MRC Clinical Trials Unit

Smarter Studies
Global Impact
Better Health

UCL

Exploring the Implications of Analysing Time-to-Event Outcomes as Binary in Meta-analysis

Thesis submitted for the degree of

Doctor of Philosophy

(Field of Study: Statistics)

Theodosia Salika

Institute of Clinical Trials and Methodology

University College London

“Οὐδὲν τοῖς θαρροῦσιν ἀνάλωτον”

Πλούταρχος *Αλέξανδρος* 58.1

“For the courageous nothing is unattainable”

Plutarch *Alexander* 58.1

Declaration

I, Theodosia Salika, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Signed:…… Date:……

Use of own published work:

The work presented in Chapters 3 and 4 was published in BMC Medical Research methodology, doi:10.1186/s12874-022-01541-9. The published article is included in the appendix (Appendix F).

This PhD is supported by award received from the Medical Research Council, Clinical Trials Unit, University College London.

Abstract

Systematic reviews and meta-analysis of time-to-event outcomes can be analysed on the hazard ratio (HR) scale but are very often dichotomised and analysed as binary using effect measures such as odds ratios (OR). This thesis investigates the impact of using these different scales by re-analysing meta-analyses from the Cochrane Database of Systematic Reviews (CDSR), using individual participant data (IPD) and a comprehensive simulation study.

For the CDSR and IPD, the pooled HR estimates were closer to 1 than the OR estimates in most meta-analyses. Important differences in between-study heterogeneity between the HR and OR analyses were observed. These caused discrepant conclusions between the OR and HR scales in some meta-analyses. Situations under which the clog-log link outperformed the logit link and vice versa were apparent, indicating that the correct method choice does matter. Differences between scales occurred mainly when event probability was high and could occur via differences in between-study heterogeneity or via increased within-study standard error in OR relative to HR analyses.

In many simulation scenarios, analysing time-to-event data as binary using the logit link did not substantially affect bias and coverage apart from those where large percentage random censoring and long follow-up time was present. The method though lacks precision particularly for small meta-analyses. Analysing the data as binary using the clog-log link consistently produced more bias, low coverage and low power.

If a HR estimate cannot be obtained per trial to perform a meta-analysis of time-to-event data, a meta-analysis using the OR scale (using the logit link) could be conducted but with awareness that this would provide less precise estimates in the analysis. Investigators should avoid performing meta-analyses on the OR scale in the presence of high event probability, large percentage random censoring and therefore longer follow-up times assuming of large event rates of the trials included.

Impact Statement

Medical research questions are usually investigated multiple times by different research groups performing separate studies. The results may be contradictory, and may not allow clear conclusions to be drawn, producing difficulties in medical decision-making.

Systematic reviews and meta-analyses of time-to-event outcomes (e.g., time to death, recurrence of symptoms, relief of pain etc.) are frequently carried out and are very common in areas such as cancer, respiratory and cardiovascular diseases. These outcomes are related to “IF” and “WHEN” an event has occurred and they are commonly analysed as binary in meta-analysis, rather than accounting for their natural properties. The work presented in this thesis focuses on time-to-event outcomes and the implications of analysing them as binary in a meta-analysis. My aim was to provide guidance to systematic reviewers and meta-analysts on the most appropriate methodology that should be used when conducting such meta-analyses.

Using empirical survival meta-analysis data from the Cochrane Database of Systematic Reviews (Issue 1, 2008) and individual participant data (IPD), I indicated that time-to-event data should ideally be analysed accounting for their natural properties and meta-analysts need to be careful about choice of method. I identified that dichotomising time-to-event outcomes may be adequate for low event probabilities but not for high event probabilities. In the IPD meta-analysis performed, I confirmed the results obtained from my empirical study, however, it was not possible to explain whether censoring and follow-up time were distinct factors affecting the discordance among the meta-analysis estimates.

My simulation study indicated that a time-to-event meta-analysis should be conducted ideally in the presence of IPD with interpretation on a HR scale, whereas in absence of IPD, extracting information from trial reports and using a log-rank test performs equally well. The logit link performed well in many simulation scenarios with some exceptions; the method though lacked precision in most scenarios. The complementary log-log link was not suitable to analyse the data as binary on a HR scale since substantial bias, low coverage, and low power were observed. If HR estimates cannot be obtained, a meta-analysis using

the OR scale (using the logit link) could be conducted but with awareness that this would provide less precise meta-analytic estimates. Investigators should avoid performing meta-analyses on the OR scale in the presence of large percentage random censoring and long follow-up times.

It is advised that systematic reviewers and meta-analysts should think carefully about the circumstances before analysing time-to-event data as binary because this may produce different conclusions than the correct time-to-event analysis. Investigators should avoid performing meta-analyses on the OR scale in the presence of high event probability, large percentage random censoring and long follow-up times of the trials included in the meta-analysis. Researchers should consider also that precision will be lower so the analysis will have lower power especially in small meta-analyses and will be less likely to detect a significant treatment effect. The complementary log-log link should not be used as an alternative to analysing time-to-event outcomes as binary on a HR scale.

Acknowledgments

I would like to express my gratitude to many people and groups who have provided me with guidance and support during my PhD journey. I am grateful to the MRC Clinical Trials Unit at UCL for the funding I have received throughout the course of my degree. I am grateful to the Nordic Cochrane Centre and the Cochrane Collaboration Steering Group for providing me with access to the Cochrane Database of Systematic Reviews. I would also like to thank the meta-analysis team for providing me with results obtained from the MRC Clinical Trials Unit's Survey of Collaborative Review Groups and with the individual participant data used in this thesis.

My principal supervisor, *Rebecca M. Turner*, who has given me both statistical and mental support during these years. This thesis would not have been achieved without Rebecca's guidance, encouragement and constructive feedback. *Ian R. White*, my subsidiary supervisor, has been exceptional in his support, input and our inspiring conversations allowed me to proceed further with this thesis. It has been a fantastic experience working with such an intelligent person as Ian. *Jayne F. Tierney* and *David Fisher*, the members of my advisory committee, who have been continuously engaged and offered valuable ideas and perspectives in my work. I would also like to thank *Claire Vale*, the department's graduate tutor, who has been always encouraging me throughout these years.

My academic mother, friend and colleague, *Hazel M. Inskip*, who has been one of the biggest supporters for me to pursue a PhD while allowing me to work flexible hours for my part-time position. I have also been lucky enough to be a member of a small *Greek* church group who have always offered an enjoyable environment and provided a balance to my academic work.

I need to thank my family and friends (especially *Roza, Georgia, Sophia and Elena*) who have been incredible pillars of strength and comfort. Finally, this thesis is dedicated to the person to whom I owe everything I have achieved to date; my Mum, *Eleni Sianidou*, for her unconditional love, constant encouragement, and patience; the biggest supporter in my life.

To my mother

Table of Contents

List of Tables

List of Figures

Abbreviations

ANOVA	Analysis of Variance
CDSR	Cochrane Database of Systematic Reviews
CI(s)	Confidence Interval(s)
Clog-log	Complementary log-log link
EM	Expectation-minimisation
HR(s)	Hazard Ratio(s)
HTA	Health Technology Assessment
IGLS	Iterative Generalized Least Squares
IPD	Individual Participant Data
IQR	Interquartile Range
MA(s)	Meta-analysis(es)
MHR	Median Hazard Ratio
NMA(s)	Network Meta-analysis(es)
NNT	Number Needed to Treat
OR(s)	Odds Ratio(s)
PH	Proportional Hazards
REML	Restricted Maximum Likelihood
RMST	Restricted Mean Survival Time
RR(s)	Risk Ratio(s) or Relative Risk(s)
SAS	Statistical Analysis System
SE(s)	Standard Error(s)
SR(s)	Systematic Review(s)
TTE	Time to an Event
WoS	Web of Science

1. Introduction

1.1 Chapter Overview

This chapter introduces the main definitions and characteristics of Time-to-Event (TTE) data along with Systematic Reviews and Meta-analyses (MA). It provides preliminary background information on how these data are used in a systematic review and MA, including also important justifications for the necessity of this research.

1.2 What is Time to an Event?

In many clinical and non-clinical studies, an important outcome of interest is measured by the time-to-event (TTE). These types of outcomes are unique in the sense that they are dependent on two essential characteristics; the first is related to “IF” and the second “WHEN” an event has occurred. Examples of such events may involve time from diagnosis of cancer to death, time to weaning of breast-fed infants, or time from start of in vitro fertilisation treatment to pregnancy.

A key feature of TTE data is that the event will not necessarily occur for all participants in the study by the end of the follow-up period, meaning that we will never know whether and when some participants experience the event[1]. Such observations are known as “censored”, indicating that the follow-up period ended before the event occurred. Other examples of censored observations include participants being lost to follow-up during the study (for example, because study participants may have moved to another country) or dying from a cause not related to the outcome of interest. There are three forms of censoring: right (where the observed survival time is less than the actual unknown survival time),

left (where the actual survival time is less than that observed) and interval (where the participant experiences the event within a known time interval)[2].

Finally, TTE are not symmetrically distributed, they are positively skewed, producing a longer “tail” to the right of the distribution; this is indicative that it is not reasonable to assume that these data follow a normal distribution[2]. Traditional linear or logistic regression methods are not suitable to model these data as they are not able to account for their natural properties.

1.3 Analysing Time-to-Event data

An important step prior to modelling TTE data is to present a numerical (e.g., life-tables) or graphical representation of the survival times (e.g., Kaplan-Meier curves) for the participants of a group in a study. TTE data can be summarised via a survival or a hazard function; details on the estimation of these functions has been described elsewhere[2-4]. Once a survival function has been defined adequately, different percentiles of the distribution of survival times can be estimated and displayed graphically. Kaplan-Meier curves are a useful representation of TTE data when we want to obtain an estimate of the proportion of participants alive at certain time points[5].

In the presence of two or more groups in a study, there are non-parametric, parametric and semi-parametric procedures available to compare formally the survival times of the groups. Non-parametric procedures include tests such as the log-rank test[6], the Wilcoxon test[2], and stratified versions of these tests. However, when additional information are recorded for each participant in the study such as demographic or disease related characteristics, parametric models (i.e. Exponential or Weibull) and semi-parametric (i.e. Cox) models are able to account for these covariates, providing more reliable results for the analyses[3, 7]. The exponential model assumes a constant hazard function over time, the Weibull model has a more flexible form of the hazard function, and the Cox model is most commonly used and does not make any assumption on the form of the underlying baseline hazard function.

A key assumption for the application of these models is that censoring is non-informative with respect to the distribution of survival time. This means that participants’ censoring time is independent of their failure time, whereas censoring is informative if patients’ censoring time depends on the failure time[8,

[9]. For example, censoring that occurs where study participants drop out of a trial comparing two treatments for cancer due to an ineffective control arm is considered as informative. In this case, the aforementioned models may produce biased results. Other methodology such as multiple imputation techniques for missing data, the use of drop-out event as a study end-point, and joint modelling of longitudinal and TTE data has been developed to examine data under these circumstances[10].

1.4 Systematic Reviews and Meta-analyses

Many medical research questions are investigated multiple times by different research groups performing separate medical studies. The results of the studies may be contradictory, and may not allow clear conclusions to be drawn, producing difficulties in medical decision-making[11].

A systematic review aims to combine empirical evidence based on pre-specified eligibility criteria, answering specific research questions that are not able to be answered by the individual studies themselves[12, 13]. Characteristics of a well-conducted systematic review include a protocol stating clearly the objectives, research questions, and methods prior to conduct of the review, a comprehensive search strategy including various bibliographical databases, explicitly stated inclusion and exclusion criteria, and development of quality criteria to evaluate research validity[14]. Different types of Cochrane reviews exist and these include intervention, diagnostic test accuracy, methodology, qualitative and prognosis reviews[13].

A MA is a statistical analysis performed within a systematic review and is able to identify whether strong evidence exists on the effectiveness of treatment for a particular disease[12]. The main aim is to mathematically summarise the results across studies, if appropriate, using suitable methodology, even if studies have used different effect measures to assess their outcomes. These summary results can provide greater statistical power on treatment effects, assess between-study heterogeneity, and identify characteristics of studies that are importantly associated with effective treatments[11]. In comparison to running a clinical trial, it is a relatively quick way to assess the effectiveness of healthcare interventions, facilitates medical decision-making, introduces new guidelines for treatments on different diseases and initiates new medical studies.

Once relevant studies from the literature have been identified and data are either extracted from published reports or collected from study authors, appropriate methodology should be applied. Two models are usually considered; one uses a common-effect approach (known also as “fixed-effect”) and the other a random-effects approach for combining study estimates[14, 15]. When treatment estimates are combined under a common-effect (or fixed-effect) model, we assume a single underlying treatment effect and no variability between the study results. This assumption frequently seems unrealistic as studies involved in a MA may differ in terms of study design, participant demographics, follow-up time and other characteristics[14]. When random-effects models are used, the true underlying treatment effects are assumed to vary at random across studies, and a normal distribution is usually assumed for these effects. Therefore, two sources of variation are observed: the within and between-study variability[14].

1.4.1 The Cochrane Collaboration

Cochrane is an international collaboration that has performed high-quality research over the last 29 years and includes independent researchers, heath care professionals, patients, carers and other stakeholders interested in improving health outcomes[13]. Their aim is to produce high-quality and accessible systematic reviews to promote evidence-based health decision-making. Official guidance including detailed useful information on the process of conducting a systematic review and MA is provided in the Cochrane Handbook of Systematic Reviews[16]. Authors are advised to follow the guidance provided by the book both on standard methods and more advanced topics[13]. The Cochrane Database of Systematic Reviews (CDSR) is a database including systematic reviews, protocols, editorials and supplements in health care and to date includes over 7,500 systematic reviews[17].

1.5 Systematic Reviews and Meta-analysis of Time-to-Event Outcomes

Special methods are required to combine studies including TTE data in a MA; censoring cannot be accommodated by analyses such as linear or logistic regression. If we treat TTE data as continuous we are assuming uncensored observations instead of allowing for censoring, we are underestimating average survival and therefore we are inadequately addressing the unique properties of these data. On the other hand, treating TTE data as binary may be sensible in

specific circumstances (discussed in 1.6); however, to account properly for the number of participants experiencing an event and the amount of time taken for the event to occur needs a more sensitive approach.

Guidelines are provided by the Cochrane Handbook of Systematic Reviews[16], providing support to researchers who are wishing to perform such MA. The easiest way to perform a MA of TTE data is to obtain a summary estimate from each study along with its standard error (SE) and combine them under a common-effect (also known as fixed-effect) or random-effects model. Specifically, if log HRs and SEs from Cox proportional hazards analyses can be obtained, study results can be combined to provide a pooled effect estimate along with its confidence interval (CI)[18]. The pooled HR obtained represents a comparison of the instantaneous risk of event in the treatment against the control group over the follow-up time[14]. If data are collected from published reports (i.e. aggregate data), we can obtain the log-rank observed minus expected events ("O-E") and variance ("V") statistics and by using appropriate statistical software we can perform a MA[18].

In more detail, the log-rank test[6] performs a comparison across the whole length of the survival curve. Time is split into intervals, observed and expected events are calculated, the "O-E" and "V" values are summarised across the studies for each time interval and finally "O-E" values are divided by the "V" values; a comparison against the standardised normal distribution is constructed to obtain a test statistic for the survival difference among the study groups. This method will give a rise to a Hazard Ratio (HR)[14].

"O-E" and "V" statistics can also be obtained by using Peto's method on dichotomous data. Peto's method was firstly proposed by Yusuf et al.[19] and gives a rise to an Odds Ratio (OR) also called a "Peto OR". In order to obtain a "Peto OR", we need the number of exposed and non-exposed participants on each study's group and the number of events and non-events that occurred; this can be easily summarised by a contingency table. "O-E" and "V" values are calculated for each study and "O-E" values are divided by the "V" values to obtain an OR. It is important to state that in the presence of substantial difference in the group sizes, serious bias can occur[20]. Failure times and censoring are also ignored under this method. In the presence of long medical studies examining mortality, for example, no difference between treatments may be observed since all

participants may experience the event, regardless of whether the treatment delayed the event or not[21].

To avoid some of these limitations, a modified version of the Peto method was specifically proposed for TTE data by Simmonds et al.[21] and provides an OR estimate. By dividing the trial into pre-defined time intervals (e.g., by year or by month), multiple contingency (two by two) tables for each study can be obtained. "O-E" and "V" statistics can be obtained for each time interval, and the log OR over all time intervals can be estimated by dividing the sum of "O-E" values by the sum of "V" values. Using this approach, a hypergeometric distribution is assumed for the observed events, misclassification bias is minimised since censored observations are excluded from the analyses for time intervals after their censoring time, and failures in the two treatment groups occurring at different times are considered[21].

It is worth mentioning though that published studies seldom report all the statistics needed to obtain a modified Peto OR or a HR and variance estimates. Both Parmar et al.[22] and Williamson et al.[23] have considered various ways to account for information from published reports and extract HR and variance estimates to facilitate MA implementation. A more detailed description of this methodology is presented in Chapter 2. The ideal scenario and the only reliable way to perform a MA according to some researchers[24] is to obtain the individual participant data (IPD) since we can then adjust for differences in case-mix by using covariates in our analyses, allowing also for variation among studies[25].

Various different software packages can be used to carry out a MA. Software such as STATA[26], SAS[27], R[28], Python[29], WinBUGS[30] are able to handle TTE data and perform MA; functions are continuously developing to accommodate new methodological developments. RevMan[31] has been specifically developed to facilitate preparation of systematic reviews and MA; however, the choice of models is restrictive when "O-E" and "V" statistics are available since only common-effect (or fixed-effect) MA is allowed[18].

1.6 Effect Measures for Time-to-Event Outcomes

For TTE outcomes, several effect measures have been used previously; the most commonly reported effect measures are the OR, HR and relative risk or risk ratio (RR). Below, I provide brief definitions of these outcome measures and then in

1.6.1 I carry out a chronological overview of important research conducted on the differences between effect measures observed in a single study.

When TTE outcome data are dichotomised and are therefore in a binary form, they can be conveniently arranged into a contingency table, and several different measures of treatment effect may be used. An OR can be calculated, providing a relative measure of the probability of an event: the odds of having the event in the treatment group relative to the odds of having the event in the control group[32]. For beneficial events, if an OR yields an estimate greater than one, a new treatment is more effective when the treatment and control groups are compared, whereas an OR of less than one indicates that a new treatment is less effective than the control group. For adverse events, interpretation is conducted the other way around. As an alternative to an OR, a RR can be calculated and is defined as the ratio of the probability of event occurring in the treatment group compared to the probability of event occurrence in the control group[32]. Risk difference, which is the difference between the probabilities of event occurring between groups, and number needed to treat, which is the reciprocal of the risk difference, are additional measures of comparative effects for binary outcomes[32].

Once the full nature and properties of TTE data are considered, the most sensible summary statistic usually employed when comparing a TTE outcome between two groups is the HR. A HR measures the instantaneous reduction in the risk of an event over a particular time frame in the treatment group relative to that of the control group[14]. If a HR is independent of time (i.e. constant), then proportional hazards are assumed among the two groups; this is the most important assumption underpinning covariates inclusion in a Cox regression model[11].

1.6.1 Past Comparisons Among Different Effect Measures for Time-to-Event Outcomes

A number of authors have previously discussed the comparison between logistic regression and Cox proportional hazards models. One of the first papers comparing the two models was written by Green and Symons[33] in 1983. Green and Symons[33] explained the mathematical relationship between logistic and Cox regression models; via an example they indicated that proportional hazards models provide relatively stable coefficients and decreased SE with increasing follow-up time, which is not the case for logistic models where SEs of the

estimates generally increase. These authors also mentioned that the two models produce similar estimates in the presence of rare incidence of a disease and short follow-up time.

Peduzzi et al.[34] evaluated the logistic and the Cox proportional hazards model when the event occurs in all participants after a fixed period of time. OR, RR and HR are discussed in the paper which shows that in the presence of rare events logistic and proportional hazard models' estimates are very similar to log RR model estimates. As the event probability increases, estimates become more discordant, however the likelihood ratio statistics are asymptotically equivalent under the null hypothesis that the regression coefficients are equal to zero. Finally, the authors indicated that these findings extend also to the multivariate case (i.e. when adjustment for baseline covariates is considered)[34].

