD, as otherwise estimated intervention effects may be prone to (data availability) bias and overly large standard errors. [132] Including AD in a two-stage meta-analysis approach is fairly straightforward, provided that the model used for generating the AD is compatible with the models for analyzing the available IPD. It is also possible to directly combine IPD and AD using a one-stage meta-analysis, although this requires more advanced models, such as Bayesian hierarchical regression. [133]

Another common type of missing data occurs when events of individual subjects are censored, for example, due to loss of follow-up. Survival models such as the Cox PH model and the AFT model readily account for this censoring, provided that it is not related to the outcome, conditional on any participant-level characteristics in the model (ie, non-informative). When the assumption of independent censoring is challenged, its implications can be evaluated by adopting multiple imputation methods. [134]

Finally, it is possible that subject-level covariates are missing for one or more studies. Although participant covariates are not commonly used when estimating relative intervention effects from RCTs, they are crucial in IPD-MA of time-to-event data because of selection differences across trials (see section 2). When relevant participant-level covariates are missing for some trial participants, it is generally recommended to apply multiple imputation. [135] Hereby, researchers should adjust for the event indicator and the Nelson-Aalen estimator of the cumulative hazard, [136,137] and also account for the presence of clustering. The latter can be achieved by adopting imputation models with mixed effects, which also facilitates imputation of covariates that have not been measured in one or more studies.[138-142]

Although the assumptions needed for multiple imputation cannot always be tested or may not always be met, several simulation studies have shown that its use is usually superior to complete-case analysis or the use of missing data indicators.[143] However, caution is still warranted when analyzing imputed data sets from IPD-MA, as in the presence of between-trial heterogeneity these are inherently prone to some degree of incompatibility with the data generation mechanism.[141,144] Further, because IPD-MA can only adjust for measured covariates and may therefore still be affected by unmeasured covariates, clustering of participants within trials should still be accounted for (section 4.3.1).[14]

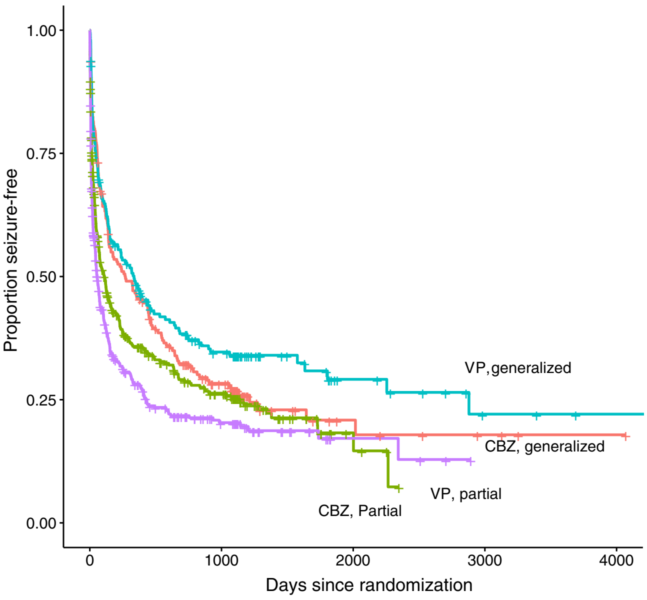

FIGURE 2 Kaplan–Meier plot of Generalized and Partial Epileptic Seizure Patients Treated with Carbamazepine (CBZ) or Valproate (VP) [Colour figure can be viewed at wileyonlinelibrary.com]

6 | APPLIED EXAMPLE

The efficacy of carbamazepine (CBZ) and valproate (VP) as interventions for epileptic seizures was compared in a systematic review and IPD-MA of RCTs.[145] IPD were obtained for a total of 1225 participants from five trials. In all these trials, one of the outcomes of interest was time to first epileptic seizure since randomization. Also, measured covariates were age at randomization, sex, type of epilepsy (partial-onset or generalized-onset), and the number of epileptic seizures before randomization. For illustrative purposes, we only consider the type of epilepsy. We use the `coxme` package of the **R** software,[146,147] to fit the mixed effects Cox PH model. Our code is given in Supporting Information 2.

As the two-stage method has been described extensively (see [104,148]) we shall restrict our analyses to illustrate some key one-stage methods. First, to evaluate the relative effects of CBZ and VP, we adopted a Cox model, as this leaves the baseline hazard unspecified. We apply a one-stage model (Equation 5.2) with a log-normal frailty and random effects for the intervention estimated with penalized partial likelihood to account for the clustering of participants within trials and to allow for heterogeneous intervention effects across trials, respectively. We find no evidence against the hypothesis that the interventions are equally effective, with a summary hazard ratio of 1.08 for valproate $(95\ \%\ \text{Confidence Interval (CI)}: 0.92\ \text{to}\ 1.27,\ p = .37)$, vs the referent, carbamazepine.

In the analysis of the effect of the intervention on the time to first epileptic seizure, we observed some statistical heterogeneity of the intervention effect. The SDs of the random intercept (ie, frailty) and drug effect (ie, random effect) equaled 0.139 and 0.099, respectively. In other words, the log hazard ratio of valproate vs carbamazepine varied with a SD of.099 between trials. This random effect of the interventions translated to a Median Hazard Ratio (MHR) of 1.10, meaning that the median relative change in the effect on time-to first epileptic seizure when comparing two identical participants from two randomly selected different trials that were ordered by intervention effect was 1.10, calculated as $\exp\left\{\sqrt{2}0.099\Phi^{-1}(0.75)\right\}$ (see section 4.3.6). In order to explain this heterogeneity in intervention effect, we added covariates and intervention-covariate interactions to the model (Figure 2). Partial epilepsy (vs generalized) was associated with a higher hazard rate ($\beta = 1.63$, 95% CI: 1.38 to 1.92, Table 8), meaning that we have found evidence that epilepsy type is a prognostic factor of time to first epileptic seizure. However, we were unable to find evidence that epilepsy type interacted with the intervention ($\beta = 1.36$, 95% CI: 0.97 to 1.89), though it should be noted that the upper bound of the CI did not exclude clinically significant effects. We note that we obtained somewhat different results than the Cochrane review,[149] as we have used a different method for analysis. Further, the low power for tests for interaction effects is a notorious issue.