In 1989, two research papers were published, one using a real-world example and the other using a simulation dataset, both comparing logistic regression and proportional hazards models. Annesi et al.[35] extended work conducted by J. Cuzick[36], by examining whether the asymptotic relative efficiency, which is defined as "ratio of the numbers of subjects necessary for the two models, to gain the same asymptotic statistical power", is close to one in the presence of high survival rate, when several risk factors are adjusted for and censoring time is identical for all subjects. The authors showed via analysis of a longitudinal dataset that the logistic model is less efficient than the Cox model; inclusion of several risk factors showed that these models are asymptotically equivalent in identifying predictors of events with low event probability[35]. However, they mentioned that logistic models may be appealing when survival times are recorded by intervals, where in the presence of many failures the model assumes no tied observations[35].

Ingram and Kleinman[37] in the same year, using simulation datasets mainly, compared estimates among a newly introduced person-time model (i.e. a modified version of logistic model), cumulative logistic and Cox models. The person-time model divides the study period into intervals, counts the numbers of risks and events in each interval, and sums these counts overall. The authors demonstrated that person-time, logistic and Cox models are identical for similar censoring rates as long as the event probability and follow-up time increase, however discordant estimates are observed in the cumulative logistic model[37].

Furthermore, when the distribution of survival time was close to exponential and in the presence of mild violations of the proportional hazards assumption, person-time and Cox models yielded little effect on parameter estimates[37].

Docksum and Gasko[38], in 1990, using a theoretical framework, discussed the fact that survival analysis models can be considered as modelling a specified transformation of the dependent variable, a linear combination of the independent variables accounting also for errors; they acknowledged similarities of the two models described by Green and Symons[33]. The authors discussed the development of Berkson's logit model three decades before the development of proportional odds model in the 1970's; they justified this since the key ingredients for model development were not in place before the 70s and 80s, such as computing tools and repurposing a model for a new application area requires a large amount of time.

In 1998, Callas et al.[39], using occupational cohort data, compared Poisson, logistic and Cox proportional hazards models. Their analyses indicated that Poisson and Cox PH models yielded nearly identical results in the presence of small sample size and rare events, in terms of coefficients and CIs, apart from a case where an age confounder included into the model in four wide intervals produced residual confounding and affected the estimates. A finding the authors found which was consistent with other studies was that logistic regression models provided discordant estimates for common outcome and strong RR; however, length of follow-up had little impact on the estimates, a finding not necessarily consistent with other studies. Authors discussed the generalisability of the results due to use only on real data conditions.

In the beginning of the new millennium, Symons and Moore[40] discussed the reporting of HR in prospective epidemiological studies. They indicated that in cases where the HR is greater than one, it consistently exceeds RR and is exceeded by OR. The authors provided evidence that similarities or differences of the three estimates is based upon the following three factors: the first is related to the length of follow-up, the second to the average rate of event occurrence, and the third to the risk of the exposed relative to the control group. Furthermore, they state that lack of preciseness in the terminology usage exist probably because these measures are often similar.

Finally, in a more recent research work conducted by Stare and Maucort-Boulch[41], using simple examples, the authors tried to challenge the misbelief that OR is a RR, and HR is a RR. They indicated that there are circumstances in which reporting these measures to be something they are not can produce misleading results. Therefore, they examined the relation between OR/RR and HR/RR by giving appropriate definitions for each one of them and describing the circumstances under which each measure approaches each other.

1.7 Real Problems Cochrane Review Groups Face with Meta-analyses of Time-to-Event Outcomes

A project[42] was conducted in 2008 at the Medical Research Council, Clinical Trials Unit with the main aim of improving the quality of analysing TTE outcomes in Cochrane Reviews based on methodology described by Parmar et al.[43] and Williamson et al.[44]. A survey was distributed to 49 different Cochrane Review groups and the response rate was 55%[42]. The responders spanned a range of health care areas including cancer, infectious diseases, genetic disorders, cardiovascular health and oral health. Among the responders most of the groups (78%) included dichotomous outcomes based on TTE data, 67% included continuous time-related outcomes and 59% included TTE outcomes, in their reviews[42]. Additionally, 69% extracted dichotomous data on the outcomes and analysed them as odds ratios, relative risks or (rarely) risk differences. However, only 37% of the groups used the methods of Parmar et al[43]. and Williamson et al.[44] despite their awareness of these methods in the literature.

The main output of the project was that although HR is considered the most appropriate scale for analysis of TTE data, in practice OR and RR are frequently used instead due to the following reasons[42]:

- unavailability of individual participant data (IPD)
- limitations on how these outcomes are reported in individual trial reports
- lack of familiarity in handling TTE outcomes for meta-analysis
- difficulties in understanding the methods of analysing such data without a statistician
- limited available training for the majority of systematic reviewers and meta-analysts who perform such analyses.

1.8 Thesis Objectives

In this thesis, I am interested to explore how TTE data are analysed within a MA and investigate the implications of analysing TTE outcomes using various meta-analytic models resulting in the different effect measures of OR and HR. To my knowledge there is no previous research on analysing TTE outcomes as binary within MA. More specifically, I am interested in answering the following questions:

- What are the implications of analysing TTE outcomes as binary in MA and how do the implications vary according to MA characteristics?
- How are TTE outcomes analysed within the biggest database publishing systematic reviews and MA, the CDSR? Are they analysed as binary or are they analysed as HR, taking into account the full properties of the data?
- Which medical areas within the database analyse the data under which scale?
- What are the assumptions made when different meta-analytic models are applied and what are the advantages and disadvantages of each one of them?
- Is there any other method that could allow us to mitigate the undesirable properties from treating the data as binary?

To answer these questions, I used real life data sets extracted from the CDSR, IPD MA data sets, and simulation studies. This thesis has seven chapters. In Chapter 2, I present a methodology review on guidance for MA of TTE outcomes. In Chapter 3, using data from the CDSR analysed originally as binary for TTE outcomes, I compare the methods of analysing these data as binary on the OR scale to an alternative option where interpretation can be performed on the HR scale. In Chapter 4, using a subset from the same database, I perform comparisons to explore differences in the results on MAs originally analysed on the HR scale using the “O-E and V” statistics, to treating the data as binary using a methodological alternative interpreting the results on a HR scale and by analysing the data as binary on the OR scale. In chapter 5, I present the results obtained under the different scales when analysing an IPD dataset obtained from the MRC Clinical Trials Unit; this chapter additionally includes the gold-standard approach of a Cox proportional hazards model. In chapter 6, I perform a simulation-based study comparing

different methods for meta-analysis in TTE outcomes allowing me to identify separately the factors affecting the potential discordance among the scales. Finally in Chapter 7, I provide a summary and discussion of the key findings obtained from the previous chapters. In the same chapter a final conclusion on the research question is drawn.

Overall, the objective of this research is to provide guidance to systematic reviewers about the implications of analysing TTE outcomes as binary, how the implications vary according to MA characteristics and in which circumstances analysing the outcome as binary may be adequate.

2. A Methodology Review for Meta-analysis of Time-to-Event Outcomes

2.1 Chapter Overview

This chapter outlines the guidance that exists in the literature for MA of TTE outcomes and any discussions raised for analysing these data as binary. A specific search strategy was followed, and research papers were extracted from the following databases: Medline (Ovid Version), Scopus and Web of Science. Literature was assessed for eligibility, was assigned to specific categories and was reviewed.

2.2 Introduction

Previous research has documented the effect measures needed for a survival analysis and provided appropriate methodology on how TTE data should be ideally handled in a single study, and under which circumstances it could be acceptable to treat these data as binary[33-35, 37, 39]. Recent reviews indicate that Kaplan-Meier methods, Cox regression models, logistic regression are the most commonly used methods in analysing TTE outcomes, while techniques avoiding making the proportional hazards assumption such as accelerated failure time models or time dependent Cox regression techniques are less frequently used[45, 46]. Additional methods for the analysis of TTE data exist including the non-parametric log-rank test and parametric proportional hazard models assuming a specific distribution for the hazard such as Weibull, Exponential, Gompertz.

TTE data MA should be ideally analysed using IPD and interpretation is performed on the HR scale assuming constant hazards over time (i.e.

proportional hazards assumption). However, access to IPD is rarely available and different techniques have been employed to obtain study level data from research publications[43, 44, 47]. The current methodology review was performed to ascertain the guidance for MA of survival outcomes and any discussions raised on MA of TTE data as binary, with the objective of informing the subsequent research reported in later chapters.

2.3 Searching the Literature

I performed a methodology review to identify all methodological publications providing guidelines for MA of TTE data. Medline (Ovid version, 1946-December 2021), Scopus (2004-December 2021) and Web of Science (1900-December 2021) were searched using keywords such as “meta-analysis”, “time-to-event”, “survival”, “methodology” via the “Advanced search” function in the electronic databases. Details on the search strategy are provided in 2.3.1.

2.3.1 Search Strategy

The following search strategy was applied in order to extract all the relevant papers from the databases.

- Search strategy for MEDLINE (Ovid version)

1. “meta-analys#s”.ti,ab.
2. (“time-to-event” or “time to event”).ti,ab.
3. (“survival outcome” or “survival endpoint” or “survival data” or “survival study” or “survival analys$”).ti,ab.
4. (“failure time” or “failure time data”).ti,ab.
5. (“guid*” or “method*” or “framework”).ti,ab.
6. 2 or 3 or 4
7. 1 and 5 and 6

- Search strategy for Scopus

TITLE-ABS-KEY (“meta-analys*”) AND TITLE-ABS-KEY (“time to event” OR “survival outcome” OR “survival endpoint” OR “survival data” OR “survival study” OR “survival analys*” OR “failure time” OR “failure time data”) AND TITLE-ABS-KEY (“method” OR “guid*” OR “framework”) AND NOT INDEX (medline) AND (LIMIT-TO (LANGUAGE, “English”)) AND (LIMIT-TO

(SUBJAREA, “MEDI”)) OR (LIMIT-TO (SUBJAREA, “MATH”)) OR (LIMIT-TO (SUBJAREA, “DECI”))

- Search strategy for Web of Science (WoS)

1. TS=(“meta-analys?s” or “meta*analys?s”) AND LANGUAGE: (English)
2. TS =(“time-to-event” or “time*to*event”) AND LANGUAGE: (English)
3. TS =(“survival outcome” or “survival endpoint” or “survival data” or “survival study” or “survival analys$”) AND LANGUAGE: (English)
4. TS =(“failure time” or “failure time data”) AND LANGUAGE: (English)
5. TS =(“guid*” or “method” or “framework”) AND LANGUAGE: (English)
6. #4 OR #3 OR #2
7. #6 AND #5 AND #1

The inclusion and exclusion criteria of the methodology review were broad (Table 2.1). I did not aim to make any comparisons or judgements on the proposed methodologies but to provide descriptions of methods.

Criteria	Inclusion	Exclusion
Journal	No restriction	No restriction
Publication Type	Full publications	Abstracts, conferences abstracts, notes
Country	No restriction	No restriction
Language	English publications	Non-English publications
Year of publication	No restriction	No restriction
Outcomes	Time-to-event	Binary, continuous, mixed, surrogate
Methods	Methodology, extensions, and comparisons	Applied methodology only, prognostic/diagnostic accuracy studies

Table 2.1: Inclusion and exclusion criteria of the methodology review used in MEDLINE (Ovid Version), Scopus, and Web of Science.

In the review, I identified 2,523 publications based on the search terms used. Among those, I removed 2,352 after title screening, 41 after abstract reading, 46 after duplicates removed, and 27 after full-text reading. I additionally included 17 publications via hand searching which were missed from the basic search terms.

Hence, I included 75 methodological publications according to the search methods described above. A full list of references is available in Appendix A. A flowchart of the identified publication is provided below in Figure 2.1.

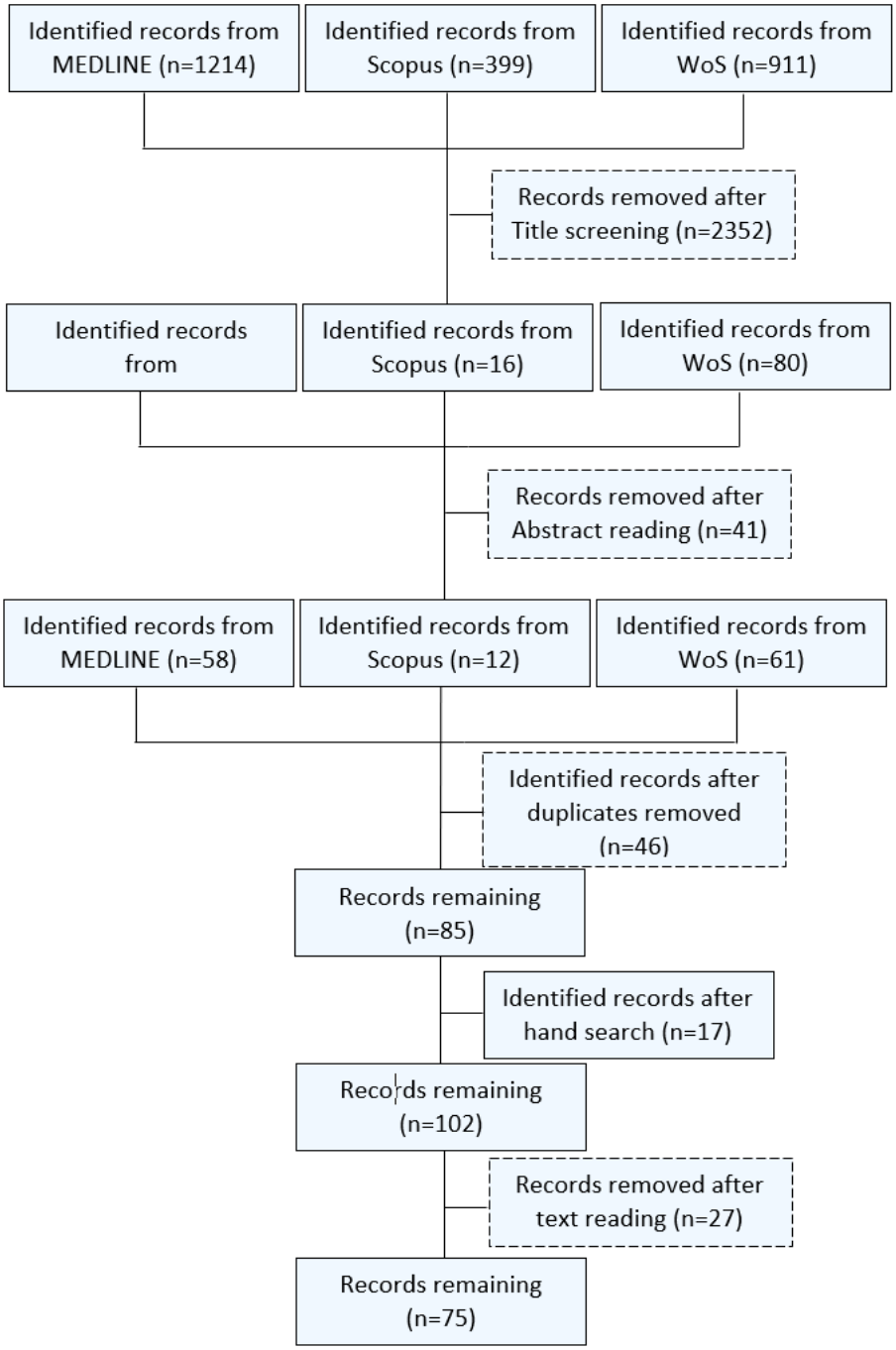


Figure 2.1: Flowchart of Methodology Review.

Literature discussing methods for performing MA of TTE data was published from 1988 onwards. Very few publications were found before 2000, whereas from the beginning of the new millennium numbers of publications increased and most research has been published during the last decade. The distribution of research publications across the years is presented in Figure 2.2.

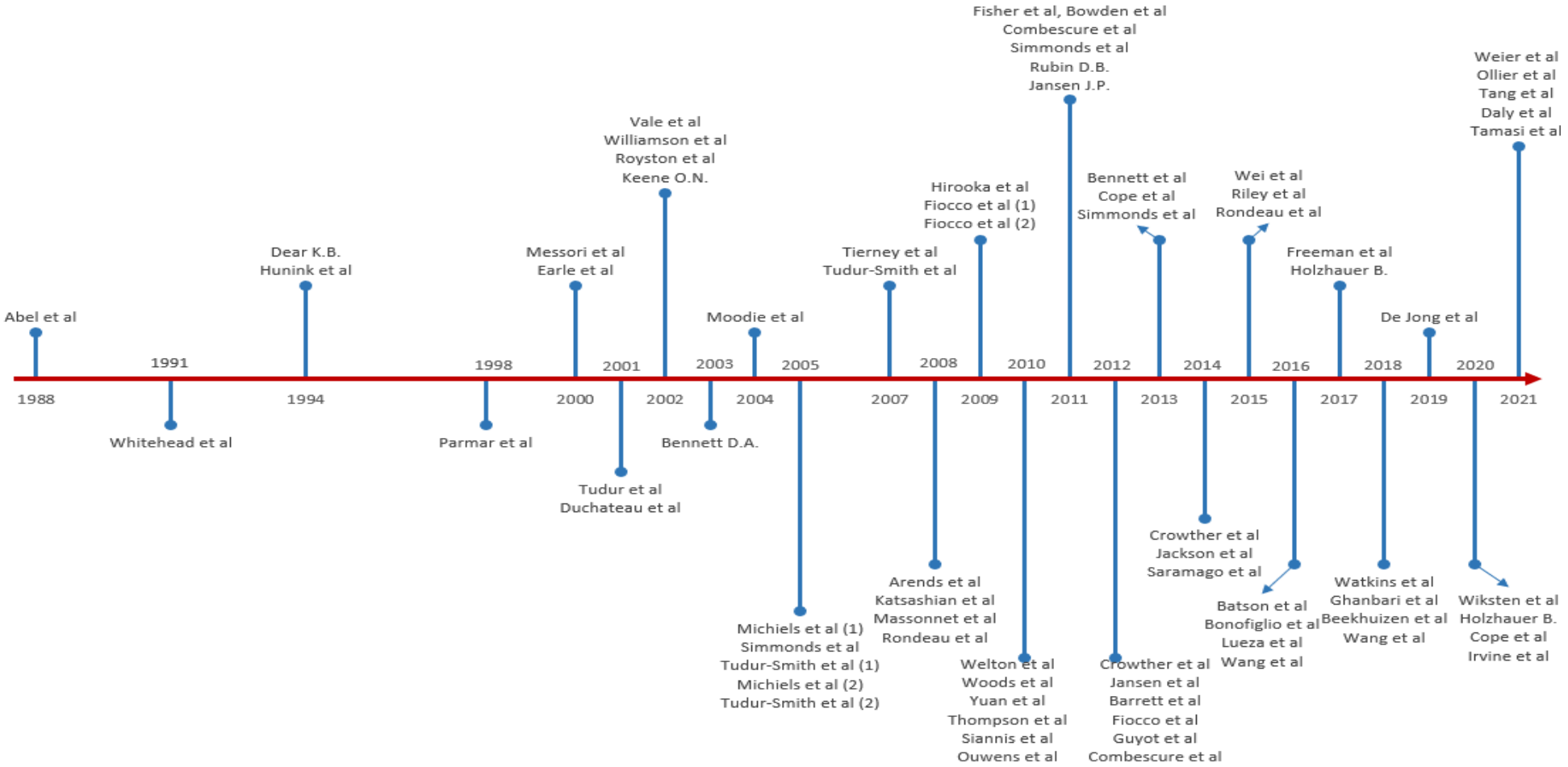


Figure 2.2: Distribution of research publications identified in methodology review

2.4 Methodology Review Results

The research papers are divided into seven main categories. These are related to models for aggregate data (11 publications), methods for reconstruction of survival data (5 publications), models for IPD (16 publications), methods for NMA (12 publications), multivariate MA (7 publications), method comparison via real life conditions and/or simulations (16 publications) and finally papers including discussions, critiques and other suggestions for MA of TTE outcomes (6 publications). I present below summaries of the methodological papers I found in the review in chronological order within each chosen theme.

2.4.1 Models for aggregate data

In 1991, Whitehead and Whitehead[48] was one of the first papers presenting comprehensive methodology on how survival (and other) data should be treated in a MA context. Particularly, they presented a general parametric approach for estimation of treatments effects based on hypothesis testing, identifying and dealing also with the issue of treatment heterogeneity in the trials. They presented both fixed and random effects models and indicated that their methodology is appropriate when large number of patients are considered; for smaller samples biased estimation should be taken into account as mentioned by Greenland and Salvan[20].

Hunink and Wong[25], in 1994, recommended a new technique combining aggregate TTE data from different sources adjusting for case-mix (i.e. different frequencies) of covariates. The authors stated that without adjusting for case-mix covariates results were misleading, providing narrow confidence intervals without accounting for the variability among subgroups. Limitations of the method such as the difficulties arising from the insufficient details from the published reports were also discussed.

One of the most important and key papers to perform MA of TTE outcomes was written by Parmar et al.[43] in 1998. The authors used a series of simple methods to extract data from publications, to facilitate the performance of MA and improve the reliability and quality of literature-based MA. More specifically, they stated that in presence of a HR and its variance from each trial these values should be used to perform a MA. If these statistics were not available other statistics could be used such as the confidence interval given for the log HR, p-values from the

log-rank[5] or Mantel-Haenszel test[49], and p-value from the Cox proportional hazard model[3] after adjustment for covariates. If limited information was available to extract from the literature for either direct or indirect estimation of the necessary statistics, published survival curves could be used to reconstruct the log HR, accounting also for censoring by adjusting for the number at risk and number of events. Full derivation of the methods was conducted by illustrative examples and the authors discuss various difficulties raised during this process such as selecting the most appropriate interval so that the event rate within each time interval is appropriate. One of the first applications of one of these methods which adjusted for censoring was conducted by Vale et al.[50] during 2002 in MAs of cancer studies, indicating that accounting for censoring answered more reliably the questions posed by systematic reviewers and meta-analysts in the presence of aggregate data.

Moodie et al.[51], in 2004, presented methodology in which the $\log(-\log)$ survival function difference for the examination of treatment effects in the MA of TTE data was used. The authors indicated the non-parametric nature of the procedure which could be used as an alternative when HRs were unavailable in the literature. The proposed methodology provided estimation of the treatment effect and was not as affected by the length of follow-up as were other effect measures such as ORs, RRs and risk difference.

In 2007, Tierney et al.[47] significantly contributed to the literature by "translating" the methods proposed by Parmar et al.[43] and Williamson et al.[44] to extract data from published reports and facilitate aggregate data MA of TTE outcomes, in a more practical way that was accessible to systematic reviewers and meta-analysts. Authors created a macro-enabled Excel spreadsheet facilitating the use of indirect methods for the calculation of HRs and their corresponding variances. This piece of research has been very widely cited (4,239 times, until 14/02/2022) and has improved the quality and interpretation of systematic reviews and MAs.

Yuan et al.[52], in 2010, discussed the fact that when different studies adjust for different covariates bias was caused when combination of the potential effect sizes is performed via MA. More specifically, authors introduced a method which combined via meta-regression incomparable Cox proportional hazards models obtained by omitting important information in an aggregate data MA.

Combescure et al.[53], in 2011 in 2011, recommended another method of meta-analysis for binary outcomes (using the relative logarithm of survival). This method was similar to Moodie's method[51] and was originally proposed by T. V. Perneger[54], when results were available in a two-by-two table, the proportionality assumption holds and information on the HR were absent. The method reduced the heterogeneity due to differential follow-up across the studies, provided a direct estimate for a HR, and can be constructed under tabulated data. In 2012, Combescure et al.[55], as an extension to Moodie's method[51], proposed a flexible method to perform meta-analysis of two-arm studies using survival probabilities collected at various time points instead of the reported measures of intervention effect, allowing detection of variation of the intervention effect over time. The authors stated that the specification of the baseline survival function was not necessary, and study level factors affecting survival were ignored. On the other hand, the method had disadvantages in the presence of non-proportional hazards, and when the assumption of a stable treatment effect over time could not be evaluated.

Bonofiglio et al.[56], in 2016, present methodology for MA of aggregate TTE data with competing risks in a frequentist framework. The authors used the cumulative incidence function to obtain cumulative incidence function ratios measuring treatment effect and developed methods to pool these ratios across studies.

B. Holzhauer[57], in 2016, discussed a Bayesian hierarchical model using aggregate data to perform TTE MA allowing differences in the follow-up period between two groups. The author performed a simulation study comparing the method to other commonly methods used for aggregate data MA, illustrating the usefulness of the exchangeability assumption that his model required, using prior information about expected control groups outcomes and increasing the power of the meta-analysis. In 2020, the same author via a simulation study compared methods for incorporating historical control data into MA. Specifically, a number of existing proposals making posterior inference more robust against prior-data conflicts were examined such as the meta-analytic combined, meta-analytic predictive in the meta-analysis setting, robust meta-analytic predictive and using a Bayesian model averaging via shrinkage priors. The simulation indicated that the last model with well-chosen hyperpriors performed best in terms of credible interval coverage and mean-squared error across scenarios.

Finally, Irvine et al[58] in 2020, developed a non-linear optimisation method for aggregate MAs that requires only the Kaplan-Meier plot and a published p-value to calculate a logHR and its variance without a requirement on the published number at risk. The method allowed for better estimation of the underlying censoring pattern and was compared to the Parmar method[43] which also did not require any published number at risk. The authors indicated that the proposed method outperformed the Parmar method, enhanced the accuracy of meta-analyses of survival outcomes and they provided the necessary R scripts for the method implementation.

2.4.2 Methods for reconstruction of survival data

Messori et al[59], in 2000, suggested an intermediate approach facilitating the reconstruction of IPD from survival curve graphs particularly in cases where the actual IPD data were not available and could not be retrieved. Via an illustrative example, the authors showed good correlation between the estimated and true IPD indicating that their approach was attractive in the absence of resources to conduct IPD MA; however, it must be noted that this approximate procedure could be less reliable than the actual IPD MA since individual survival times are approximately estimated and are not directly measured from the participants[59].

Williamson et al.[44], in 2002, extended and improved a method proposed by Parmar et al.[43], on estimating log HR from survival curves. Specifically, the extension assumed a constant censoring rate within trial intervals and varying censoring rate among time intervals. The authors examined also differences in proportionality of hazards across trials within a MA as a potential source of heterogeneity.

D. B. Rubin[60], in 2011, proposed an unbiased non-parametric approach to synthesize survival curves across studies. A counterfactual model for TTE MA was applied, avoiding confounding in non-randomised trials, a method similar to the one proposed by Xie and Liu[61] on Kaplan-Meier adjustment. Even though this method was immune to biases, the authors suggested that their method still required trials with similar characteristics and assessing similar interventions.

In 2012, Guyot et al.[62] created an algorithm which attempted to reconstruct Kaplan-Meier data obtained from survival curves, using DigitizeIt software to read

the coordinates of the Kaplan-Meier curves. A high degree of accuracy and reproducibility was observed especially for median survival times and probabilities of survival; however, reconstruction of HRs was less adequate without complete information on censoring patterns and finally the algorithm could not be used in the absence of numbers at risk and events.

Ghanbari et al.[63] proposed a new method in 2018 for combining survival curves in MA of TTE endpoints in the absence of IPD, bypassing limitations of other approaches. Specifically, the authors combined survival curves using functional data analysis which was corrected by single exponential smoothing and then a test similar to log-rank test was performed. Via a simulation study and an example using clinical data, the authors indicated the method was useful in groups with small or moderate hazard rates, moderate or high sample sizes across all studies.

2.4.3 Models for individual participant data (IPD)

Royston and Parmar[64], in 2002, suggested extensions to the Weibull and log-logistic models aiming for the estimation of hazard, density and survival functions by smoothing the cumulative odds or hazard functions. This was implemented by modelling as a natural cubic spline function of log time the logarithm of the baseline cumulative odds or hazard failure functions. Extensions of the models were suggested, facilitating allowance of non-proportional effects of the covariates.

In 2005, Michiels et al.[65] extended the use of Cox proportional hazards model by suggesting Cox random (frailty) effects models to assess the heterogeneity obtained from the variation in treatment effects from the difference in baseline hazard rates. The methods applied (under frequentist and Bayesian frameworks) used treatment-study interaction terms to adjust for study specific covariates. An assumption of a common baseline hazard function shape across trials was necessary.

In the same year, Tudur-Smith et al.[66] introduced and compared five hierarchical Cox regression models aiming to explore potential heterogeneity by including patient level covariates in IPD MA. Trial effects were included either by fixed trial effects using indicator variables or using stratification or via inclusion of random trial effects. Authors indicated that stratified models with random treatment effects

were more appropriate since they maintained the likelihood construction and allowed for different baseline hazard functions for the studies. Hazards could also be proportional within each study, and important assumptions could be relaxed when synthesising studies conducted under different settings.

During 2008, Massonnet et al.[67] presented an alternative way to fit frailty models as an equivalent linear mixed-effects model assuming a clustered data structure with random cluster and random treatment effects. The authors suggested that the method was useful for large number of clusters in the datasets and relatively large sample sizes within covariate-level subgroups in the clusters. They also highlighted the fact that standard statistical software was limited to fitting conditional random-effects survival models.

In the same year, Rondeau et al.[68] proposed a one-stage additive random-effects Cox model, similar to those proposed by Vaida and Xu[69], modelling simultaneously a random treatment-study interaction and a random trial effect. Their approach jointly accounted for different sources of heterogeneity such as a) heterogeneity in MA of treatment effects and b) heterogeneity obtained from baseline risk. The authors indicated that in the presence of MA with a large number of trials or large sample sizes with a non-zero correlation (ρ) between the two random-effects, the results obtained were more accurate.

In 2010, Thompson et al.[70] extended statistical methods for IPD MA of TTE outcomes from multiple epidemiological studies accounting for a) the shape of exposure-risk association, b) inclusion of interaction terms dividing the within and between-study information, c) the regression dilution bias that could occur from measurement error and within-person variation in confounders.

Siannis et al.[71] proposed methodology for IPD MA of TTE data using percentile ratios which are a function of survival percentile, allowing reduction to the median ratio (i.e. 50th survival percentile) estimated through an accelerated failure time model. This was an alternative methodology to the restrictive semi-parametric proportional hazard model which allowed shape parameters to vary among the trials, offering greater flexibility in the parametric representation of treatment effects. Barrett et al.[72], in 2012, extended this methodology[71] into a two-stage method for meta-analysis of percentile ratios using only Kaplan-Meier estimates. Specifically, by using these estimates for the survival function to calculate the

percentile ratios in a first stage and combining them in a second stage by univariate or multivariate random-effects meta-analysis, the authors avoided the need to make any distributional assumptions at the study level.

Crowther et al.[73], in 2012, introduced a Poisson regression model to perform one-stage and two-stage IPD MA of TTE outcomes, accounting also for random-effects, non-proportional hazards and treatment-effect modifiers. Their approach provided identical estimates to the Cox model both under frequentist and Bayesian frameworks. Heterogeneity estimates were slightly underestimated under the frequentist approach; the performance of the model under a Bayesian framework was improved.

In 2013, Simmonds et al.[74] recommended a new approach using the expectation-minimisation (EM) algorithm to fit a random-effects proportional hazards model treating random effects as missing data. The authors suggested that that their method provided suitable estimates for random effects without any biases or loss of precision.

In 2014, Crowther et al.[75] extended the methodology of parametric frailty models, incorporating multiple normally distributed random effects (including exponential, Weibull, Gompertz proportional hazards models, log logistic, log normal and generalised gamma accelerated failure time models) and using adaptive or non-adaptive Gauss-Hermite quadrature. The authors extended also the Royston and Parmar flexible parametric survival model[64] to include random effects and time-dependent effects (i.e. non-proportional hazards). They indicated that results were in agreement with previous findings; small bias and good coverage was observed on the baseline hazard parameters of the Weibull distributions.

Rondeau et al.[76], in 2015, proposed a joint frailty model for clustered TTE outcomes and a dependent terminal event (e.g. time-to-progression and death). Using a semi-parametric penalized likelihood approach, the authors showed that they could calculate the joint model parameters simultaneously, accounting for data clusters and the relationship between the outcomes.

In 2016, Wang et al.[77] introduced a method examining the patterns of heterogeneity in treatment effect across covariate values in TTE MA of IPD with a continuous covariate of interest. The Meta-STEPP (subpopulation treatment effect pattern plot for meta-analysis) method estimated treatment effects using a

continuous variable by “forming overlapping subpopulations” and estimating treatment effects for a particular subpopulation via the use of fixed-effects meta-analysis. The method yielded a weighted average of specific study treatment effects for different subpopulations. In 2018, the same authors extended their work by assessing treatment effect variation across a continuous covariate for TTE outcomes in the presence of IPD[78]. Authors indicated that Meta-STEPP tool with random effects was more conservative in assessing treatment effect variability than the fixed-effect approach, due to the larger variances produced.

De Jong et al[79] in 2019, performed a narrative review of the methods related to IPD MA of TTE data. Specifically, the authors focused on modelling frailty of trial participants across trials, heterogeneity in treatment effects, interactions, dealing with censoring and follow-up times using parametric and semi-parametric methods both in a one- and two-stage IPD MA framework. They recommended exploring heterogeneity via interactions and non-linear terms and highlighted the importance of random-effects models which account for residual heterogeneity.

In 2021, Tamasi et al., presented a one-stage IPD MA model for TTE outcomes that incorporates general normally distributed random effects into linear transformation models. The authors stated that the model could handle arbitrary random censoring patterns, could model between-study heterogeneity in baseline risks and the assumption of proportional hazards could be relaxed via the use of time-varying prognostic factor effects.

2.4.4 Methods for Network meta-analysis

Welton et al.[80], in 2010, initially developed a framework under which the synthesis of outcomes reported in clinical trials would be feasible using summary statistics such as mean or median TTE to perform NMA on the log HR scale, taking into account important assumptions for each of the models. Woods et al[81], in the same year, provided a tutorial which allowed HRs and cumulative count survival statistics to be combined in an evidence network, accounting for multi-arm trials and allowing results to be interpreted on the HR scale. They noted that median survival times could be incorporated in the models; however, they required strong assumptions of a constant hazard in each arm.

Ouwens et al.[82], suggested a NMA framework of analysing TTE data with treatment effects based on shape and scale parameters of survival curves. The authors extended the use of the Weibull model by proposing the use of two-dimensional treatment effects; shape and scale parameters were reformulated to understand better the relative treatment effects rather than assuming constant hazards. Building on the Weibull model presented initially by Arends et al.[83], the authors diversified it by using hazard over time instead of log(-log) survival proportions, evaluated random treatment-log(time) interactions, and considered evidence networks of more than two treatments.

In the absence of proportional hazards, J. P. Jansen[84] introduced a NMA technique which used fractional polynomials to model treatment effect via several parameters. The author showed that less bias occurred when IPD and aggregate data were incorporated in NMA rather than aggregate data alone especially when indications of heterogeneity, inconsistency and confounding bias were observed.

Jansen and Cope[85], in 2012, extended the models originally proposed by Ouwens et al.[82] and J.P. Jansen[84]; they presented multidimensional NMA models for TTE outcomes, accounting for treatment-covariate interactions and adjusting for bias caused by the differences in treatment effect modifiers. Their models did not rely on a proportional hazards' assumption and adjustments for any imbalances could be made. Jansen and Cope criticized their work, indicating that covariate adjustment was based on aggregate level data, and recommended that their models can only be used with study level data in the presence of limited variation in effect modifiers within studies.

Saramago et al.[86] introduced a Bayesian framework, combining jointly aggregate data (for specific follow-up) and IPD of censored TTE outcomes in NMA by extending work conducted by Woods et al.[81], and Soares et al.[87]. In their framework, IPD directly informed the distribution (likelihood), and aggregate data informed a probability estimate using a binomial likelihood for a specific subset of evidence. A common distribution for TTE data was assumed accounting for duration of follow-up in each study provided by the aggregate data. The authors discussed strengths such as the flexibility in modelling the IPD and aggregate data using treatment-covariate interactions and exploring within and across study interactions.

Freeman and Carpenter[88], in 2017, described a flexible and computationally practical approach to apply a Bayesian one-step IPD NMA of TTE data using Royston-Parmar models. The authors showed that the model presented allowed for inclusion of patient level covariates and examination of non-proportional hazards a) via inclusion of treatment-log(time) interaction and b) by allowing interaction to vary by trial in order to understand which trial is driving non-proportionality.

Watkins et al.[89] proposed a method for HR and variance derivation from reported binomial data based on a Taylor series expansion for the approximation of variance. Proportional hazards and minimal non-informative right censoring at the binomial data measurement time were necessary requirements that need to hold for the application of the method. Other methods such as digitised Kaplan-Meier curves and Bayesian analysis methods could provide more accurate estimates; however, according to the authors these could be time consuming and most of the time curve data were not always published.

Cope et al[90] in 2020, proposed a two-step approach for NMA of TTE data with a multidimensional treatment effect to overcome limitations (e.g. approximate likelihood on discrete hazards instead of a likelihood for individual event times) introduced by Ouwens et al[82] and Jansen[84] models. On the first step, for each trial arm reconstructed patient data were fit to alternative survival distributions (i.e., exponential, Weibull, Gompertz, log-normal, log-logistic); on the second step the scale and shape parameter estimates were synthesized and using a multivariate NMA model were indirectly compared across trials providing time-varying treatment effects for the competing interventions.

In the same year Wiksten et al[91] reformulated fractional polynomial and piecewise constant NMA models as generalised linear models with time-varying covariates initially introduced by Jansen[84]. The authors indicated that the proposed method allowed for rapid exploration of different frequentist NMA models allowing for the best one to be refitted in a Bayesian framework.

In 2021, Ollier et al[92] extended work conducted by Crowther et al[73] in a NMA framework. Specifically, the authors introduced a Poisson regression model for IPD NMA of TTE data allowing implementation of one-stage MA while accounting for heterogeneity and non-proportionality. The authors indicated that the

quantification of model's heterogeneity and model selection were performed simultaneously by a penalised fixed effect model, overcoming the optimization problem met when random-effects models were applied.

Tang and Trinquart[93] introduced a Bayesian multivariate NMA model for the difference in restricted mean survival times (RMST). The model synthesized simultaneously all the necessary evidence from multiple time points; via the between and within-study covariance for the difference in RMST it borrowed information across different time points. A simulation study indicated lower mean squared error and increased precision compared to a single time point model. Comparisons to previous methodology focusing on the synthesis of survival functions (e.g. Ouwens et al[82], Cope et al.[90], Jansen[84], Wiksten et al[91], Freeman et al[88]) rather than reporting differences in RMST in NMA showed improved interpretation of the findings. Their work could be described as an extension of the Weir et al[94] research (described in 2.4.5) which introduced a meta-analysis model for the difference in RMST by borrowing strength from multiple time points for conventional meta-analysis of pairwise treatment comparisons.

Finally, Daly et al[95] extended the RMST approach of Wei et al[96] (discussed in 2.4.6) in a NMA framework. Their approach jointly synthesised relative treatment effects from progression-free and overall survival Kaplan-Meier curves in a NMA without any parametric and proportionality assumptions; it also respected the constraints related to the overall survival that should be equal or greater than progression-free survival.

2.4.5 Methods for Multivariate meta-analysis

K.B. Dear[97], in 1994, presented a generalised least-squares algorithm for the analysis of survival proportions reported at multiple times, accounting for single arm trials and including also between and within trial covariates. Multiple outcomes were considered as repeated measures. Multi-arm studies and non-randomised historical controls did not require additional considerations. K.B. Dear claimed particularly that they could not incorporate a random-effects baseline term.

Arends et al.[83], in 2008, suggested a multivariate, random, mixed-effects model for simultaneous analysis of survival proportions at different time points, which

was an extension of the proposed fixed-effect model[97]. At a fixed time point the model reduced to the DerSimonian-Laird random-effects model and therefore could be seen as a generalisation of it. The authors indicated that their model allowed for investigation of non-proportional hazards and inclusion of trial and treatment interactions.