A recent investigation of the intervention-covariate interaction on the time to remission of epilepsy demonstrated that bias occurs when within-trial and across-trial information is not separated.[88] Such separation can be performed by centering the covariates, hence we have centered the covariates in in our analysis (Table 8). The possible bias that may occur when within-trial and across-trial information are amalgamated can be

TABLE 8 Intervention, Covariates and Intervention-Covariate Interactions in a Multivariable Mixed Effects Cox Model

Variable	Variable Type	HR	95% CI	*p*
VP (vs CBZ)	Intervention	1.05	0.86 to 1.28	0.65
Partial epilepsy (vs generalized), centered[a]	Individual-level covariate	1.63	1.38 to 1.92	< .001
Partial epilepsy (vs generalized), trial mean[b]	Trial-level covariate	1.47	0.99 to 2.19	0.06
Partial epilepsy (vs generalized), centered[a] * VP (vs CBZ)	Intervention-covariate interaction	1.36	0.97 to 1.89	0.07

VP: Valproate, CBZ: Carbamazepine, HR: Hazard ratio, given by $exp(\beta)$, CI: Confidence interval. Standard deviations of random intercept (ie, frailty) and random effect of VP (vs CBZ) equal 0.126 and 0.164, respectively. *P*-values are for Wald type tests of the null hypothesis that the log HR equals zero.

[a] Covariates are centered within trials, to avoid ecological bias (see [88]).

[b] Trial mean value for the covariate is entered in the analysis, to quantify the bias that would occur if centering of the covariate were not performed.

quantified by including the trial-mean in the model, [88] as we have done here (Table 8).

7 | DISCUSSION

Our search has identified a wide range of articles on topics regarding TTE IPD-MA, and is the first comprehensive review on this topic to our knowledge. However, the basics of the methodology regarding TTE data was excluded from our search as it did not concern MA or clustered data. Covering all methodological works regarding TTE data would have been an immense task. As such, we were forced to include relevant literature based on our own opinion to introduce this topic, and restrict our systematic search through Pubmed and Web of Science to works that simultaneously concerned IPD, meta-analysis and time-to-event data. We did not cover every article that covers these three topics, as this was not our aim. Instead, we our purpose was to achieve theoretical saturation, that is, that an extended search would be unlikely to add important information.

The general consensus in the reviewed works was that the Cox model should be the default model of choice for TTE IPD-MA. Though, it is also criticized for not yielding a valid estimate of intervention effect when not all (un-)measured predictive covariates are accounted for, mostly on theoretical grounds. The literature is currently missing information on the impact of this issue in real life data, leading us to suggest that further research should focus thereon. As such, we have provided a comprehensive review of current methods for IPD-MA of TTE data.

Although the statistical properties of the meta-analysis estimators for the two-stage approach have been well studied and simulation studies have investigated the performance for meta-analysis of dichotomous and continuous outcome data, this is not the case for time-to-event data. Further, although aggregate data (ie, estimates from the literature) can readily be included in the two-stage approach (provided that the models are specified the same), as well as in Bayesian one-stage models, there appears to be no method yet for doing so in a Frequentist model.

Another issue is to what extend one should try to borrow information across trials in the one-approach. In the two-stage approach, no information is borrowed (apart from the intervention effect and its uncertainty), as all parameters are naturally estimated per trial. To what extend one should account for this in the one-stage approach, by stratifying the baseline and covariate effects or by applying random effects and a frailty, deserves extra attention in the literature. For the meta-analysis of trials with adequate sample sizes, the safest choice is to stratify all included parameters as this accounts for all differences in baselines between trials. In a simulation study where IPD from a total of 600 participants from 3-20 trials were generated, both the frailty and the stratified baseline method worked well, [29] though exactly what sample sizes are necessary for this strategy, and especially for the stratification of covariates as well, has apparently not yet been identified.

8 | CONCLUDING REMARKS

We have discussed numerous models in this manuscript, the choice between which is not always straightforward. For this reason, we provide some recommendations below. First, intervention effect conditional on covariates and/or frailties have different interpretations from marginal ones (ie, averaged over the entire sample and follow-up time), and yield different estimates. Before embarking on an IPD-MA, researchers should decide whether a conditional or a marginal effect is of interest. As assumptions may be satisfied on one scale but not the other, this may lead to a different choice of model.

Additionally, one can choose between one-stage and two-stage models. In the two-stage method participants within trials are compared, which inherently yields a conditional intervention effect and stratified baselines. The one-stage approach offers more possibilities as it allows for conditional intervention effects as well as marginal ones, and frailties for the baseline. When the same (or similar) model assumptions are made for these models and the same estimation methods are used, these two approaches generally lead to the same estimates of intervention effect. [13,28] Though, the one-stage approach can have better convergence properties when the included studies are very small, [28,150] or at least one of the studies has zero events.

Further, when a conditional effect is desired (in contrast to a marginal one), we recommend to apply random effects instead of common effects, as common effects models are only valid when no heterogeneity is present, which is unlikely in our experience. When a marginal effect is desired, only a correction for the variance is necessary. As described in section 2, when an intervention effect is present the estimated intervention effect in PH models may be time-dependent, depending on the distribution of prognostic factors that are not accounted for (even if balanced across intervention groups). This may lead to heterogeneity in intervention effects across trials that have different follow-up lengths. Further, differences in trial design and methodology and clinical procedures may contribute to the heterogeneity of the intervention effect. [12] Random effects models can account for heterogeneity of the intervention effect and lead to the same solution as common effect models when no heterogeneity is present. However, if a formal test of heterogeneity is desired, a variety of tests can be used. For one-stage meta-analysis, the common effect model (without trial effects) is nested in the frailty model, and therefore a comparison of these models can be made using the log-likelihood ratio test. [14] Alternatively, a score test, [151-153] or a small sample test can be used. [154] A permutation test for testing of the presence of heterogeneity in time-to-event data was recently proposed, and a simulation showed that the method is more powerful and has a better type I error rate than likelihood ratio tests of a random effect. [155]

Finally, when comparing non-nested (eg, PH vs AFT) models, more general methods are needed. In such cases, one may select the model with lowest value for Akaike's Information Criterion (AIC) [156,157] or the Bayesian Information Criterion (BIC) [156-158]. Though, due to the correlated nature of participants within trials a correction for clustering should be made, which is not straightforward in the frequentist estimation framework as quantification of the number of degrees of freedom is difficult. For subject-specific inferences, the conditional (cAIC) can be used, whereas for inferences on the population level the marginal AIC can be used. [64,74,159,160]

Further, one should be cautious regarding model selection. If one model is rejected, bias will appear in the estimated intervention effect and significance in a second model if the second model is not independent of the test that was used to reject the first, such as when a non-PH effect is included in the model after a statistical test indicated non-proportionality. [161] This bias can be alleviated by bootstrapping the model selection procedure. On the other hand, this bias does not occur when the second model is independent of the test used to reject the first model. [161]

ACKNOWLEDGEMENTS

This work is financially supported by the Netherlands Organization for Health Research and Development grant 91617050 for TD and grant 91810615 for VJ and KM, and the European Union's Horizon 2020 Research and Innovation Programme under ReCoDID Grant Agreement no. 825746 for VJ and KM. The data that support the findings of this study are not publicly available, according to the conditions determined by the Epilepsy Monotherapy Trial Group, but are available on request from AM, by e-mailing a.g.marson@liverpool.ac.uk. We would like to thank the editor and reviewers for thoughtful comments on the manuscript and suggesting items for the review.

CONFLICT OF INTEREST

The author reported no conflict of interest.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are not publicly available, according to the conditions determined by the Epilepsy Monotherapy Trial Group, but are available on request from AM, by e-mailing A.G.Marson@liverpool.ac.uk.

ORCID

Valentijn M.T. de Jong https://orcid.org/0000-0001-9921-3468
Richard D. Riley https://orcid.org/0000-0001-8699-0735
Marinus J.C. Eijkemans https://orcid.org/0000-0001-9400-0615
Thomas P.A. Debray https://orcid.org/0000-0002-1790-2719

REFERENCES

1. Riley RD, Steyerberg EW. Meta-analysis of a binary outcome using individual participant data and aggregate data. *Res Synth Methods*. 2010;1(1):2-19.
2. Stewart LA, Parmar MKB. Meta-analysis of the literature or of individual patient data: is there a difference? *The Lancet*. 1993;341(8842):418-422.