In 2009, Fiocco et al.[98] proposed a new correlated gamma frailty Poisson model using a newly constructed multivariate gamma distribution allowing for between-subjects correlation within a study. Composite likelihood was approximated using two factors, one related to parameter estimation and the other related to correlation parameter estimation[98]. The method was less sensitive to rounding errors due to the absence of quite large terms in a simulation. Bootstrap standard errors and CIs could be obtained for the parameters via simulation of the multivariate gamma distribution. This work was extended to MA of pairs of survival curves under heterogeneity, using aggregate TTE data, and suggested a simultaneous analysis of survival proportions reported at multiple time-points using a multivariate random-effects model[99]. The authors stated that their method did not require an assumption on the shape of individual survival curves, as K.B. Dear[97] and Arends et al.[83] proposed in the past; it could also deal with missing data and allowed for heterogeneity in baseline risk. Restrictions related to the estimation of lower dimension models and incomplete follow-up were discussed.

Jackson et al.[100] developed a random-effects multivariate aggregate MA model for TTE outcomes. The authors suggested modelling event probabilities using multiple time-points instead of the hazard function, avoiding any proportional hazards assumptions. Their method could be preferable in situations where crude overall survival was desired or when inferences on specific time-points' survival probabilities were needed.

Riley et al.[101] described the methodology needed for a multivariate MA, using IPD to estimate within-study correlations using also non-parametric bootstrapping methods, particularly for survival outcomes. Methodology for other outcomes has been described including continuous, binary and mixed outcomes. The method produced more appropriate standard errors particularly in the presence of longitudinal data and allowed for adjusted estimates and treatment-covariate

interactions. The authors indicated that their method could be used both under two- and one-stage model approaches.

Finally, Weir et al[94] in 2019, introduced a multivariate random-effects model for meta-analysis of the difference in RMST with IPD. The model borrowed strength from all available data at different follow-up times across trials and incorporated between-time point covariances. Unlike previous methods such as Wei et al[96] it did not rely on predictions from fitted models but incorporated all observed data at all time points of interest. In a simulation study their approach yielded in smaller mean standard error at all time points when compared to other univariate methods.

2.4.6 Method comparison and application to real life conditions and/or simulations

In 2000, Earle et al.[102] assessed five methods combining published survival curves in medical research; the iterative generalized least squares, MA of TTE data adjusting for covariates, non-linear regression, the log relative risk, and the weighted log relative risk. The authors suggested that all methods maintained reproducibility of summary survival curves from published literature, however, the best method was dependent on the data characteristics and the aim of the analysis.

Duchateau et al.[103], during 2001, compared the results from TTE outcomes from IPD MAs to those obtained from aggregate data MAs. The authors indicated that the differences mainly occurred since IPD MAs were based on duration of survival whereas aggregate data MAs were based on the cumulative mortality at specific time points.

Tudur et al.[104], in 2001, compared three indirect methods proposed by Parmar et al.[43] and an extension of the survival curve approach proposed by Williamson (via internal communication during that time). The authors indicated that estimating the variance of the log HR from a CI and estimating log HR and variance using the p-value from the log-rank test performed better compared to estimating log HR and variance from survival curves where variability in the estimates was present. In the presence of low event probability, the indirect method using

survival curves was not reliable. Finally, they suggested situations under which an aggregate data approach was adequate.

In 2005, Michiels et al.[65] compared results obtained from MAs when median survival times were used as an alternative to HRs, or ORs of survival rates. Authors found that both median survival times and OR methods could result in an important loss of statistical power and under- or overestimation of treatment effects. In the presence of lower event rates, median survival time method provided more biased results. They highlighted the necessity of collecting important information on measures such as the degrees of freedom, HR with 95% CIs or the exact p-value allowing for replication of the HRs directly from the summary statistics of the trial report.

Tudur-Smith et al.[105] compared methodology investigating heterogeneity in TTE MA for aggregate data and IPD in 2005. Aggregate data meta-regression was accurate in the presence of within-study treatment-covariate interaction in addition to the between-trial variations for the aggregate value of the covariate. Additionally to previous evidence from Lambert et al[106], the authors indicated that IPD should be used to study patient characteristics reliably and assess heterogeneity since adequate summary data are usually limited.

Tudur-Smith and Williamson[107], in 2007, compared three methods for fixed-effect IPD MA using TTE outcomes: the stratified log-rank analysis, stratified Cox regression and inverse variance weighted average of estimates. The authors indicated circumstances under which the models could produce similar estimates of the pooled log HR and its variance (when the underlying treatment effect was close to zero and the degree of heterogeneity across trials was minimal). The stratified log-rank analysis biased the results for larger treatment effects; all methods were approximately equivalent for modest treatment effects and low heterogeneity.

In 2008, Katsahian et al.[108] compared four approaches for IPD MA of TTE endpoints via simulation: the fixed-effect, random-effects (frailty), stratified and marginal models. The conditional model results differed substantially from marginal models since they were trying to address different questions. Frailty and random-effects models behaved fairly well even if few trials (i.e., not less than three) per study were present. Stratified models performed similarly to frailty

models; heterogeneity though could not be measured. Smaller type I errors and greater power were obtained from the random-effects compared to fixed-effect models when heterogeneity was explored. Finally, frailty models appeared to be the best suited models as they could handle trial-treatment interactions.

Hirooka et al.[109], during 2009, reported the performance of the estimation methods for literature based MA suggested by Parmar et al.[43] via a simulation study comparing Cox regression analysis to direct method (i.e. log HR and variance were calculated from the log-rank score and its variance calculated from data), indirect method (log HR was approximately calculated as the "(total observed events)/4"), survival curve method and the survival curve method involving Mantel-Haenzel method (i.e. modified survival curve method). The authors indicated that the direct method performed similarly to Cox regression; the indirect method was highly accurate but underestimated the effect size in presence of a large effect with large sample size and high event probability. Finally, the survival curve and modified survival curve methods underestimated the effect size for large effect size with small sample size and low event rate.

In 2011, Fisher et al.[110], evaluated four methods (the pooling of within-trial covariate interactions (PWT), one-stage model with a treatment-covariate interaction term (OSM), testing for difference between covariate subgroups in their pooled treatment effects (TDCS) and combining PWT with meta-regression (CWA)) assessing patient-level interactions in IPD MA and provided guidance on method selection. They indicated that method selection should be based upon whether across-trial information is accounted for in the analysis. PWT and CWA methods were considered important initial steps of any analysis; OSM could be a more attractive approach since it allowed for multiple parameters to be simultaneously estimated, however, methodology and software issues exist for the application of this method. TDSC could identify treatment-covariate interaction, however, it could increase the risk of ecological bias since estimates could contain both within and between-trial information.

In 2011, Bowden et al.[111] compared the performance of two-stage log-rank and Cox model methods to the one-stage methods using Cox proportional hazards model and made use of the restricted maximum likelihood (REML). Negligible bias was present in the two-stage and one-stage Cox model estimates whereas

a small amount of bias was observed with the log-rank method; the estimates were though quite similar. The coverage of the model reduced when the sample size increased in all methods, and more conservative effect estimates were obtained because of the increased variance of the HR under the random-effects model used.

Simmonds et al.[112] compared three two-stage common methods for analysis of TTE data via simulation in 2011: a hypergeometric proportional odds model (Peto method), a Cox proportional hazards and an interval-censored logistic model. The Cox proportional hazards model and interval-censored logistic regression provided generally unbiased results, with the log-rank method yielding bias for large HRs. The authors discussed the relevant implications on a meta-analysis level and suggested that maximum-likelihood methods such as Cox and interval censored logistic models should be preferred over log-rank test for MA of TTE endpoints since they are able to test if the proportionality assumption holds.

Fiocco et al.[113] evaluated in 2012 three models for MA of survival curves. The authors compared the results from the model proposed by K.B. Dear[97] using iterative least squares, a multivariate random effects model which was suggested by Arends et al[83] as an extension of the previous model and a model proposed by Fiocco et al.[98] using a Poisson correlated gamma frailty model. The same trend was observed in the estimated overall survival in the presence of heterogeneity. The Poisson correlated gamma frailty model could deal with the proportionality assumption as indicated in a simulation study; potential sources of heterogeneity between the studies were explained via inclusion of covariates at the study-level.

Bennett et al.[114], in 2013, assessed three Cox proportional hazard models for TTE data MA, two from a frequentist and one from a Bayesian perspective, considering also how these methods perform in the presence of low event rates. Based on simulation studies the Heinze and Schemper method[115] with firth correction was consistently better in predicting log HR when the event rate was low, however, all methods performed equally well when the event count was large enough.

In 2015, Wei et al.[96] evaluated one flexible parametric and two non-parametric estimation methods using restricted mean survival time (RSMT) for MA of survival

outcomes as an alternative way to the calculation of HR in the two-stage IPD MA. RMST did not require the proportional hazard assumption, allowed the treatment effect evaluation to rely on the difference in TTE, facilitated interpretation of the results and allowed trial data inclusion in a MA in the presence of trials with shorter follow-up. The authors compared the three methods via simulation, concluding that methods perform similarly well in terms of the coverage, and flexible parametric method produced smaller mean square errors under specific scenarios.

Lueza et al.[116] ,in 2016, compared methodology used to estimate the difference in the RMST from IPD MA of TTE outcomes, as Wei et al[96] did, looking at a different range of scenarios. Specifically, the authors compared the "Naïve Kaplan-Meier", the "Peto-quintile", the "Pooled Kaplan-Meier" method, and the "Pooled exponential" method. Simulation studies indicated that the Pooled Kaplan-Meier with DerSimonian-Laird random effects performed better in terms of bias and variance. Pooled exponential method showed bias in presence on non-proportional hazards; Peto-quintile underestimated the RMST apart from the case where non-proportional hazards considered; fixed effects underestimated the standard error of the RMST in most cases apart from the Pooled Kaplan-Meier and Pooled Exponential with DerSimonian-Laird random effects.

Finally in 2018, van Beekhuizen et al.[117] compared three methods for NMA of TTE outcomes: the HR NMA, the parametric survival NMAs (PNMA) and finally fractional polynomial NMAs (FPNMA). Using datasets where the proportionality assumption was either valid or violated and making outcome comparisons based on RMST, the authors indicated that all methods predicted equally well mean survival in the presence of proportional hazards, however in the absence of them, HR NMA performed worse than PNMA and FPNMA. PNMA was not very good in selecting the best fit due to option limitations and having fewer opportunities to predict survival plateaus.

2.4.7 General Discussions, Suggestions and/or Critiques on meta-analysis of Time-to-Event Outcomes

In 1988, Abel and Edler[118] were among the first researchers identifying issues with the conventional approaches of measuring relative risk in MA. Using a simple

example, the authors indicated that in the presence of time-dependent treatment effect, when the number of individuals at risk changes markedly, the true relative risk was likely to be underestimated. They claimed that estimation of relative risk should account for the time-dependence of observed and expected events at each time point.

O.N. Keene[119] described in 2002 classic approaches to the analysis of TTE data when the proportionality assumption seemed questionable and explored alternative estimates of efficacy including the HR. The author suggested that using median times to event could result in a robust measure of efficacy within a non-parametric framework, using a bootstrap method for the calculation of confidence intervals. A calculation of bootstrap standard errors for the difference in medians was also performed. In 2003, D.A. Bennett[120] provided an outline of methods necessary for the analysis of observational TTE outcomes. The authors discussed the implications for MA and the relevant areas of concern involved in MA such as publication bias, heterogeneity, misclassification and measurement error.

In 2005, Simmonds et al.[121] performed a methods review used for IPD MA. The authors indicated that MA of TTE outcomes were more apparent when IPD MA was performed; the majority of them used the Peto method, the log-rank and Cox proportional hazard models. Review authors discussed the need of developing methods incorporating heterogeneity via random-effects models; they also suggested that more clear strategies were needed in the absence of IPD.

Cope and Jansen[122], using a fractional polynomial Bayesian NMA of parametric survival curves, discussed different approaches to present rank probabilities. They looked into effect measures such as: median survival, expected survival, mean survival, mean survival of the trial with the shortest follow up time point, hazard or hazard ratio over time, cumulative survival or survival proportions over time and finally mean survival at subsequent time points. The authors indicated that the first half of the rank probabilities were easier to understand, communicated and did not vary over time whereas the other three improved the information related to the relative treatment effects over time, facilitating decision making by providing a more transparent approach.

Finally, Batson et al.[46], in 2016, published a review of the methodology reported in oncology clinical trials and its suitability for informing their inclusion in a MA. The authors indicated that serious limitations were observed in the reporting of clinical trials; they were most influenced by traditional approaches such as Cox, stratified Cox, log-rank test without justifying important assumptions of the models posed such as the proportional hazard assumption. The authors suggested that statistical methodology should be assessed by goodness of model fit and alternative approaches for MA of TTE outcomes where the proportional hazards assumption does not hold such as accelerated failure time models should be considered for valid decision making.

2.5 Discussion

This chapter aimed to identify and describe methodological research papers describing the methods for MA of TTE outcomes, without providing direct comparison or judgements among proposed methodologies. The review of the articles included was based on searches conducted in Medline (Ovid version), Scopus and Web of Science from the earliest date up to December 2021. I reviewed a total of 75 articles. The purpose of carrying out the review was to obtain an in-depth summary of relevant published literature, to inform the subsequent research presented in later chapters.

I categorised the publications into seven main categories: Models for aggregate data (11 publications), methods for reconstruction of TTE data (5 publications), models for IPD (16 publications), methods for NMA (12 publications), multivariate MA (7 publications), method comparison via real life conditions and/or simulations (16 publications) and finally papers including discussions, critiques and other suggestions for MA of TTE outcomes (6 publications). I described various methodologies including proportional hazards, non-proportional hazards, RMST, data extraction from survival curves to conduct MA, methodology for IPD and aggregate data MA, and frailty models for the examination of heterogeneity. The review identified limited publications focusing on the issue of analysing TTE outcomes as binary such Michiels et al[65]. I was able also to extract information from some research publications on the significance of the use of different effect measures.

For example, Abel and Edler[118] discussed that the estimation of the effect measure in TTE outcomes has to account for the time-dependence of observed and expected events at each time point. Via an example they indicated that calculation of the cumulative observed and expected events could lead to underestimation of the true risk. Duchateau et al.[103] claimed that in an IPD MA interpretation is conducted in terms of a HR, therefore taking into account patient's time to death, whereas aggregate data MA is often interpreted using ORs, leading to non-representative conclusions on the overall treatment effect size[123].

Additionally, Combescure et al.[53] indicated that MA of binary outcomes when censoring is present could affect the reliability of MA on TTE data, highlighting the necessity of further research in assessing the implications of censored data being present in aggregate data MA. Cope and Jansen[122], from a NMA perspective, discussed the potential advantages and disadvantages of different effect measures related to treatment rankings. Specifically, the authors claimed that ranking based upon one-dimensional measures did not yield the necessary information needed by rank probabilities whereas those based upon the HR could provide important information over time on the treatment effect at each time point[122]. Finally, a recent review conducted by Otwombe et al.[45] stressed the fact that the research using logistic regression on TTE outcomes is classified as "suboptimal" due to their failure in accounting for follow-up.

The methodology review identified the research that exists in the literature to support systematic reviewers and meta-analysts to perform MA of TTE outcomes. It has also described more complex methodologies with regards to different modelling techniques that are not necessarily aimed to be applied by systematic reviewers and meta-analysts. The review identified that most publications in the past were focusing mainly on models for aggregate data, whereas recently publications are focusing mainly on meta-analysis of IPD or NMA. The use of Bayesian techniques in recent years has been explored.

This review did not aim to collect all empirical evidence from a certain topic, but to evaluate a broad pre-specified methodological question[124]. Therefore, I did not intend to consider it as a systematic review, even though I took a systematic approach to searching and screening to identify the necessary evidence.

Furthermore, I excluded any methodological studies that were reported in languages other than English and this may have introduced language bias.

Even though the use of the complementary log-log link was not directly explored here due to the limited publications identified in the literature I focus on this for the rest of this PhD since it provides a direct interpretation on a HR scale, it is closely related to continuous-time models, has a built-in proportional hazards assumption, and therefore has important application in survival analysis.[125] The use of RR has not been explored further since according to the literature[34, 35, 39, 41], is placed in between the OR and HR measures and therefore, it is expect to capture any biases within these extremes.

To my knowledge there is one previous PhD report by Sarah Nevitt (published in 2017) identifying methodology that exists on MA of TTE data or application of the existing methods; this included more than a hundred publications[126]. However, this review had different inclusion and exclusion criteria and different databases were searched. It is important to note that the core of methodological papers (up to January 2017) that were described here were similar to those identified in the previous report, while additional methodological publications were identified in more recent years. Additionally, important contribution to the literature on analysing TTE outcomes as binary in meta-analysis was conducted by research produced by Tudur-Smith et al[127].

In conclusion, I explored and described methodological papers for MA of TTE outcomes, including discussions on the significance of the correct choice of the effect measure, including quite limited discussions on the element of analysing TTE outcomes as binary. My review indicated that many different methodologies have been proposed specifically for MA of TTE outcomes, however past reviews have indicated that their application to date is still quite limited[45, 46]. Further research is needed in order to understand how these methodologies perform comparatively when applied to different MA datasets having various characteristics, using effect measures such as the HR and OR.

3. Analysing Time-to-Event Endpoints Originally Treated as Binary on a HR scale Using Empirical Data from the CDSR

3.1 Chapter Overview

This chapter provides an empirical comparison between TTE MA analysed originally as binary in the CDSR and interpreted on the OR scale with MA results from analyses performed using the complementary log-log link (clog-log) and interpreted on the HR scale. I describe the reasons for performing these analyses, the statistical models and comparisons conducted, and present the results, together with a discussion of the findings.

3.2 Introduction

Systematic reviews and MA of TTE outcomes (e.g. time to death, recurrence of symptoms, time to conception, relief of pain etc.) are frequently carried out and are very common in areas such as cancer, respiratory and cardiovascular, since event timings are crucial to assessing the impact of an intervention[47].

The decision on how TTE outcomes are handled in a particular meta-analysis largely depends on how eligible studies are reported and is often out of the control of the meta-analyst except if individual participant data (IPD) are available. The information extracted by systematic reviewers may include the total number of participants and events per arm, and/or the HR alongside its CI, and/or the log-rank observed minus expected statistic ("O-E") and its variance ("V") (which are useful alternative statistics if a hazard ratio is not directly reported[47]).

Time-to-event data can be analysed using the effect measure of hazard ratio (HR), or can be dichotomised and analysed as binary using effect measures such as the odds ratio (OR) or risk ratio (RR)[16]. Although HR is considered the most appropriate scale for analysis of TTE data, in practice OR and RR are frequently used instead due to the following reasons: unavailability of individual participant data (IPD); limitations on how these outcomes are reported in individual trial reports; lack of familiarity in handling TTE outcomes for meta-analysis; difficulties in understanding the methods of analysing such data without a statistician; limited available training for the majority of systematic reviewers and meta-analysts who perform such analyses[42].

Discussions have been raised in the past and are still ongoing over how TTE outcomes should be analysed in a MA. Since TTE data take into account the timing and censoring of the events, strong assumptions are made if these data are dichotomised, ignoring their natural properties and treating them as any other binary outcome. This could have a serious impact on the final pooled effect estimates, potentially producing misleading decisions on the appropriateness of healthcare interventions, which in turn could adversely affect patient heath and healthcare services, or lead to initiation of new trials which may not be cost-effective.

In the past, research was conducted comparing the differences between the OR using logistic regression models and the HR using proportional hazard (PH) models within individual studies. Green and Symons[33] showed that logistic and Cox proportional hazard models produce similar results when the event is rare and for shorter follow-up times under a constant hazard rate. Ingram and Kleinman[37] added that important differences among the methods occur in the presence of varying censoring rates and length of follow-up. However, it has not been established yet how such results transfer to the context of an aggregate data meta-analysis for which summary data is extracted from trial reports. Further, in this context it is of interest to examine potential alternatives such as the use of the complementary log-log (clog-log) link, which may reduce the difference in the results between the two effect measures used. The overall meta-analytic estimate can be affected due to changes to the weighting allocated to each study, and therefore changes to the results of a meta-analysis can be unpredictable.