3. Tierney JF, Vale C, Riley R, et al. Individual participant data (IPD) meta-analyses of randomised controlled trials: guidance on their use. *PLoS Med*. 2015;12(7):e1001855.
4. Lyman GH, Kuderer NM. The strengths and limitations of meta-analyses based on aggregate data. *BMC Med Res Methodol*. 2005;5:14.
5. Tudur Smith C, 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. *J Eval Clin Pract*. 2005;11(5):468-478.
6. Higgins JPT, Whitehead A, Turner RM, Omar RZ, Thompson SG. Meta-analysis of continuous outcome data from individual patients. *Stat Med*. 2001;20(15):2219-2241.
7. Debray TPA, Moons KGM, van Valkenhoef G, et al. Get real in individual participant data (IPD) meta-analysis: a review of the methodology. *Res Synth Methods*. 2015;6(4):293-309.
8. Thomas D, Platt R, Benedetti A. A comparison of analytic approaches for individual patient data meta-analyses with binary outcomes. *BMC Med Res Methodol*. 2017;17:28.
9. Jackson D, Law M, Stijnen T, Viechtbauer W, White IR. A comparison of seven random-effects models for meta-analyses that estimate the summary odds ratio. *Stat Med*. 2018;37(7):1059-1085.
10. Legha A, Riley RD, Ensor J, Snell KI, Morris TP, Burke DL. Individual participant data meta-analysis of continuous outcomes: a comparison of approaches for specifying and estimating one-stage models. *Statistics in Medicine*. 2018;37:4404-4420.
11. Riley RD, Kauser I, Bland M, et al. Meta-analysis of randomised trials with a continuous outcome according to baseline imbalance and availability of individual participant data. *Stat Med*. 2013;32(16):2747-2766.
12. Tudur Smith C, Williamson PR, Marson AG. Investigating heterogeneity in an individual patient data meta-analysis of time to event outcomes. *Stat Med*. 2005;24(9):1307-1319.
13. Bowden J, Tierney JF, Simmonds M, Copas AJ, Higgins JP. Individual patient data meta-analysis of time-to-event outcomes: one-stage versus two-stage approaches for estimating the hazard ratio under a random effects model. *Res Synth Methods*. 2011;2(3):150-162.
14. Wienke A. *Frailty Models in Survival Analysis*. Boca Raton, FL: CRC Press; 2011.
15. Klein JP, van Houwelingen HC, Ibrahim JG. *Scheike TH*. Handbook of Survival Analysis: Chapman and Hall/CRC; 2013.
16. Duchateau L, Janssen P. *The Frailty Model*. New York, NY: Springer-Verlag; 2008.
17. Hougaard P. *Analysis of Multivariate Survival Data*. New York, NY: Springer-Verlag; 2012.
18. Cox DR. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*. 1972;34(2):187-220.
19. Vaupel JW, Manton KG, Stallard E. The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography*. 1979;16(3):439-454.
20. Gail MH, Wieand S, Piantadosi S. Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates. *Biometrika*. 1984;71(3):431-444.
21. Hougaard P. Fundamentals of survival data. *Biometrics*. 1999;55(1):13-22.
22. Lin NX, Logan S, Henley WE. Bias and sensitivity analysis when estimating treatment effects from the Cox model with omitted covariates. *Biometrics*. 2013;69(4):850-860.
23. Hougaard P. Modelling Heterogeneity in Survival Data. *Journal of Applied Probability*. 1991;28(3):695-701.
24. Abbring JH, van den Berg GJ. The unobserved heterogeneity distribution in duration analysis. *Biometrika*. 2007;94(1):87-99.
25. Hernández AV, Eijkemans MJC, Steyerberg EW. Randomized controlled trials with time-to-event outcomes: how much does Prespecified covariate adjustment increase power? *Ann Epidemiol*. 2006;16(1):41-48.
26. Hjort NL. On inference in parametric survival data models. *Int Stat Rev*. 1992;60(3):355-387.
27. Royston P, Parmar MKB. Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. *Stat Med*. 2002;21(15):2175-2197.
28. Burke DL, Ensor J, Riley RD. Meta-analysis using individual participant data: one-stage and two-stage approaches, and why they may differ. *Stat Med*. 2017;36(5):855-875.
29. Katsahian S, Latouche A, Mary JY, Chevret S, Porcher R. Practical methodology of meta-analysis of individual patient data using a survival outcome. *Contemp Clin Trials*. 2008;29(2):220-230.
30. Abo-Zaid G, Guo B, Deeks JJ, Debray TPA, Steyerberg EW, Moons KGM, Riley RD Individual participant data meta-analyses should not ignore clustering. *Journal of Clinical Epidemiology*. 2013;66(8):865–873.e4.
31. Riley RD, Lambert PC, Staessen JA, et al. Meta-analysis of continuous outcomes combining individual patient data and aggregate data. *Stat Med*. 2008;27(11):1870-1893.
32. Firth D. Bias reduction of maximum likelihood estimates. *Biometrika*. 1993;80(1):27-38.
33. Heinze G, Schemper M. A solution to the problem of monotone likelihood in Cox regression. *Biometrics*. 2001;57(1):114-119.
34. Lambert PC, Sutton AJ, Abrams KR, Jones DR. A comparison of summary patient-level covariates in meta-regression with individual patient data meta-analysis. *J Clin Epidemiol*. 2002;55(1):86-94.
35. Localio AR, Berlin JA, Ten Have TR, Kimmel SE. Adjustments for center in multicenter studies: an overview. *Ann Intern Med*. 2001;135(2):112-123.
36. Aalen OO. Effects of frailty in survival analysis. *Stat Methods Med Res*. 1994;3(3):227-243.
37. Whitehead A, Whitehead J. A general parametric approach to the meta-analysis of randomized clinical trials. *Stat Med*. 1991;10(11):1665-1677.
38. Hartung J. An alternative method for meta-analysis. *Biom J*. 1999;41(8):901-916.
39. Sidik K, Jonkman JN. Robust variance estimation for random effects meta-analysis. *Computational Statistics & Data Analysis*. 2006;50(12):3681-3701.
40. Knapp G, Hartung J. Improved tests for a random effects meta-regression with a single covariate. *Stat Med*. 2003;22(17):2693-2710.
41. Hartung J, Knapp G. On confidence intervals for the among-group variance in the one-way random effects model with

unequal error variances. *Journal of Statistical Planning and Inference*. 2005;127(1–2):157-177.

42. Jackson D, Law M, Rücker G, Schwarzer G. The Hartung-Knapp modification for random-effects meta-analysis: a useful refinement but are there any residual concerns? *Stat Med*. 2017;36(25):3923-3934.
43. Higgins JPT, Thompson SG, Spiegelhalter DJ. A re-evaluation of random-effects meta-analysis. *J R Stat Soc A Stat Soc*. 2009; 172(1):137-159.
44. Riley RD, Higgins JPT, Deeks JJ. Interpretation of random effects meta-analyses. *BMJ*. 2011;342:d549.
45. Berlin JA, Santanna J, Schmid CH, Szczech LA, Feldman HI. Individual patient- versus group-level data meta-regressions for the investigation of treatment effect modifiers: ecological bias rears its ugly head. *Stat Med*. 2002;21(3):371-387.
46. Thompson S, Kaptoge S, White I, et al. Statistical methods for the time-to-event analysis of individual participant data from multiple epidemiological studies. *Int J Epidemiol*. 2010;39(5): 1345-1359.
47. DerSimonian R, Laird N. Meta-analysis in clinical trials. *Control Clin Trials*. 1986;7(3):177-188.
48. Langan D, Higgins JPT, Simmonds M. Comparative performance of heterogeneity variance estimators in meta-analysis: a review of simulation studies: a review of simulation studies. *Res Synth Methods*. 2017;8(2):181-198.
49. Veroniki AA, Jackson D, Viechtbauer W, et al. Methods to estimate the between-study variance and its uncertainty in meta-analysis. *Res Synth Methods*. 2016;7(1):55-79.
50. Brockwell SE, Gordon IR. A comparison of statistical methods for meta-analysis. *Stat Med*. 2001;20(6):825-840.
51. Sidik K, Jonkman JN. A comparison of heterogeneity variance estimators in combining results of studies. *Stat Med*. 2007;26 (9):1964-1981.
52. Langan D, Higgins JPT, Jackson D, et al. A comparison of heterogeneity variance estimators in simulated random-effects meta-analyses. *Research Synthesis Methods*. 2018;.
53. Ioannidis JPA, Patsopoulos NA, Evangelou E. Uncertainty in heterogeneity estimates in meta-analyses. *BMJ*. 2007;335 (7626):914-916.
54. Viechtbauer W. Confidence intervals for the amount of heterogeneity in meta-analysis. *Stat Med*. 2007;26(1):37-52.
55. Sidik K, Jonkman JN. A simple confidence interval for meta-analysis. *Stat Med*. 2002;21(21):3153-3159.
56. Hartung J, Knapp G. A refined method for the meta-analysis of controlled clinical trials with binary outcome. *Stat Med*. 2001;20(24):3875-3889.
57. 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 Med Res Methodol*. 2014;14 (1):25.
58. Normand SLT. Meta-analysis: formulating, evaluating, combining, and reporting. *Stat Med*. 1999;18(3):321-359.
59. Veroniki AA, Jackson D, Bender R, et al. Methods to calculate uncertainty in the estimated overall effect size from a random effects meta analysis. *Res Synth Methods*. 2019;10(1):23-43.
60. Fisher DJ. Two-stage individual participant data meta-analysis and generalized forest plots. *Stata Journal*. 2015;15 (2):369-396.
61. Tudur Smith C, Williamson PR. A comparison of methods for fixed effects meta-analysis of individual patient data with time to event outcomes. *Clin Trials*. 2007;4(6):621-630.
62. Michiels S, Baujat B, Mahé C, Sargent DJ, Pignon JP. Random effects survival models gave a better understanding of heterogeneity in individual patient data meta-analyses. *J Clin Epidemiol*. 2005;58(3):238-245.
63. Rondeau V, Filleul L, Joly P. Nested frailty models using maximum penalized likelihood estimation. *Stat Med*. 2006;25(23): 4036-4052.
64. Donohue MC, Overholser R, Xu R, Vaida F. Conditional Akaike information under generalized linear and proportional hazards mixed models. *Biometrika*. 2011;98(3):685-700.
65. Glidden DV, Vittinghoff E. Modelling clustered survival data from multicentre clinical trials. *Stat Med*. 2004;23(3):369-388.
66. Munda M, Legrand C. Adjusting for Centre heterogeneity in multicentre clinical trials with a time-to-event outcome. *Pharm Stat*. 2014;13(2):145-152.
67. Johnson B, Carlin BP, Hodges JS. Cross-study hierarchical Modeling of stratified clinical trial data. *J Biopharm Stat*. 1999;9(4):617-640.
68. Lanke J. How to Describe the Impact of the Family-Specific Frailty (Appendix of "Quantifying the Family Frailty Effect in Infant and Child Mortality by Using Median Hazard Ratio (MHR)"). *Historical Methods: A Journal of Quantitative and Interdisciplinary History*. 2010;.
69. Bengtsson T, Dribe M. Quantifying the family frailty effect in infant and child mortality by using median Hazard ratio (MHR). *Historical Methods: A Journal of Quantitative and Interdisciplinary History*. 2010;43(1):15-27.
70. Austin PC, Wagner P, Merlo J. The median hazard ratio: a useful measure of variance and general contextual effects in multilevel survival analysis. *Statistics in Medicine*. 2017;36(6): 928-938.
71. Cai J, Prentice RL. Estimating equations for Hazard ratio parameters based on correlated failure time data. *Biometrika*. 1995;82(1):151-164.
72. Spiekerman CF, Lin DY. Marginal regression models for multivariate failure time data. *J Am Stat Assoc*. 1998;93(443):1164-1175.
73. Lin DY. Cox regression analysis of multivariate failure time data: the marginal approach. *Stat Med*. 1994;13(21):2233-2247.
74. Gardiner JC, Luo Z, Roman LA. Fixed effects, random effects and GEE: what are the differences? *Stat Med*. 2009;28(2):221-239.
75. McGilchrist CA. REML estimation for survival models with frailty. *Biometrics*. 1993;:221–225.
76. Ripatti S, Palmgren J. Estimation of multivariate frailty models using penalized partial likelihood. *Biometrics*. 2000;56 (4):1016-1022.
77. Therneau TM, Grambsch PM, Pankratz VS. Penalized survival models and frailty. *J Comput Graph Stat*. 2003;12(1):156-175.
78. Rondeau V, Michiels S, Liquet B, Pignon JP. Investigating trial and treatment heterogeneity in an individual patient data meta-analysis of survival data by means of the penalized maximum likelihood approach. *Stat Med*. 2008;27(11):1894-1910.
79. Vaida F, Xu R. Proportional hazards model with random effects. *Stat Med*. 2000;19(24):3309-3324.
80. Simmonds MC, Higgins JPT, Stewart LA. Random-effects meta-analysis of time-to-event data using the expectation-

maximisation algorithm and shrinkage estimators. *Res Synth Methods*. 2013;4(2):144-155.