In this chapter I aimed to carry out an empirical “meta-epidemiological” study using survival meta-analysis data from the Cochrane Database of Systematic Reviews (CDSR) (Issue 1, 2008) to explore the implications of analysing TTE outcomes as binary in meta-analysis. Since only binary data were available I examined the impact of using alternative methodology such as the complementary log-log link (clog-log), proven to facilitate interpretation of the results on a HR scale.[125, 128] I assess only the differences between the OR and the HR, as the RR, according to the literature[34, 35, 39, 41], is placed in between these measures and therefore, I expect to capture any bias within these extremes. I perform these analyses under both two- and one-stage models.

The rest of this chapter is set out as follows. In Section 3.3, I describe the dataset I used and the statistical models that I applied. In Section 3.4, I present descriptive statistics of the database and then I describe the results obtained from re-analysing the data originally analysed as binary on an HR scale in two subsections: one for the two-stage and one for the one-stage models. These results are followed by a discussion exploring the strengths and limitations of my findings in Section 3.5, together with conclusions and plans for further work.

3.3 Methods

3.3.1 Data

The Nordic Cochrane Centre provided the content of the first issue from 2008 of the CDSR and includes meta-analyses within reviews which have been classified previously by outcome type, medical specialty and types of interventions included in the pairwise comparisons[129]. The database did not record whether data type was TTE; however, based on the outcome classification I was able to identify (using words such as “survival”, “death”, “fatality”) meta-analyses with outcome classification “all-cause mortality” where the information recorded was based only on the number of events and participants per arm. Therefore, a first subset of TTE MA was identified; those recorded as binary summaries.

3.3.2 Eligibility Criteria

Rebecca M. Turner previously extracted these binary data and conducted initial cleaning including examination of the outcome classification; I repeated the data extraction to confirm the information obtained were accurate. The dataset could contribute more than one meta-analysis per Cochrane review.

RMT identified 30 misclassifications due to disagreement with the original outcome classification as listed in the datasets, conflicting information in the database or unavailability of the correct version of the Cochrane review, leaving 1,102 MA in the dataset. I excluded 1,252 studies including double zero events, since they do not contribute to the meta-analysis results[48, 129]. I removed another 352 meta-analyses including fewer than 3 studies because some of the models applied below (i.e. generalised linear mixed models) will be affected by estimation issues and inevitable failures using small numbers of studies[130]; hence I wanted to make fair comparisons between the models applied. Derivation of the analysis sample is provided in Figure 3.1.

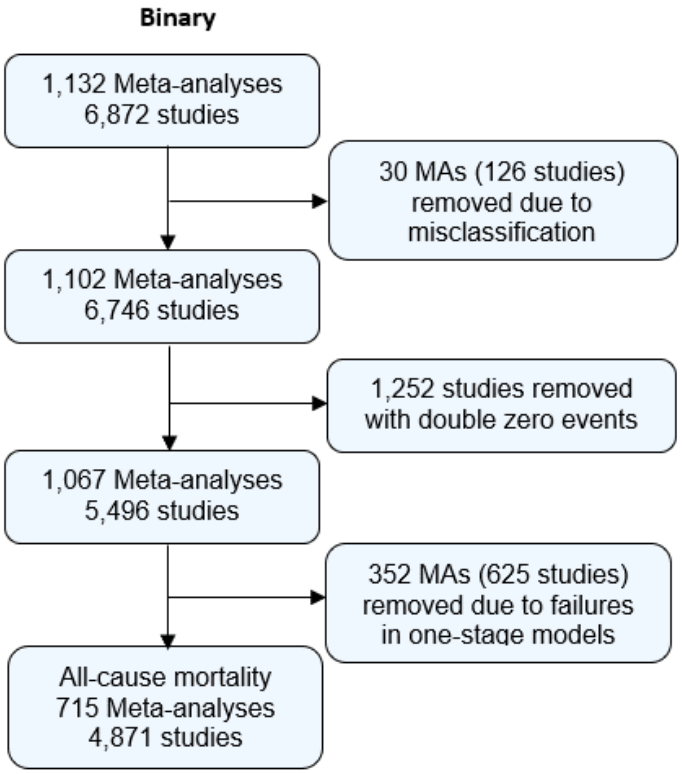


Figure 3.1: Analysis sample of binary dataset from the CDSR (2008, Issue 1).

3.3.2 Descriptive Statistics

Prior to analysis, I obtained descriptive statistics on the number of studies per meta-analysis, number of events and study size by the median and interquartile range (IQR). I identified also the number of medical specialities, and median number of events (and IQR) per medical specialty.

3.3.3 Model description

I used the following meta-analysis models to analyse the data on the OR or HR scale. The model was initially presented by Jackson et al.[130] on meta-analysis level and by T.V Perneger[54] on a single study level; I applied it on the HR scale. The first was a model proposed for binary data (assuming a binomial likelihood with a logit link) which is based only on the number of patients and number of events which occurred. Interpretation for the treatment effect is conducted in terms of the logarithm of an OR.

In the second approach, I modelled the binary data using a normal approximation to binomial likelihood with a complementary log-log link (clog-log), where treatment effect interpretation was based on the logarithm of a HR. This method is also based only on the number of patients and events which occurred, and ignores censoring and the time element; however it is closely related to continuous-time models, has a built-in proportional hazards assumption, and therefore has important application in survival analysis[125]. More details on this approach are presented in Appendix B.3 (part A).

3.3.3.1 Fitting two-stage random-effects models for binary data

Prior to fitting the two-stage random-effects models, study arms with zero events were identified for the binary data. For 771 studies, a "treatment arm" continuity correction was applied as proposed by Sweeting et al.[131] and was constrained to sum to one as this ensures that the same amount of information is added to each study. Specifically, the reciprocal of the size of the opposite treatment arm was added to both cells (m/n where m is a constant of a chosen size and n represents the total amount of patients randomised in the opposite arm).

Let $\mathrm{i} = 1,2,\dots,\mathrm{n}$ denote the study. The estimated log odds and log hazard ratios were given by:

$$\mathrm{y_i} = \begin{cases} \log\left(\frac{\mathrm{A_i}}{\mathrm{B_i}}\right) - \log\left(\frac{\mathrm{C_i}}{\mathrm{D_i}}\right) \text{ for ORs} & (3.1) \\ \log[-\log(1-\mathrm{P_{Ti}})] - \log[-\log(1-\mathrm{P_{Ci}})] \text{ for HRs} & (3.2) \end{cases}$$

where $\mathrm{A_i}, \mathrm{C_i}$ represented number of events, $\mathrm{B_i}, \mathrm{D_i}$ represented number of non-events in the treatment and control groups respectively, $\mathrm{P_{Ti}} = \frac{\mathrm{A_i}}{\mathrm{A_i}+\mathrm{B_i}}$ was the proportion of events on the treatment arm of the i^{th} study, and $\mathrm{P_{Ci}} = \frac{\mathrm{C_i}}{\mathrm{C_i}+\mathrm{D_i}}$ was the proportion of events on the control arm of the i^{th} study.

The corresponding variances were given by:

$$s_i^2 = \begin{cases} \frac{1}{A_i} + \frac{1}{B_i} + \frac{1}{C_i} + \frac{1}{D_i} & \text{for ORs (3.3)} \\ \left(\frac{1}{\log(1-P_{Ti}) * (P_{Ti}-1)}\right)^2 * \left(\frac{P_{Ti} * (1-P_{Ti})}{A_i + B_i}\right) + \left(\frac{1}{\log(1-P_{Ci}) * (P_{Ci}-1)}\right)^2 * \left(\frac{P_{Ci} * (1-P_{Ci})}{C_i + D_i}\right) & \text{for HRs (3.4)} \end{cases}$$

Equations (3.2) and (3.4) provided a HR estimate via the use of the clog-log link considered as a useful link function for the discrete-time hazards models as recommended by Hedeker et al.[128] and Singer et al.[125]. More information on the derivation of the HR estimate using the clog-log link is provided in Appendix B.3 (part B). I estimated the study-specific log odds ratios or log hazard ratios, y_i and their within-study variances s_i^2 as shown above and fitted a standard two-stage random-effects model to these. Additionally, I obtained the I^2 statistic from the fitted models as follows:

$$I^2 = \frac{\hat{\tau}^2}{\hat{\tau}^2 + \hat{\sigma}^2} \quad (3.5)$$

where τ^2 denotes the variance of the underlying true effects across studies and σ^2 the typical within-study variance.

To avoid downward bias in the variance components estimates, I used the restricted maximum likelihood (REML) estimator for model implementation[132]. The models were implemented via the "rma.uni" command from "metafor" package in R (Appendix B.2).

3.3.3.2 Fitting one-stage random-effects models for binary data

The following model is a generalised linear mixed model which enables us to perform the analysis in one-stage with the possibility to provide us with more accurate inferences. Let $i = 1,2,\dots,n$ denote the study and $k = 0,1$ denote the treatment group ($k = 0$ indicates control and $k = 1$ indicates active treatment group). Assuming that π_{ik} is the event probability in the i^{th} study for the k^{th} treatment group.

A generalised linear mixed model was fitted first from Jackson et al.[130] for the binary data and I extended it to the HR scale. This model uses the exact binomial

likelihood allowing us to provide more accurate results, especially with sparse data[130]. According to the Simmonds and Higgins model[133] I assumed that:

$$g(\pi_{ik}) = \gamma_i + k \cdot \theta_i,\ \theta_i \sim N(\theta, \tau^2) \quad (3.7)$$

where $g(\pi_{ik})$ is a link function with:

$$g(\pi_{ik}) = \begin{cases} \text{logit}(\pi_{ik}) \text{ for ORs} \\ \log[-\log(1-\pi_{ik})] \text{ for HRs} \end{cases}$$

γ_i was the baseline risk of event in study i, θ was the overall treatment effect across studies, τ^2 was the heterogeneity across studies, θ_i was the true study-specific treatment effect which varies between studies. Using the "glmer" function in R, I obtain the following:

$g(\pi_{ik}) = \gamma_i + k \cdot \theta + k \cdot \varepsilon_i$, where $\varepsilon_i \sim N(0, \tau^2)$ and all ε_i are independent. I applied to this dataset a modification of the Simmonds and Higgins model with random treatment effects and fixed study-specific effects indicating that there is a separate baseline risk parameter γ_i for each study as follows:

$$g(\pi_{ik}) = \gamma_i + k \cdot \theta + z_{ik}\varepsilon_i \quad (3.8)$$

I replaced $k \cdot \theta$ from the above equation with $z_{ik}\theta = (k - 0.5)\theta$. The model's form does not change and $z_{ik}\theta$ is only a re-parameterisation of the model as described in detail by Jackson et al.[130]. The difficulty related to using common study specific effects is that of the number of parameters needing to be estimated since the asymptotic theory of maximum likelihood requires the number of parameters to remain stable as the sample size increases[130]. The original and the modified versions of the Simmonds and Higgins model are similar with the same mean and variance, however they have different bivariate structures (i.e. the variance covariance matrix is defined in a different way). No serious bias in the estimation of variance components was found using REML in fitting this model in a small scale simulation study using continuous data conducted by Jackson et al.[130].

The "rma.glmm" command from "metafor" package was used to calculate the one-stage ORs and the "glmer" command from "lme4" package was used for the corresponding HR estimates (Appendix B.2). Estimation of between-study heterogeneity (I^2) for the one-stage HR models was considered computationally intensive[134] and was computed outside the model specification (Appendix B.4);

to provide justification for the method of calculation, I^2 estimates were obtained similarly for the one-stage OR models and were compared to the directly modelled ones, indicating almost identical results.

3.3.3 Model Comparison

The following model comparisons were performed. Initially, I examined whether the results from analysing TTE data as binary on an OR scale are similar to results from analysing on the HR scale using the clog-log link, both under two-stage and one-stage models.

First, I calculated the proportion of significant and non-significant meta-analytic pooled effect estimates under the different scales used (OR vs HR scale); I identified the number of meta-analyses which were significant under one scale and non-significant under the other at a two-sided 5% level of significance.

Bland-Altman plots with associated 95% limits of agreement were constructed, with the aim of facilitating interpretation of results and producing fair comparisons between the two scales[135]. In order to create these plots, results were standardised by dividing the logarithm of the estimate by its standard error. Plots were produced for the standardised treatment effect estimates and for the I^2 statistics. I^2 represents the percentage of variability that is due to between-study heterogeneity rather than chance; I^2 values range from 0% to 100%. This measure was chosen for model comparison as it enables us to compare results directly between the two scales used. The variance of underlying true effects across studies (τ^2) was not used as it does not allow direct comparison between different outcome measures. Finally, I examined whether the difference between standardised estimates on the treatment effects between the OR and HR scales is associated with level of baseline risk in individual studies.

I identified “outliers” as meta-analyses outside the 95% limits of agreement, and I examined their characteristics. The meta-analysis characteristics I examined were the following:

- between-scale differences in the magnitude of the pooled treatment effect estimate and its 95% confidence intervals
- the levels of within-study standard error and between-study heterogeneity and study weights in the meta-analysis

- study-specific event probabilities and baseline risk

I summarised these differences by meta-analysis and reported those characteristics which were mostly associated with substantial differences between OR pooled effect estimates and corresponding HR pooled effect estimates.

3.4 Results

For the outcome of "all-cause mortality", 1,132 meta-analyses within the Cochrane database were originally analysed as binary; after applying the exclusion criteria 715 meta-analyses were explored further. The median number of meta-analyses per review was 1 with IQR (1,2). The median number of studies and the median number of events are provided in Table 3.1.

Outcome	**All-cause Mortality**
Total Number of MA	715
Number of studies per MA: Median (IQR)	5 (3, 8)
Number of events per MA: Median (IQR)	13 (4, 40)
Median Study Size (IQR)	124 (60, 312)

Table 3.1: Descriptive statistics for binary data from the Cochrane Database of Systematic Reviews (Issue 1, 2008).

The distribution of medical specialities of the meta-analyses is presented in Table 3.2. For these data, "Cardiovascular" (23%) is the most frequently occurring category, followed by "Cancer" (13%), "Gynaecology, pregnancy and birth" (12%) and "respiratory diseases" (12%). The median number of events in cancer substantially exceeded the median number of events in other medical areas.

Medical Specialty	**ACM^{+} Number (%) of MAs**	**Events per MA: Median (IQR)**
Cancer	95 (13%)	49 (17, 120)
Cardiovascular	168 (23%)	14 (4, 43)
Central nervous system/musculoskeletal	44 (6%)	12 (5, 33)
Digestive/endocrine, nutritional and metabolic	71 (10%)	7 (3, 18)
Gynaecology, pregnancy and birth	87 (12%)	7 (2, 20)

Infectious diseases	46 (6%)	18 (8, 47)
Mental health and behavioural conditions	21 (3%)	2 (1, 5)
Pathological conditions, symptoms and signs	5 (1%)	9 (2, 15)
Respiratory diseases	87 (12%)	11 (5, 36)
Urogenital	30 (4%)	4 (2, 12)
Other*	61 (9%)	9 (3, 27)

*Other: Blood and immune system, General heath, Injuries, Mouth and dental, and Cystic fibrosis.

[+]ACM: All-cause mortality

Table 3.2: Distribution of medical specialties for the binary data meta-analyses in the CDSR.

Once the models were applied, we compared results between OR and HR analyses. Table 3.3 provides the percentages of significant and non-significant meta-analyses at a two-sided 5% level of significance indicating that there are few discrepancies present under two- or one-stage models.

Outcome		**Two-stage OR**		**One-stage OR**	
		Significant	Non-Significant	Significant	Non-Significant
HR (clog-log) All-cause Mortality	Significant	106 (15%)	2 (0.1%)	123 (17%)	2 (0.3%)
	Non-Significant	4 (0.6%)	603 (84%)	4 (0.6%)	589 (82%)

Table 3.3: Number (%) of (non-)significant meta-analyses under different scales for two- and one-stage models.

3.4.1 Results for Two-stage models

According to the Bland-Altman plot (Figure 3.2), the average difference between the two methods for the standardised pooled effect estimates was -0.004 units (-0.222 units, 0.214 units) and -0.1% (-10.6%, 10.3%) for the estimation of I^2 for two-stage models; this indicates a relatively small percentage difference between the two methods in the estimation of the measure of impact of heterogeneity I^2. The width of the 95% limits of agreement is small, indicating acceptable

agreement between the two methods except in specific circumstances mentioned below.

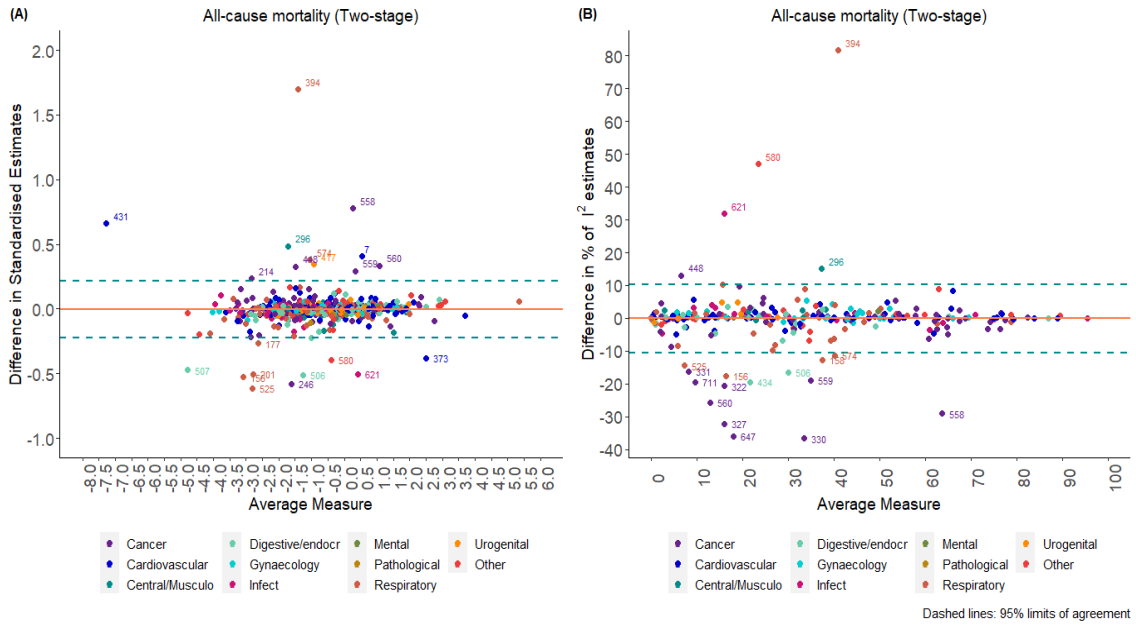


Figure 3.2: Bland-Altman plots comparing standardised pooled effect and I^2 estimates for two-stage models.

Based on Bland-Altman plots, 6% (n=47) of the meta-analyses were considered as outliers both under two- and one-stage models. In 21% of the outlying meta-analyses (e.g., MA 327; outlier obtained from I^2 estimates) a high event probability (defined here as probability greater than 0.7 for the majority of the individual studies) was observed. For example, meta-analysis 327 consists of 7 studies for which the event probability was greater than 0.7 for 5 out of 7 studies; consequently, high event probability affected substantially the difference in the individual study estimates between the OR and HR analyses, leading to different allocated relative weights for the studies, and discrepancies in the pooled effect estimates as shown in Figure 3.3. Other examples of meta-analyses under the same category include MA 246, 322, 331, 394,448, 559, 711 (Appendix B.5).