81. Morris C, Christiansen C. Fitting Weibull duration models with random effects. *Lifetime Data Anal*. 1995;1(4):347-359.
82. Crowther MJ, Riley RD, Staessen JA, Wang J, Gueyffier F, Lambert PC. Individual patient data meta-analysis of survival data using Poisson regression models. *BMC Med Res Methodol*. 2012;12:34.
83. Crowther MJ, Look MP, Riley RD. Multilevel mixed effects parametric survival models using adaptive gauss-Hermite quadrature with application to recurrent events and individual participant data meta-analysis. *Stat Med*. 2014;33(22):3844-3858.
84. Clayton DG. A Monte Carlo method for Bayesian inference in frailty models. *Biometrics*. 1991;47:467-485.
85. Sargent DJ. A general framework for random effects survival analysis in the Cox proportional hazards setting. *Biometrics*. 1998;54(4):1486-1497.
86. Legrand C, Ducrocq V, Janssen P, Sylvester R, Duchateau L. A Bayesian approach to jointly estimate Centre and treatment by Centre heterogeneity in a proportional hazards model. *Stat Med*. 2005;24(24):3789-3804.
87. Hobbs BP, Sargent DJ, Carlin BP. Commensurate priors for incorporating historical information in clinical trials using general and generalized linear models. *Bayesian Anal*. 2012;7 (3):639-674.
88. Hua H, Burke DL, Crowther MJ, Ensor J, Tudur Smith C, Riley RD. One-stage individual participant data meta-analysis models: estimation of treatment-covariate interactions must avoid ecological bias by separating out within-trial and across-trial information. *Stat Med*. 2017;36(5):772-789.
89. Royston P, Lambert PC. *Flexible Parametric Survival Analysis Using Stata: beyond the Cox Model*. College Station, Texas, USA: Stata Press; 2011.
90. Kent DM, Steyerberg E, van Klaveren D. Personalized evidence based medicine: predictive approaches to heterogeneous treatment effects. *BMJ*. 2018;363:k4245.
91. van Houwelingen HC, Eilers PHC. Non-proportional hazards models in survival analysis. In: Bethlehem JG, Heijden PGM, eds. *COMPSTAT*. Heidelberg: Physica; 2000:151-160.
92. Kay R, Kinnersley N. On the use of the accelerated failure time model as an alternative to the proportional hazards model in the treatment of time to event data: a case study in influenza. *Drug Inf J*. 2002;36(3):571-579.
93. Hess KR. Assessing time-by-covariate interactions in proportional hazards regression models using cubic spline functions. *Stat Med*. 1994;13(10):1045-1062.
94. Giorgi R, Abrahamowicz M, Quantin C, et al. A relative survival regression model using B-spline functions to model non-proportional hazards. *Stat Med*. 2003;22(17):2767-2784.
95. Thomas L, Reyes EM. Tutorial: survival estimation for Cox regression models with time-varying coefficients using SAS and R. *J Stat Softw*. 2014;61(c1):1-23.
96. White IR, Kaptoge S, Royston P, Sauerbrei W, Collaboration ERF. Meta-analysis of non-linear exposure-outcome relationships using individual participant data: a comparison of two methods. *Statistics in Medicine*. 2019;38(3):326-338.
97. Simmonds MC, Tierney J, Bowden J, Higgins JP. Meta-analysis of time-to-event data: a comparison of two-stage methods. *Res Synth Methods*. 2011;2(3):139-149.
98. Bennett S. Log-logistic regression models for survival data. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*. 1983;32(2):165-171.
99. Bennett S. Analysis of survival data by the proportional odds model. *Stat Med*. 1983;2(2):273-277.
100. 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. *Stat Med*. 2011;30(19):2409-2421.
101. Wei Y, Royston P, Tierney JF, Parmar MKB. 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):2881-2898.
102. Lueza B, Rotolo F, Bonastre J, Pignon JP, Michiels S. Bias and precision of methods for estimating the difference in restricted mean survival time from an individual patient data meta-analysis. *BMC Med Res Methodol*. 2016;16:37.
103. Siannis F, Barrett JK, Farewell VT, Tierney JF. One-stage parametric meta-analysis of time-to-event outcomes. *Stat Med*. 2010;29(29):3030-3045.
104. Barrett JK, Farewell VT, Siannis F, Tierney J, Higgins JPT. Two-stage meta-analysis of survival data from individual participants using percentile ratios. *Stat Med*. 2012;31(30):4296-4308.
105. Beyersmann J, Allignol A. *Schumacher M*. Springer Science & Business Media: Competing Risks and Multistate Models with R; 2011.
106. Noordzij M, Leffondré K, van Stralen KJ, Zoccali C, Dekker FW, Jager KJ. When do we need competing risks methods for survival analysis in nephrology? *Nephrology Dialysis Transplantation*. 2013;28(11):2670-2677.
107. Andersen PK, Geskus RB, de Witte T, Putter H. Competing risks in epidemiology: possibilities and pitfalls. *Int J Epidemiol*. 2012;41(3):861-870.
108. Fine JP, Gray RJ. A proportional hazards model for the subdistribution of a competing risk. *J Am Stat Assoc*. 1999;94 (446):496-509.
109. Satagopan JM, Ben-Porat L, Berwick M, Robson M, Kutler D, Auerbach AD. A note on competing risks in survival data analysis. *Br J Cancer*. 2004;91(7):1229-1235.
110. Geskus RB. *Data Analysis with Competing Risks and Intermediate States*. Chapman and Hall/CRC: Chapman and Hall/CRC; 2015.
111. Lee KH, Dominici F, Schrag D, Haneuse S. Hierarchical models for semicompeting risks data with application to quality of end-of-life care for pancreatic cancer. *J Am Stat Assoc*. 2016;111(515):1075-1095.
112. Jung TH, Peduzzi P, Allore H, Kyriakides TC, Esserman D. A joint model for recurrent events and a semi-competing risk in the presence of multi-level clustering. *Stat Methods Med Res*. 2019;28(10-11):2897-2911.
113. Putter H, Fiocco M, Geskus RB. Tutorial in biostatistics: competing risks and multi-state models. *Stat Med*. 2007;26(11):2389-2430.
114. Riley RD, Thompson JR, Abrams KR. An alternative model for bivariate random-effects meta-analysis when the within-study correlations are unknown. *Biostatistics*. 2008;9(1):172-186.
115. Riley RD, Price MJ, Jackson D, et al. Multivariate meta-analysis using individual participant data. *Res Synth Methods*. 2015;6(2):157-174.

116. Efthimiou O, Debray TP, Valkenhoef G, et al. GetReal in network meta-analysis: a review of the methodology. *Res Synth Methods*. 2016;7(3):236-263.
117. Riley RD, Jackson D, Salanti G, et al. Multivariate and network meta-analysis of multiple outcomes and multiple treatments: rationale, concepts, and examples. *BMJ*. 2017;358:j3932.
118. Saramago P, Chuang LH, Soares MO. Network meta-analysis of (individual patient) time to event data alongside (aggregate) count data. *BMC Med Res Methodol*. 2014;14:105.
119. Veroniki AA, Straus SE, Soobiah C, Elliott MJ, Tricco AC. A scoping review of indirect comparison methods and applications using individual patient data. *BMC Med Res Methodol*. 2016;16:47.
120. Freeman SC, Fisher D, Tierney JF, Carpenter JR. A framework for identifying treatment-covariate interactions in individual participant data network meta-analysis. *Research Synthesis Methods*. 2018;9(3):393-407.
121. Freeman SC, Carpenter JR. Bayesian one-step IPD network meta-analysis of time-to-event data using Royston-Parmar models. *Res Synth Methods*. 2017;8(4):451-464.
122. ICH. *General Considerations for Clinical Trials*. 1997.
123. Shi Q, Sargent DJ. Meta-analysis for the evaluation of surrogate endpoints in cancer clinical trials. *Int J Clin Oncol*. 2009; 14(2):102-111.
124. Renfro LA, Shi Q, Xue Y, Li J, Shang H, Sargent DJ. Center-within-trial versus trial-level evaluation of surrogate endpoints. *Computational Statistics & Data Analysis*. 2014;78:1-20.
125. Buyse M, Piedbois P. On the relationship between response to treatment and survival time. *Stat Med*. 1996;15(24):2797-2812.
126. Schluchter MD, Konstan MW, Davis PB. Jointly modelling the relationship between survival and pulmonary function in cystic fibrosis patients. *Stat Med*. 2002;21(9):1271-1287.
127. Luo S, Su X, DeSantis SM, Huang X, Yi M, Hunt KK. Joint model for a diagnostic test without a gold standard in the presence of a dependent terminal event. *Stat Med*. 2014;33 (15):2554-2566.
128. Shi Q, Renfro LA, Bot BM, Burzykowski T, Buyse M, Sargent DJ. Comparative assessment of trial-level surrogacy measures for candidate time-to-event surrogate endpoints in clinical trials. *Computational Statistics & Data Analysis*. 2011; 55(9):2748-2757.
129. Renfro LA, Shi Q, Sargent DJ, Carlin BP. Bayesian adjusted R2 for the meta-analytic evaluation of surrogate time-to-event endpoints in clinical trials. *Stat Med*. 2012;31(8):743-761.
130. Buyse M, Molenberghs G, Paoletti X, et al. Statistical evaluation of surrogate endpoints with examples from cancer clinical trials. *Biom J*. 2016;58(1):104-132.
131. Brilleman SL, Crowther MJ, Moreno-Betancur M, et al. Joint longitudinal and time-to-event models for multilevel hierarchical data. *Stat Methods Med Res*. 2019;28(12):3502-3515.
132. Poppe KK, Doughty RN, Yu CM, et al. Understanding differences in results from literature-based and individual patient meta-analyses: an example from meta-analyses of observational data. *Int J Cardiol*. 2011;148(2):209-213.
133. Riley RD, Simmonds MC, Look MP. Evidence synthesis combining individual patient data and aggregate data: a systematic review identified current practice and possible methods. *Journal of Clinical Epidemiology*. 2007;60(5):431. e1-431.e12.
134. 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(27):4681-4694.
135. Buuren S, Groothuis-Oudshoorn K. mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*. 2011;45(3).
136. White IR, Royston P. Imputing missing covariate values for the Cox model. *Stat Med*. 2009;28(15):1982-1998.
137. Falcaro M, Nur U, Rachet B, Carpenter JR. Estimating excess hazard ratios and net survival when covariate data are missing: strategies for multiple imputation. *Epidemiology*. 2015;26 (3):421-428.
138. Resche-Rigon M, White IR, Bartlett JW, Peters SAE, Thompson SG, PROG-IMT Study Group. Multiple imputation for handling systematically missing confounders in meta-analysis of individual participant data. *Stat Med*. 2013;32(28): 4890-4905.
139. Jolani S, Debray TPA, Koffijberg H, Buuren S, Moons KGM. Imputation of systematically missing predictors in an individual participant data meta-analysis: a generalized approach using MICE. *Stat Med*. 2015;34(11):1841-1863.
140. Resche-Rigon M, White IR. Multiple imputation by chained equations for systematically and sporadically missing multilevel data. *Stat Methods Med Res*. 2018;27(6):1634-1649.
141. Audigier V, Resche-Rigon M. *micemd: Multiple Imputation by Chained Equations with Multilevel Data*. 2017.
142. Kline D, Andridge R, Kaizar E. Comparing multiple imputation methods for systematically missing subject-level data. *Research Synthesis Methods*. 2017;8(2):136-148.
143. Giorgi R, Belot A, Gaudart J, Launoy G. French network of cancer registries FRANCIM. The performance of multiple imputation for missing covariate data within the context of regression relative survival analysis. *Stat Med*. 2008;27(30): 6310-6331.
144. Grund S, Lüdtke O, Robitzsch A. Multiple imputation of missing covariate values in multilevel models with random slopes: a cautionary note. *Behav Res Methods*. 2016;48(2): 640-649.
145. Marson AG, Williamson PR, Clough H, Hutton JL, Chadwick DW. On behalf of the epilepsy Monotherapy trial group. Carbamazepine versus valproate Monotherapy for epilepsy: a meta-analysis. *Epilepsia*. 2002;43(5):505-513.
146. Therneau TM. *coxme: Mixed Effects Cox Models*. R package version 2.2–7; 2018.
147. R Core Team. *R: A Language and Environment for Statistical Computing*. 2018.
148. Jones E, Sweeting MJ, Sharp SJ, Thompson SG, EPIC-InterAct Consortium. A method making fewer assumptions gave the most reliable estimates of exposure-outcome associations in stratified case-cohort studies. *J Clin Epidemiol*. 2015;68(12): 1397-1405.
149. Marson AG, Williamson PR, Hutton JL, Clough HE, Chadwick DW. Carbamazepine versus valproate monotherapy for epilepsy. *Cochrane Database Syst Rev*. 2000;3:1-20.
150. Lin DY, Zeng D. On the relative efficiency of using summary statistics versus individual-level data in meta-analysis. *Biometrika*. 2010;97(2):321-332.