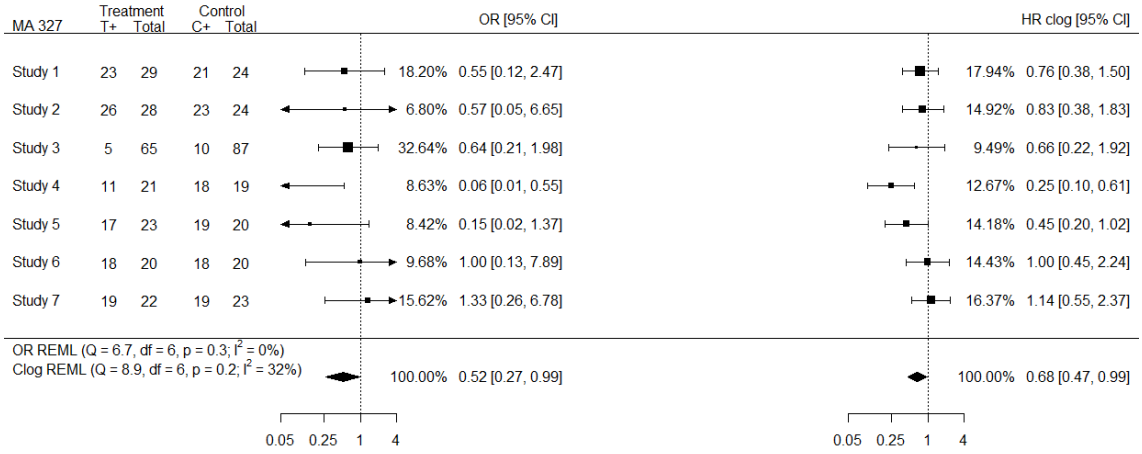


Figure 3.3: Forest plot (MA 327) indicating discrepancies in the presence of high event probability.

The pooled HR estimates were closer to 1 than the OR estimates in the majority of meta-analyses with some exceptions such as MA 574 (outlier obtained from standardised and I^2 estimates) for binary data where, even though most of the individual study HR estimates are closer to 1 than the individual OR estimates, the pooled HR estimate is further from 1 than the pooled OR estimate. Other MA under the same category include MA 417, 621, and 647 (Appendix B.5).

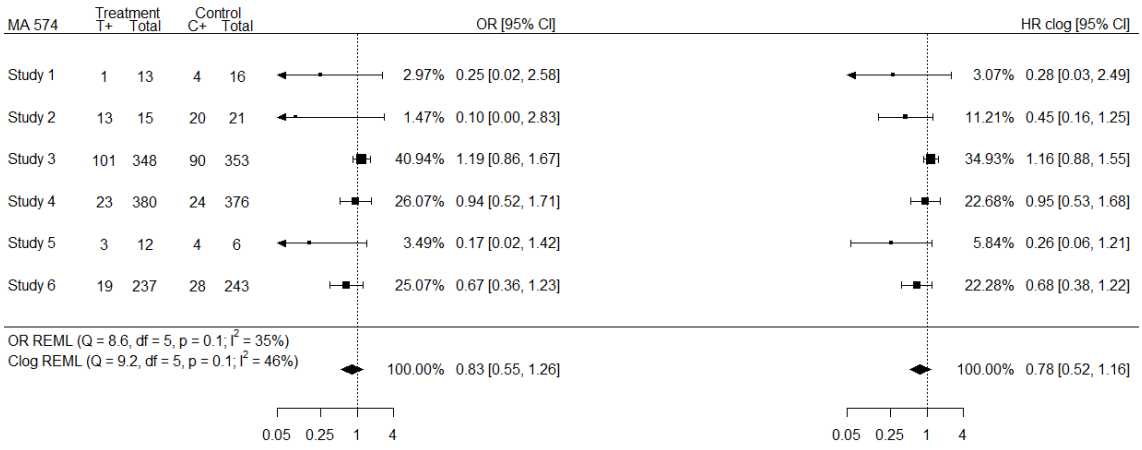


Figure 3.4: Forest plot (MA 574) in which pooled OR estimate is closer to one than pooled HR estimate.

Increased within-study variability on the OR scale relative to the HR scale may affect the weighting more than the actual estimates in the studies, for example within binary data meta-analysis 7 (outlier obtained from standardised estimates),

producing some differences in the pooled effect estimates between the two scales. Other outlier MAs lying under the same category are 156, 201, 214, 373, 431.

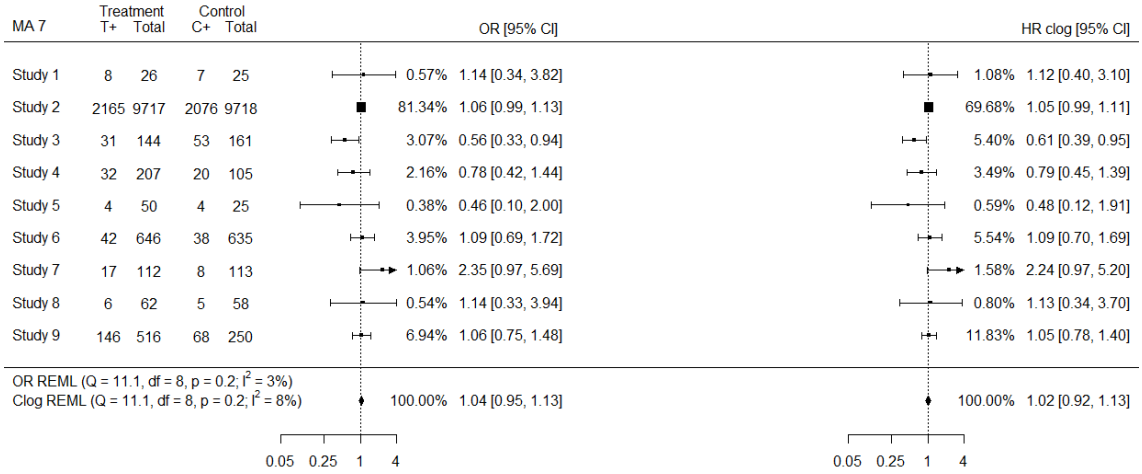


Figure 3.5: Forest plot (MA 7) showing increased within-study variability on the OR scale relative to the HR scale.

Important differences in between-study heterogeneity between the HR and OR analyses were also observed (MA: 330, 434, 506). For example, meta-analysis 330 (outlier obtained from I^2 estimates) consists of 8 studies of which 6 are smaller studies which received increased weight in the HR analysis compared to the OR analysis while the two larger studies received smaller weights; this affected both the individual HR estimates that have moved closer to each other and the relevant weights of the studies as presented in Figure 3.6.

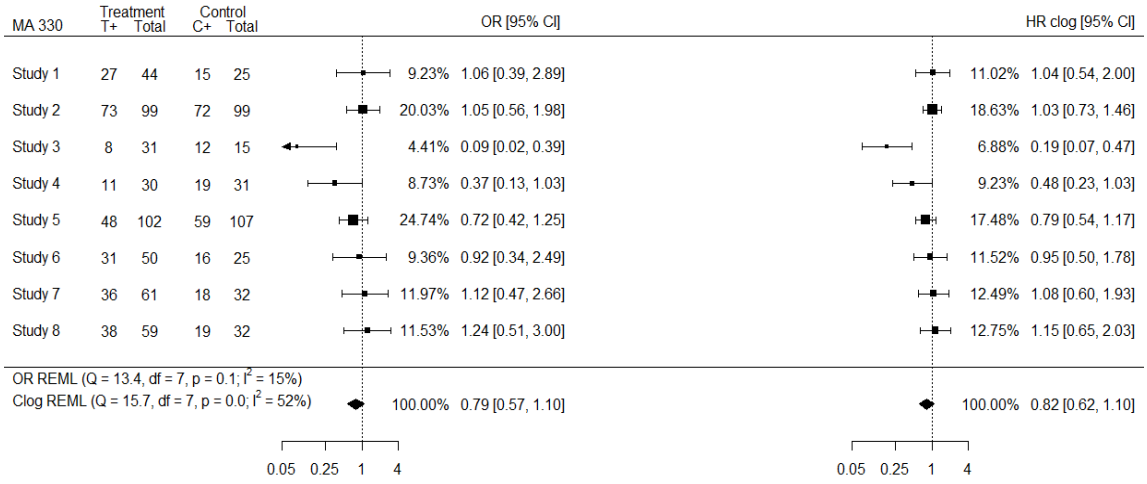


Figure 3.6: Forest plot (MA 330) indicating discrepancies arising from differences in between-study heterogeneity.

In 34% of the outlying meta-analyses (e.g., MAs 158, 177, 296, 507, 525, 558, 560), the individual study estimates and the corresponding weights were affected by a combination of differing event probability across study arms, differences in between-study heterogeneity or increased within-study variability on the OR relative to the HR scale (Figure 3.7). In the presence of a limited amount of studies in the meta-analyses this was even more evident. Additional examples of forest plots indicating the discrepancies among the results are shown in the Appendix B.5, including tables presenting the treatment effect, I^2 and τ^2 estimates.

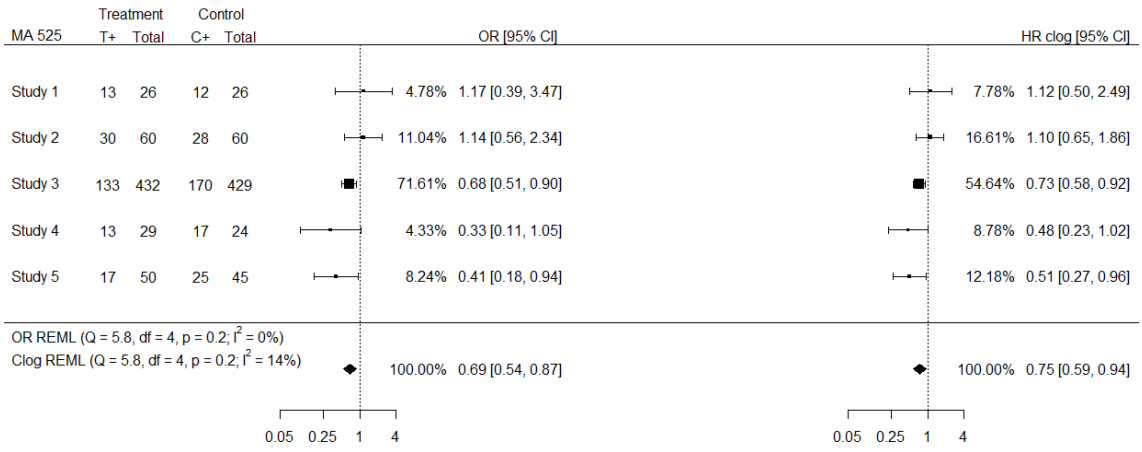


Figure 3.7: Forest plot in which a combination of reasons affect differences between the OR scale and the HR scale.

Finally, examination of the effect of baseline risk on the differences between OR and HR estimates shows somewhat greater differences for central values of control risk (Figure Appendix B.1, values between 0.3 and 0.7).

3.4.2 Results for One-stage models

In a similar way to the two-stage models, Bland-Altman plots were obtained to identify any potential discrepancies in the results once one-stage models were applied. According to the Bland-Altman plot (Figure 3.8), the average difference between the two methods for the standardised pooled effect estimates was -0.008 units (-0.340 units, 0.324 units) and -0.1% (-6.5%, 6.4%) for the estimation of I^2 for one-stage models; this indicates an even smaller percentage difference between the two methods in the estimation of the measure of impact of heterogeneity I^2 . The width of the 95% limits of agreement is small, indicating acceptable agreement between the two methods.

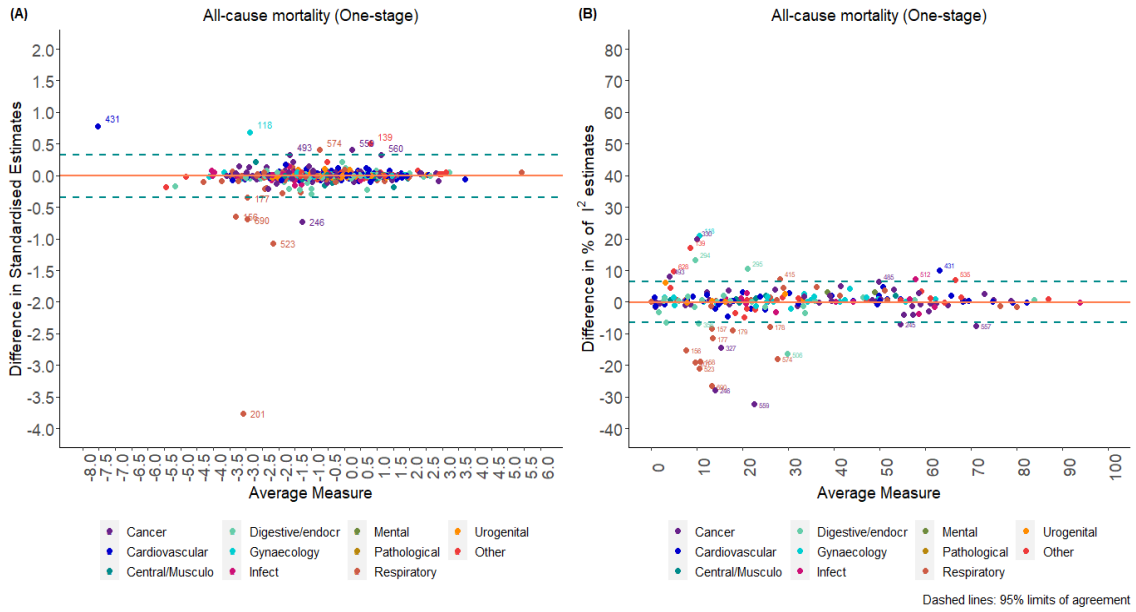


Figure 3.8: Bland-Altman plots comparing standardised pooled effect and I^2 estimates for one-stage models.

Since there were no available forest plots for one-stage models, only a table was produced for the differences in the treatment effect, I^2 and τ^2 estimates between the OR and HR analyses (Table 3.4). Similarly to two-stage models, high event probability was defined here as probability greater than 0.7 for the majority of the individual studies) was observed for some of the outlying meta-analyses (e.g. MA 245, 246, 327).

The pooled HR estimates were closer to 1 than the OR estimates for the majority of the MA with the exception of MA 201 and 574 where, even though most of the individual study HR estimates are closer to 1 than the individual OR estimates, the pooled HR estimate is further from 1 than the pooled OR estimate.

Increased within-study variability on the OR scale relative to the HR scale may have affected the weighting more than the actual estimates in the studies, for example within meta-analysis 628. Important differences in between-study heterogeneity between the HR and OR analyses were also observed for a lot of meta-analyses such as MA 157, 178, 179, 294, 295, 330, 356, 415, 431, 557, 690.

In 43% of the outlying meta-analyses, (MA 118, 139, 156, 158, 177, 485, 493, 506, 512, 523, 535, 559, 560) the individual study estimates and the corresponding weights were affected by a combination of differing event probability across study arms, differences in between-study heterogeneity or increased within-study variability on the OR relative to the HR scale.

One-Stage Random-Effects Model			
MA Identifier	OR (95% CI) vs. HR (95% CI)	τ^2 OR vs. τ^2 HR	I^2 OR vs. I^2 HR
118	0.531 (0.326, 0.864) vs. 0.565 (0.399, 0.798)	0.089 vs. 0.000	21% vs. 0%
139	1.165 (0.876, 1.550) vs. 1.027 (0.935, 1.128)	0.065 vs. 0.000	17% vs. 0%
156	0.601 (0.457, 0.790) vs. 0.653 (0.494, 0.863)	0.000 vs. 0.021	0% vs. 15%
157	0.757 (0.521, 1.099) vs. 0.815 (0.582, 1.140)	0.018 vs. 0.025	9% vs. 17%
158	0.730 (0.538, 0.990) vs. 0.795 (0.615, 1.029)	0.001 vs. 0.011	1% vs. 20%
177	0.657 (0.506, 0.854) vs. 0.701 (0.546, 0.899)	0.009 vs. 0.019	8% vs. 19%
178	0.711 (0.545, 0.927) vs. 0.753 (0.592, 0.958)	0.024 vs. 0.026	22% vs. 30%
179	0.729 (0.568, 0.935) vs. 0.773 (0.620, 0.964)	0.009 vs. 0.011	13% vs. 22%
201	0.235 (0.133, 0.416) vs. 0.210 (0.017, 2.645)	0.000 vs. 0.278	0% vs. 19%
245	0.577 (0.348, 0.957) vs. 0.738 (0.548, 0.993)	0.262 vs. 0.096	51% vs. 58%
246	0.717 (0.482, 1.066) vs. 0.902 (0.723, 1.124)	0.000 vs. 0.028	0% vs. 28%

294	0.700 (0.390, 1.256) vs. 0.760 (0.463, 1.247)	0.101 vs. 0.012	16% vs. 3%
295	0.701 (0.385, 1.278) vs. 0.793 (0.470, 1.337)	0.166 vs. 0.055	26% vs. 16%
327	0.446 (0.232, 0.855) vs. 0.676 (0.478, 0.956)	0.070 vs. 0.049	8% vs. 22%
330	0.768 (0.532, 1.110) vs. 0.873 (0.725, 1.051)	0.046 vs. 0.000	20% vs. 0%
356	0.818 (0.657, 1.018) vs. 0.846 (0.689, 1.039)	0.005 vs. 0.007	7% vs. 14%
415	0.880 (0.407, 1.900) vs. 0.924 (0.473, 1.804)	0.224 vs. 0.097	32% vs. 24%
431	0.500 (0.414, 0.605) vs. 0.569 (0.495, 0.654)	0.106 vs. 0.048	68% vs. 58%
485	1.179 (0.910, 1.529) vs. 1.102 (0.949, 1.279)	0.104 vs. 0.029	53% vs. 47%
493	0.792 (0.584, 1.074) vs. 0.857 (0.726, 1.011)	0.008 vs. 0.000	8% vs. 0%
506	0.664 (0.383, 1.151) vs. 0.771 (0.512, 1.162)	0.144 vs. 0.160	21% vs. 38%
512	0.503 (0.217, 1.165) vs. 0.576 (0.311, 1.066)	0.675 vs. 0.329	61% vs. 54%
523	0.733 (0.585, 0.918) vs. 0.805 (0.619, 1.045)	0.000 vs. 0.018	0% vs. 21%
535	0.390 (0.023, 6.565) vs. 0.550 (0.081, 3.729)	3.889 vs. 1.759	70% vs. 63%
557	1.293 (0.469, 3.566) vs. 1.269 (0.523, 3.082)	0.660 vs. 0.469	67% vs. 75%
559	1.246 (0.462, 3.359) vs. 1.008 (0.631, 1.611)	0.040 vs. 0.056	6% vs. 39%
560	1.437 (0.830, 2.487) vs. 1.231 (0.809, 1.872)	0.000 vs. 0.000	0% vs. 0%
574	0.870 (0.534, 1.418) vs. 0.788 (0.485, 1.280)	0.035 vs. 0.066	19% vs. 37%
628	0.730 (0.168, 3.165) vs. 0.821 (0.449, 1.502)	0.125 vs. 0.000	10% vs. 0%
690	0.434 (0.264, 0.712) vs. 0.479 (0.276, 0.832)	0.000 vs. 0.129	0% vs. 27%

Table 3.4: Results from meta-analyses outside the 95% limits of agreement based on difference of standardised estimates and difference in I^2.

MA coloured in **blue** represent results from studies outside the 95% limits of agreement based on difference of standardised estimates. MA coloured in **red** represent results from studies outside the 95% limits of agreement based on difference in I^2. MA coloured in **black** represent results from studies outside the 95% limits of agreement based on both difference of standardised estimates and difference in I^2.

3.5 Discussion

Using meta-analysis data from the CDSR of 2008, I investigated how TTE outcomes are sometimes treated within meta-analysis; I explored the differences that occur when data are analysed as binary as opposed to analysing the data using the complementary log-log link where interpretation is conducted on a HR scale. For this dataset, I identified important reasons associated with discordance among the results, indicating that the correct choice of the method does matter and may affect the interpretation and conclusions drawn from the results.

My analyses highlighted that high event probability was an important factor associated with discordant effect estimates; changes to between and within-study variation were important mechanisms producing differences in the results as well. However, there were occasions where there was no clear single factor driving the differences, since there was a combination of reasons affecting the individual study estimates and corresponding weights.

While most of the meta-analyses within the database were analysed originally as binary, with an outcome classification of all-cause mortality it is worth mentioning that these meta-analyses could include the outcome of short-term mortality (e.g. 30 days) or longer-term mortality (e.g. 5 years); therefore some of these meta-analyses with short follow-up may have been appropriately analysed as binary. The outcome classification of all-cause mortality was considered a representative sample of survival meta-analysis up to 2008, however results might be different for other outcomes and results might have changed in later reviews where more information on methodology was available.

I did not assess other reasons for differences between the results due to lack of information on censoring and follow-up times. Interpretation of the results was conducted with caution as I was interpreting the results based on known factors, without excluding other unknown factors that may have affected the results. I was not able to examine whether current practice of analysing TTE data has changed and whether methodological choices have improved since 2008. Further work examining the differences observed between analyses on the OR and HR scales in the presence of IPD is necessary.

The model used to analyse TTE data as binary is the conventional approach widely used by many systematic reviewers and meta-analysts[133]. It is quick,

inexpensive and study results are obtained from appropriately synthesized study publications or by contacting study authors[136]. This approach to analysis ignores censored observations[137] and treats them as missing and has also been criticised for the within-study normality assumptions required[136].

The use of a clog-log link function, facilitating the results' interpretation in a HR scale for the binary data, was the best alternative approach enabling us to make comparisons between the scales used if only binary summaries are available. In the past, the clog-log link has been proven to provide a close approximation to Cox regression invoking a proportional hazards assumption, rather than a proportional odds assumption[125]. However, for these data, I was not able to assess whether the HR obtained from the clog-log link is a close approximation to the HR estimate that would be obtained under a proportional hazards model; therefore, this magnifies the importance of extracting appropriate information when conducting TTE MA. Similarly, I was not able to identify a clear pattern under which the complementary log-log link could be employed since we were not able to compare the clog-log approach to an approach including for example information on "O-E" and "V" statistics or HR summaries. This will be explored in later chapters.

A limitation observed by T.V Perneger[54] who conducted research on an individual level basis indicated that the use of the clog-log function is useful when the duration of follow-up is the same for all individuals and whenever the traditional two-by-two table is a fair summary of results. However, when duration varies from observation-to-observation Kaplan-Meier curves or incidence rates could be obtained. This could be another justification on the mixed results observed in some of the meta-analyses performed.

For these data, I also used a one-stage random-effects model with fixed study-specific effects describing the baseline risk probability of the event in each study. These models use exact binomial likelihoods and may therefore be more accurate, especially with sparse data[130]. The fixed study-specific effects cause difficulties in estimation since the number of parameters increases with the number of studies, but maximum likelihood theory requires the number of parameters to remain stable as the sample size increases. A random-effects model with random study-specific effects could be applied, however based on

simulation studies this model performed better than others without any serious biases present[130].

To my knowledge, no research has been conducted using such a large database assessing the differences between a) analysing the data as binary and interpreting the results in an OR scale and b) analysing the data either using the clog-log link facilitating interpretation on the HR scale.

I have demonstrated the impact of re-analysing binary TTE meta-analyses within the Cochrane Database on a different scale, identifying the main drivers influencing discrepancies between the meta-analytic results. My findings provided useful insights into changes to meta-analytical results; however, additional research is needed in order to proceed with more extensive comparisons within meta-analyses where data such as “O-E” and “V” statistics or individual participant data are available.

Parts of this Chapter were presented as an oral presentation at the 42nd conference of International Society of Clinical Biostatistics and at the 2021 annual meeting of the Society for Research Synthesis Methodology. The results of the chapter were published in BMC Medical Research methodology, doi:10.1186/s12874-022-01541-9.

---

4. Comparing Methodology of Analysing Time-to-Event Outcomes as Binary in Meta-analysis using Empirical Data from the CDSR

4.1 Chapter Overview

As an extension to the previous chapter, here I compare MA results for a subset of TTE data initially analysed using “O-E” and “V” statistics in the CDSR and interpreted on the HR scale to results from analysing these data as binary on the OR (via the logit link) or HR (via the clog-log link) scale. A detailed comparison using various statistical methods for TTE MA is conducted. At the end of the chapter, I conclude with a discussion of the findings.

4.2 Introduction

As discussed in previous chapters, TTE data is a unique category of data recording “IF” and “WHEN” the event occurred. Various techniques have been developed for analysing TTE data enabling us to utilize multiple time points, account for censoring across study subjects and provide unbiased survival estimates[3].

The Cochrane handbook[138] (version 5.1.0) suggests two approaches for MA of TTE outcomes. Depending on whether data are extracted from the literature or IPD are obtained, either “O-E” and “V” statistics (which are useful alternative statistics if a hazard ratio is not directly reported[47]) can be used or estimates of the $\log \mathrm{HR}$ and its corresponding standard error can be obtained for these analyses. In the presence of “O-E” and “V” statistics, Peto’s method can provide

us with an OR estimate, the log-rank approach[139] with a HR estimate and a variation of Peto's method with something in between[21]. Simpler than the previous statistics, if $\log \mathrm{HR}$ and its standard error is provided, the exponential of the $\log \mathrm{HR}$ will give rise to a HR estimate.

According to Davey et al[140]**,** in January 2008, the Cochrane database contained 22,453 MAs containing 112,600 studies, and the authors performed classifications of MAs by outcome type, medical specialty and types of interventions compared[7]. Systematic reviews and MA of TTE data are published frequently within that database. Among these classified MAs, 1,693 MAs containing 10,959 studies involved TTE outcomes such as “all-cause mortality”, “composite mortality/morbidity only”, and “cause-specific mortality”. Surprisingly, more than 90% of them dichotomised their TTE outcomes and only a small proportion of them accounted for the natural properties of the data.

Having previously analysed TTE data from the CDSR which had been originally analysed as binary, it was important for me to perform comparisons and understand the differences I would obtain in the results if MAs analysed using “O-E” and “V” statistics giving rise to a HR estimate were contrasted to the meta-analytic estimates of OR (via the logit link) or HRs (via the clog-log link) when data were analysed as binary accounting only for the total number of participants and events per arm.

A substantial amount of research was conducted in the past examining the differences between logistic and proportional hazards models in presence of TTE outcomes. In Chapters 1 and 3, I provided detailed comparisons on the differences between the two models using individual studies, while in Chapter 3, I presented “meta-epidemiological” results exploring differences between the models in meta-analyses of outcomes originally analysed as binary. In this chapter, I aimed to extend my previous “meta-epidemiological” study by re-analysing meta-analyses originally analysed using “O-E” and “V” statistics; using additional data from the CDSR, I am able to provide a more accurate conclusion for the analyses performed within the database.

The rest of this chapter is set out as follows. In Section 4.3, I present descriptive statistics of the database and describe the two-stage model I applied. Section 4.4 describes the results obtained from re-analysing the data originally analysed

using “O-E” and “V” data on an OR (via the logit link) and HR (via the clog-log link) scale. In Section 4.5, I perform discussion of the results, present the advantages and disadvantages of the findings and I finish with some conclusions and plans for further work.

4.3 Methods

4.3.1 Data

As described in detail in Chapter 3, the Nordic Cochrane Centre permitted access to the data of the CDSR 2008. Details related to outcomes, outcome classifications, range of sample sizes in the MAs and medical areas are described by Davey et al.[140]. From the database, I was interested in study MAs analysed as HRs with outcome classifications of overall survival and progression/disease free survival. Based on the database using the outcome classification I was able to identify (using words such as “survival”, “death”, “fatality”) another two sets of TTE meta-analyses:

- “OEV” meta-analyses: Those with outcome classifications “overall survival” and “progression/disease free survival” where the information recorded was based on “binary” data (as in Chapter 3) in addition to log-rank “O-E” and “V” statistics”; these were originally analysed as HRs in the RevMan software;
- Meta-analyses with estimated log HR and its standard error. These were removed from further analyses since there was no available information on the number of events and participants per arm and therefore no binary data meta-analysis could be conducted.

Therefore, I identified another subset of TTE MAs: those with binary summaries in addition to “OEV” data. I analysed the outcome types of overall survival and progression/disease free survival separately to assess whether differences exist due to different characteristics of the outcomes. I also examined whether the information available as “OEV” data was based on aggregate data or IPD by examining the individual Cochrane reviews.

4.3.2 Eligibility Criteria

I initially extracted the "OEV" data and conducted cleaning including examination of the outcome classification; Rebecca M. Turner confirmed the choice of included meta-analyses obtained from "OEV" data extraction. I identified 16 misclassifications due to disagreement with the original outcome classification as listed in the datasets, conflicting information in the database or unavailability of the correct version of the Cochrane review. Another 158 studies were excluded due to the fact that even though an estimate of the log HR and its variance were available, the number of events and total number of patients in the studies were not available. Finally, 32 studies were excluded because they had double zero events, therefore not contributing to any analyses, and another 7 MAs were excluded because they contained fewer than three studies, for which we know that a synthesis is usually questionable[130]. Figure 4.1 shows the sample derivation.

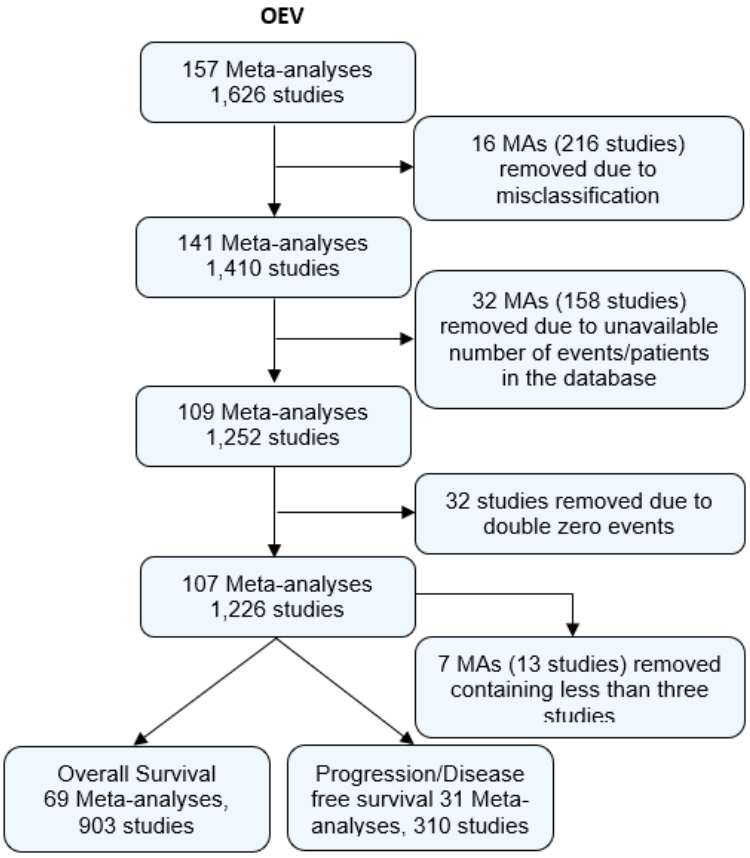


Figure 4.1: Analysis sample of "OEV" dataset from the CDSR (2008, issue 1).

4.3.3 Descriptive Statistics

For both outcome classifications, as in the binary data analysed in Chapter 3, I obtained descriptive statistics such the number of studies per meta-analysis, number of events and study size by the median and interquartile range. I also identify the number of medical specialities, and median number of events (and IQR) per medical specialty.

4.3.4 Model description for "OEV" data

For these data, the "O-E" and "V" statistics were available in the Cochrane database alongside the number of patients and events. They came either from published reports or from IPD; I examined the individual reviews from the Cochrane database and assessed the data origin. Since there was more available information for these data than for the binary data (Chapter 3), the following three models were applied, using only two-stage meta-analysis models.

I initially analysed the "OEV" data as binary data and modelled them as described in detail in Chapter 3. I also used the log-rank Observed - Expected events (O-E) and the log-rank Variance (V) statistics calculated previously from the number of events and the individual times to event on each research arm of the trial; I used the log-rank approach[139] in order to obtain another type of HR estimate. I used random-effects models to analyse the data throughout, including between-study heterogeneity to account for variation across studies.

4.3.5 Fitting two-stage random-effects models for "OEV" data

Similarly to Chapter 3, the estimated log odds and log hazard ratios were given by equations (3.1) and (3.2) for the binary summaries while the "O-E" and "V" statistics were used as follows:

$$y_i = \frac{\text{logrank Observed} - \text{Expected events (O} - \text{E)}}{\text{logrank Variance (V)}} \text{ for HRs } (4.1)$$

The corresponding variances were given by equations (3.3) and (3.4) as shown in Chapter 3 for binary summaries while for "O-E" and "V" statistics as follows:

$$s_i^2 = \frac{1}{\text{logrank Variance (V)}} \text{ for HRs } (4.2)$$

where V denotes the variance of the logrank statistic.

To avoid biases such as downward bias on the unknown variance components estimates, I used the REML estimator as the most suitable approach for the model implementation[132]. Between the treatment and control arms, a hypothesis test is performed assuming identical proportions of patients for whom the event occurs versus the alternative hypothesis of higher or lower proportion of patients experiencing the event. Test accuracy is based on the following assumptions: a) drop out times are identically distributed across treatment and control groups and b) drop out times are independent of the occurring event time[137]. The model was implemented via the "rma.uni" command from "metafor" package in R as shown in Appendix C.2.

4.4.6 Model comparison for "OEV" data

For the "OEV" data set, comparisons on overall and progression/disease free survival outcomes were conducted separately; this was because differences between these outcomes might be observed in the presence of different disease severities, and therefore this would be associated with different length of follow-up and risk of the outcome.

For both outcomes, I performed comparisons by examining the differences between analysing the data as binary on an OR scale, analysing the data as binary using the clog-log link on a HR scale, or analysing the data using the "O-E" and "V" statistics on a HR scale. I assessed whether the differences observed from analysing the data as binary on an OR scale could be reduced by the use of the clog-log link. I present only comparisons of the results under two-stage models since there were no available IPD to perform comparisons under one-stage models.

Similarly to Chapter 3, I examined the proportion of significant and non-significant meta-analytic pooled effect estimates under the different scales used and identified the number of meta-analyses which were significant under one scale and non-significant under the other. I created Bland-Altman plots for the standardised treatment effect estimates and for the I^2 statistics to explore the agreement among the methods producing fair comparisons between the two scales[135]. Meta-analyses outside the 95% limits of agreement were examined for their characteristics.

4.4 Results

In the Cochrane database, 157 meta-analyses were originally analysed using the "O-E" and "V" statistics on a HR scale. After applying the exclusion criteria, 100 MAs remained for further analysis. The median number of meta-analyses per review was 2 with IQR (2, 3). The median number of studies and the median number of events are provided in Table 4.1. The overall survival category contains 38 (55%) IPD and 31 (45%) non-IPD MAs whereas progression/disease free survival contains 17 (55%) IPD and 14 (45%) non-IPD MAs. The median number of events for IPD MAs was 122 with an IQR (57, 278) while non-IPD MAs had a median of 93 with an IQR of (41, 202).

After applying the same exclusion criteria to the excluded MAs from this chapter, which presented estimated log HR and its standard error only (32 MAs including 158 studies, see Figure 4.1), overall survival was represented in the sample with 13 MAs and 84 studies whereas progression/disease free survival was represented with 8 MAs and 40 studies. The median number of studies in the former outcome was 7 IQR (6,12) and for the latter outcome 6 IQR (4,8), slightly less but still in the same range as the included MAs providing the "O-E" and "V" statistics. The median number of events and median study size could not be obtained and therefore we could not assess further how similar the excluded MAs are to those included in the sample analysed.

	"OEV"	
Outcome	**Overall Survival**	**Progression/Disease Free Survival**
Total Number of MA	69	31
Number of studies per MA: Median (IQR)	10 (6, 14)	10 (7, 14)
Number of events per MA: Median (IQR)	108 (58, 254)	104 (70, 192)
Median Study Size (IQR)	182 (93, 369)	185 (90, 317)

Table 4.1: Descriptive statistics for "OEV" data from the CDSR.

The distribution of medical specialities of the meta-analyses is presented in Table 4.2. I observed that analysing TTE outcomes as HRs is restricted to very few medical specialties; “Cancer” was still the most frequent medical specialty for both outcome types as observed in Chapter 3.

	“OEV”			
Medical Specialty	**OS[++]: Number (%) of MAs**	**Events per MA: Median (IQR)**	**PDFS[++]: Number (%) of MAs**	**Events per MA: Median (IQR)**
Cancer	60 (87%)	104 (45, 221)	31(100%)	116 (56,243)
Digestive/endocrine, nutritional and metabolic	1 (1%)	52 (35, 64)	-	-
Infectious diseases	8 (12%)	482 (160,1109)	-	-

[++]OS: Overall Survival, PDFS: Progression/Disease free survival

Table 4.2: Distribution of medical specialties for the “OEV” data meta-analyses in the CDSR.

Table 4.3 provides the percentages of significant and non-significant meta-analyses for each outcome for two-stage models at a two-sided 5% level, indicating that discrepancies are more prevalent in the “OEV” data compared to the “binary” data (in Chapter 3); additionally the amount of discrepancies observed in statistical significance from the comparison of OR and HR obtained from the clog-log link was smaller than the amount of discrepancies observed between the OR and HR analyses. With regards to dichotomisation of MAs into IPD and non-IPD, IPD MAs had 9 (16%) MAs having a non-significant OR and significant HR and another 1 (2%) MA with a significant OR and non-significant HR; the corresponding numbers for non-IPD MAs were 6 (13%) and 3 (7%) respectively.

			OR		**HR (O-E & V)**	
“OEV”			Sig*	Non-Sig*	Sig*	Non-Sig*
HR (clog-log)	Overall Survival	Sig*	20 (29%)	1 (0.2%)	18 (26%)	10 (14%)
		Non-Sig*	1 (0.2%)	47 (68%)	3 (4%)	38 (55%)
	Progression/Disease free Survival	Sig*	9 (29%)	0 (0%)	8 (26%)	6 (19%)
		Non-Sig*	1 (3%)	21 (68%)	1 (3%)	16 (52%)

HR (O-E &V)	Overall Survival	Sig*	18 (26%)	10 (14%)
		Non-Sig*	3 (4%)	38 (55%)
	Progression/Disease free Survival	Sig*	9 (29%)	5 (16%)
		Non-Sig*	1 (3%)	16 (52%)

*Sig/Non-Sig: Significant; Non-Significant

Table 4.3: Number (%) of (non-)significant meta-analyses under different scales for two-stage models ("OEV" data).

Bland-Altman plots produced for this subset indicated that the average difference between each pair of methods is larger than those obtained from the "binary" data (Figures 4.2 – 4.4). For overall survival, the average difference between the two methods for the standardised pooled effect estimates was 0.2 units (-1.8 units, 2.1 units) for OR versus HR and 0.2 units (-2.2 units, 2.5 units) for HR using clog-log versus HR; however, for OR vs HR clog-log differences the average bias was 0 units (-2.6 units, 2.7 units) indicating that clog-log is a closer approximation to OR rather than HR analyses (Figure 4.2). For the estimation of I^2, the average difference between the methods is -6% (-41%, 29%) for OR versus HR, -6% (-42%, 31%) for HR using clog-log versus HR, and 0% (-21%, 21%) for OR vs HR clog-log differences; similarly the clog-log seems a closer approximation to OR analyses rather than HR analyses (Figure 4.3). For progression/disease free survival, the average difference between the two methods for the standardised pooled effect estimates was 0.4 units (-1.5 units, 2.2 units) for OR versus HR, 0 units (-2.9 units, 2.9 units) for HR using clog-log versus HR, and 0.4 units (-2.7 units, 3.5 units) for OR vs HR clog-log differences (Figure 4.4). For the estimation of I^2, the average difference between the methods is -16% (-86%, 53%) for OR versus HR, -16% (-89%, 57%) for HR using clog-log versus HR, and 0% (-13%, 13%) for OR vs HR (Figure 4.5). Bland-Altman plots produced for the average difference between each pair of methods for IPD and non IPD datasets are presented in Appendix C.1.