151. Commenges D, Andersen PK. Score test of homogeneity for survival data. *Lifetime Data Anal* 1995;1(2):145–156; discussion 157–159.
152. Gray RJ. Tests for variation over groups in survival data. *J Am Stat Assoc*. 1995;90(429):198-203.
153. Claeskens G, Nguti R, Janssen P. One-sided tests in shared frailty models. *Test*. 2008;17(1):69-82.
154. Economou P, Stehlík M. On small samples testing for frailty through homogeneity test. *Communications in Statistics - Simulation and Computation*. 2015;44(1):40-65.
155. Biard L, Porcher R, Resche-Rigon M. Permutation tests for Centre effect on survival endpoints with application in an acute myeloid leukaemia multicentre study. *Stat Med*. 2014;33(17):3047-3057.
156. Hox JJ. *Multilevel Analysis: Techniques and Applications*. 2nd ed. New York, NY: Routledge, Taylor & Francis; 2010.
157. Goldstein H. *Multilevel Statistical Models*. 4th ed. Hoboken, N.J.: Wiley; 2011.
158. Volinsky CT, Raftery AE. Bayesian information criterion for censored survival models. *Biometrics*. 2000;56(1):256-262.
159. Vaida F, Blanchard S. Conditional Akaike information for mixed-effects models. *Biometrika*. 2005;92(2):351-370.
160. Greven S, Kneib T. On the behaviour of marginal and conditional AIC in linear mixed models. *Biometrika*. 2010;97(4):773-789.
161. Campbell H, Dean CB. The consequences of proportional hazards based model selection. *Stat Med*. 2014;33(6):1042-1056.
162. Harrell FE. *Regression Modeling Strategies: with Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis*. 2nd ed. Cham, Switzerland: Springer; 2015.
163. Keiding N, Andersen PK, Klein JP. The role of frailty models and accelerated failure time models in describing heterogeneity due to omitted covariates. *Stat Med*. 1997;16(2):215-224.
164. Gompertz B. On the nature of the function expressive of the Law of human mortality, and on a new mode of determining the value of life contingencies. *Philosophical Transactions of the Royal Society of London*. 1825;115:513-583.
165. Makeham WM. On the Law of mortality and the construction of annuity tables. *The Assurance Magazine, and Journal of the Institute of Actuaries*. 1860;8(6):301-310.
166. Grambsch PM, Therneau TM. Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika*. 1994;81(3):515-526.
167. Wei LJ. The accelerated failure time model: a useful alternative to the cox regression model in survival analysis. *Stat Med*. 1992;11(14–15):1871-1879.
168. Lambert P, Collett D, Kimber A, Johnson R. Parametric accelerated failure time models with random effects and an application to kidney transplant survival. *Stat Med*. 2004;23(20):3177-3192.
169. Aalen OO, Cook RJ, Røysland K. Does Cox analysis of a randomized survival study yield a causal treatment effect? *Lifetime Data Anal*. 2015;21(4):579-593.
170. Wolkewitz M, Cooper BS, Palomar-Martinez M, et al. Multilevel competing risk models to evaluate the risk of nosocomial infection. *Crit Care*. 2014;18(2):R64.
171. Su X, Zhou T, Yan X, Fan J, Yang S. Interaction trees with censored survival data. *The International Journal of Biostatistics*. 2008;4(1):Article 2.
172. Andreano A, Rebora P, Valsecchi MG. Measures of single arm outcome in meta-analyses of rare events in the presence of competing risks. *Biometrical Journal Biometrische Zeitschrift*. 2015;57(4):649-660.
173. Debray TPA, Moons KGM, Ahmed I, Koffijberg H, Riley RD. A framework for developing, implementing, and evaluating clinical prediction models in an individual participant data meta-analysis. *Stat Med*. 2013;32(18):3158-3180.
174. Therneau T. *Grambsch PM*. Extending the Cox Model. Springer, New York: Modeling Survival Data; 2000.
175. Bennett MM, Crowe BJ, Price KL, Stamey JD, Seaman JW. Comparison of Bayesian and frequentist meta-analytical approaches for analyzing time to event data. *J Biopharm Stat*. 2013;23(1):129-145.
176. Sobel M, Madigan D, Wang W. Causal Inference for Meta-Analysis and Multi-Level Data Structures, with Application to Randomized Studies of Vioxx. *Psychometrika*. 2016;.
177. Rotolo F. Paoletti X. *Michiels S Surrosurv: An R Package for the Evaluation of Failure Time Surrogate Endpoints in Individual Patient Data Meta-Analyses of Randomized Clinical Trials Computer Methods and Programs in Biomedicine*. 2018;155:189-198.
178. Boutitie F, Gueyffier F, Pocock SJ, Boissel JP. Assessing treatment-time interaction in clinical trials with time to event data: a meta-analysis of hypertension trials. *Stat Med*. 1998;17(24):2883-2903.
179. Cox C, Chu H, Schneider MF, Muñoz A. Parametric survival analysis and taxonomy of hazard functions for the generalized gamma distribution. *Stat Med*. 2007;26(23):4352-4374.
180. Goldstein H, Browne W, Rasbash J. Multilevel modelling of medical data. *Stat Med*. 2002;21(21):3291-3315.

SUPPORTING INFORMATION

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

How to cite this article: de Jong VMT, Moons KGM, Riley RD, et al. Individual participant data meta-analysis of intervention studies with time-to-event outcomes: A review of the methodology and an applied example. *Res Syn Meth*. 2020;11:148–168. https://doi.org/10.1002/jrsm.1384