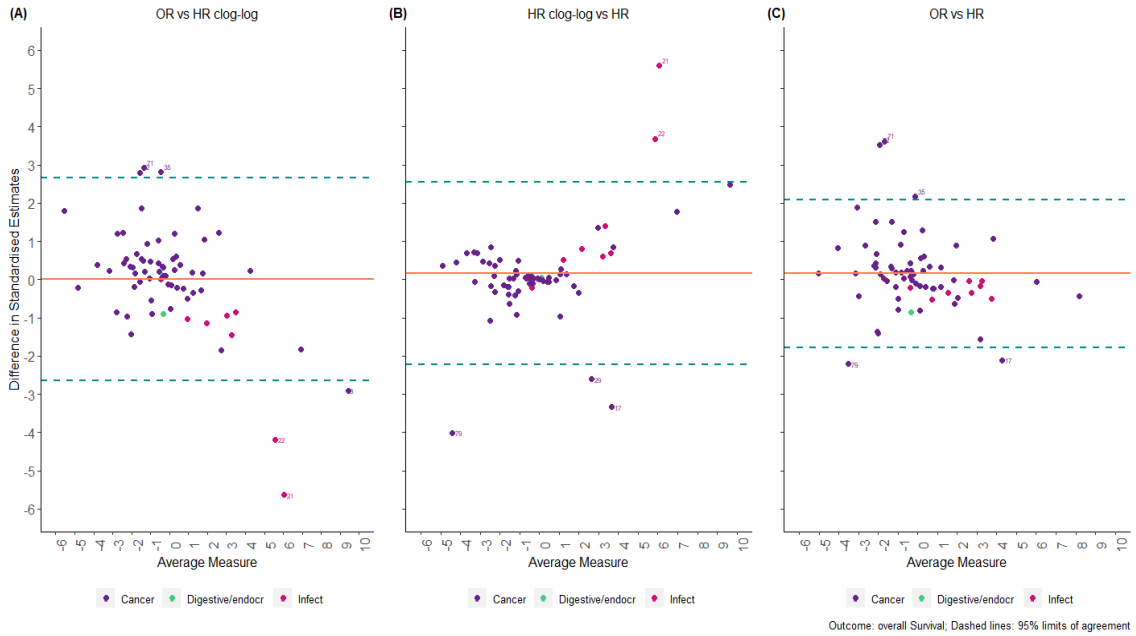


Figure 4.2: Overall Survival - Bland-Altman Plot comparing standardised OR vs. HR estimates for two-stage models in “OEV” data.

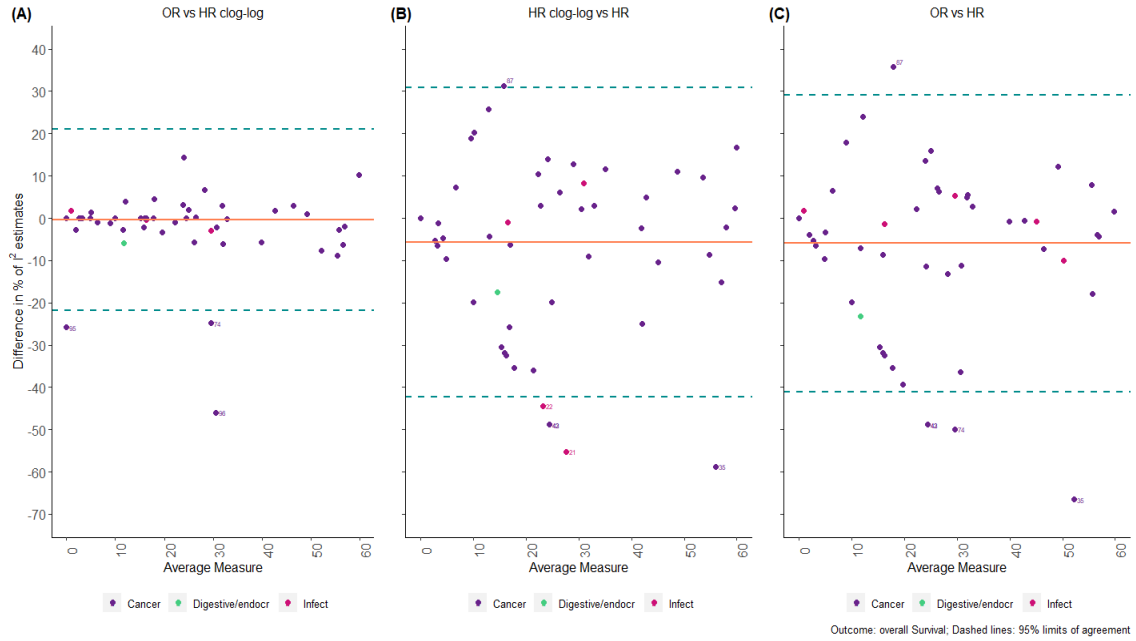


Figure 4.3: Overall Survival - Bland-Altman Plot comparing I^2 estimates (OR vs. HR) for two-stage models in “OEV” data.

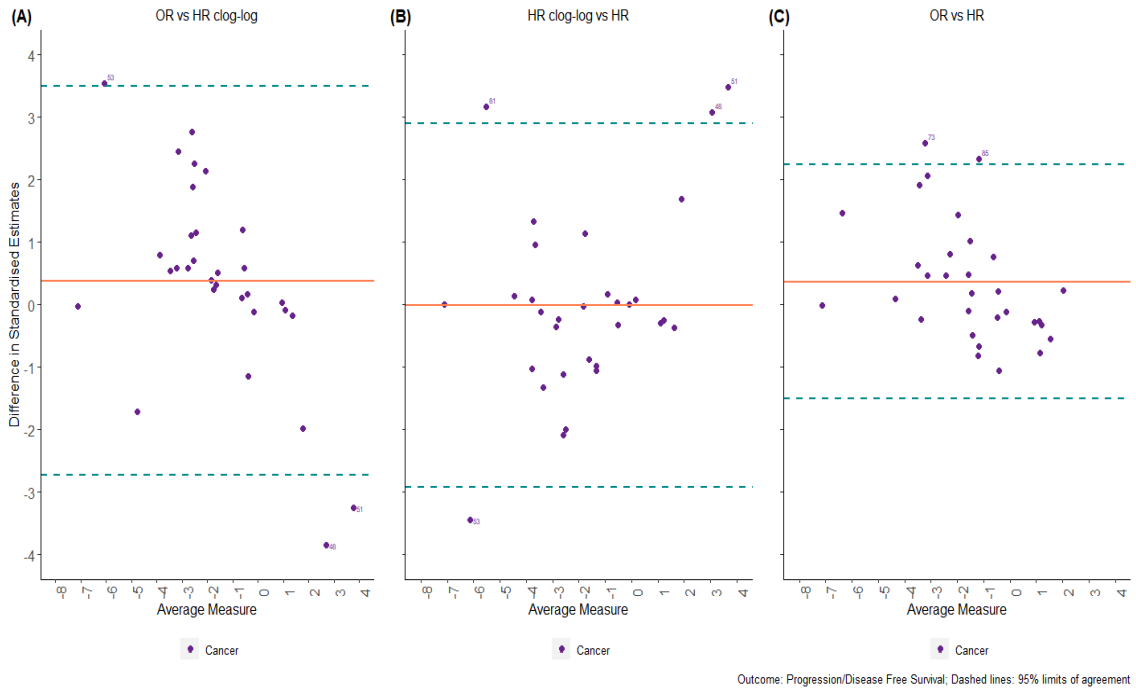


Figure 4.4: Progression/Disease Free Survival - Bland-Altman Plot comparing standardised OR vs. HR estimates for two-stage models in "OEV" data

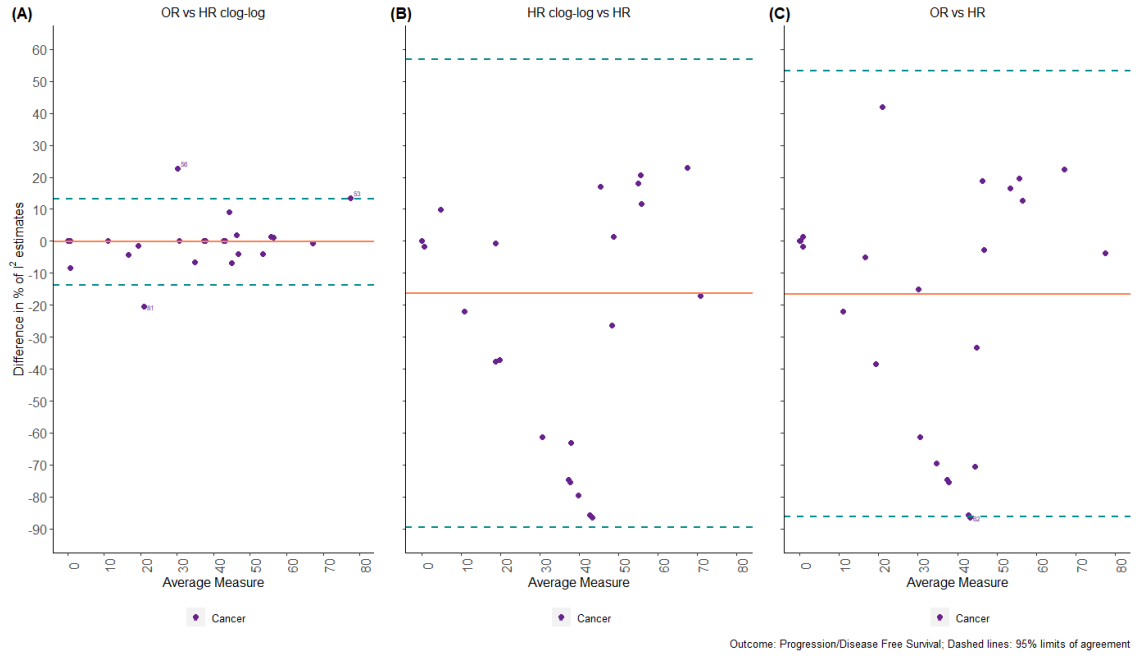


Figure 4.5: Progression/Disease Free Survival - Bland-Altman Plot comparing I^2 estimates (OR vs. HR) for two-stage models in "OEV" data.

Outliers were considered 28% of the "OEV" meta-analyses. Of these, 57% were from IPD rather than non-IPD and 54% of them were for the outcome of overall

survival. In 50% of the outliers a high event probability (defined here as probability greater than 0.7) was observed, suggesting that this may be an important factor associated with differences among the scales used. For example, meta-analysis 45 (outlier obtained from standardised estimates) consists of 7 studies for which the event probability was greater than 0.7 for all the studies; consequently, high event probability affected substantially the differences in the individual study estimates between the OR and HR analyses, leading to different allocated relative weights for the studies, and discrepancies in the pooled effect estimates as shown in Figure 4.6. Even though the individual HR clog-log estimates were closer to the individual OR estimates, the final pooled effect estimate was closer to the pooled HR estimate; this was not though the case for all meta-analyses.

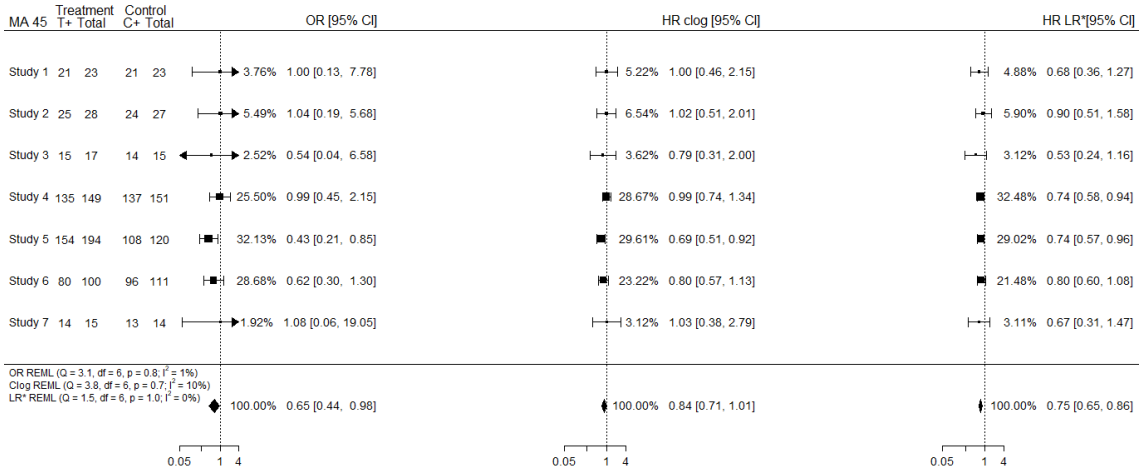


Figure 4.6: Forest plot (MA 45) indicating discrepancies in the presence of high event probability.

Increased within-study variability on the OR scale relative to the HR scale may affect the weighting more than the actual estimates in the studies, for example for meta-analysis 17 (outlier obtained from standardised estimates), producing differences in the pooled effect estimates between the two scales (Figure 4.7).

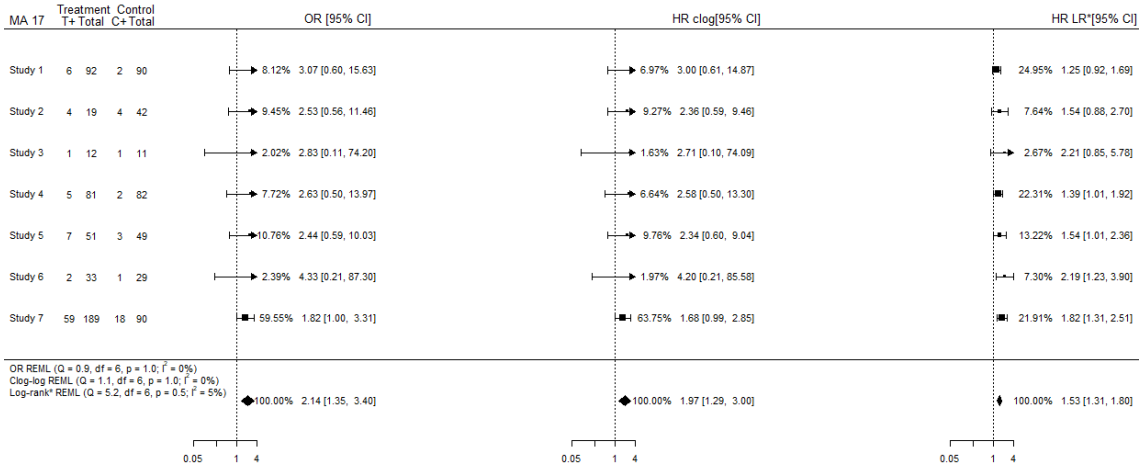


Figure 4.7: Forest plot (MA 17) indicating increased within-study variability on the OR scale relative to the HR scale.

Similarly, even though the individual study estimates and weights of OR and HR clog-log were closer to each other, the HR clog-log pooled effect estimate was closer to the pooled HR estimate; however, this was not the case for all meta-analyses (e.g., MA 83, 85). Important differences in between-study heterogeneity between the HR and OR analyses were observed in meta-analyses such as 42, 90. For example, meta-analysis 90 (outlier obtained from I^2 estimates) consists of 11 studies out of which 8 are smaller studies and 3 are larger studies. Smaller studies received increased weight in the HR analysis compared to the OR analysis, while larger studies received smaller weights in the HR scale compared to OR scale. However, this was not the case on the HR clog-log scale as presented in Figure 4.8.

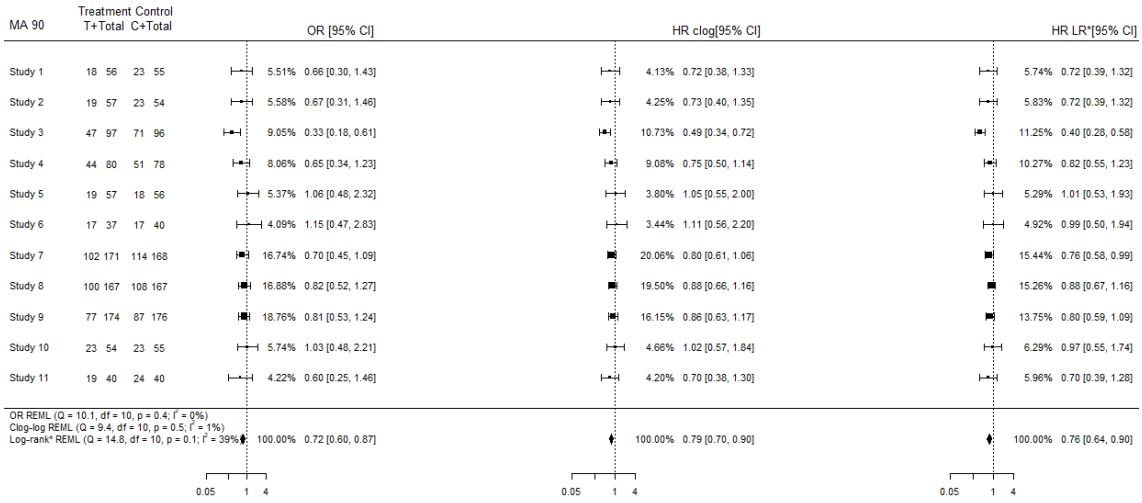


Figure 4.8 Forest plot (MA 90) indicating discrepancies arising from differences in between-study heterogeneity.

In 46% of the outlying meta-analyses, the individual study estimates, and the corresponding weights were affected by a combination of differing event probability across study arms, differences in between-study heterogeneity or increased within-study variability on the OR relative to the HR scale (e.g., MA 3, 35, 56, 68, 71-74, 79, 87). In the presence of a limited amount of studies in the meta-analyses this was even more evident. Additional examples of forest plots and the exact tables indicating the discrepancies among the results are shown in Appendix C.3.

Overall, using the “OEV” data, a mixed pattern was observed. In 39% (n=11) of outlying meta-analyses the OR pooled effect estimate was closer to HR pooled effect estimate; however, in 4 out of 11 outlying meta-analyses the individual study estimates obtained from the HR clog-log link were a closer approximation to the individual study HR estimates. Similarly, even though in 61% (n=17) of the outlying meta-analyses the HR clog-log pooled effect estimate was closer to the pooled HR estimate, 3 of outlying meta-analyses provided individual study OR estimates closer to individual study HR estimates, and another 3 individual study HR clog-log estimates were closer to individual study OR estimates. Finally, observing Figures 4.13 - 4.14 from Appendix C.1, I was not able to identify whether the differences between the OR and HR scales are associated with the level of baseline risk in individual studies.

4.5 Discussion

Using an additional subset of meta-analysis data from the CDSR of 2008 (“OEV” data), I compared different methods for handling TTE outcomes within meta-analysis. I identified the differences that occur when these data are analysed as binary as opposed to analysing the data using the clog-log link additionally to using the “O-E” and “V” statistics where interpretation is conducted on a HR scale. As in Chapter 3, the analysis confirmed that there are important reasons associated with discordance among the results, indicating that the correct choice of the method does matter affecting also the interpretation and conclusions drawn from the results.

High event probability, changes to between and within-study variation were confirmed to be important factors producing differences in the results in this subset of meta-analyses as well. However, in this dataset there were more occasions under which there was no clear indication of one single factor driving these differences and a combination of reasons affected the individual study estimates and corresponding weights. Therefore, regarding method selection, based on the “OEV” data I identified that a mixed pattern was observed and there was no clear indication of the exact conditions under which the clog-log link outperforms logit link on an OR scale or vice versa.

The data used for the comparison of OR/HR scale in the “OEV” data were slightly different; I used the number of events and non-events for the OR and HR clog-log calculation (as in Chapter 3 using binary data) and calculated a HR based on “O-E” and “V” statistics. Therefore, there is a possibility that for some cases the two data sets entered by Cochrane reviewers may not completely correspond to each other. Lack of information on censoring pattern and follow-up times was present in this subset as well and therefore interpretation was conducted carefully since I interpreted the results based on known factors and cannot exclude other unknown factors affecting the results.

Two out of the three models applied were the same as those applied in Chapter 3 (i.e. the conventional two-stage approach and the use of the clog-log link function). The third approach used for the “OEV” data was the log-rank approach; “O-E and V” data provide the best method to analyse aggregate data and facilitate interpretation of results on the HR scale, but in the absence of IPD important biases may occur when large treatment effects and unbalanced data are

present[20]. I was not able to identify a clear pattern under which the clog-log link could be employed since there were circumstances under which it performed better or worse than an OR analysis; therefore, I was not able to identify whether the clog-log approach is useful when a MA includes binary summaries alongside "OEV" or HR summaries. IPD and simulation studies are required to assess in more detail the conditions determining where this method would be acceptable.

Finally, I was not able to make comparisons using one-stage models in the "OEV" data. I would be able to apply one-stage models when the data were analysed as binary, but I did not have the IPD required to fit one-stage models on the HR scale.

To my knowledge, no previous research has been conducted using a large database subset such as this to assess the differences between a) analysing the data as binary and interpreting the results in an OR scale and b) analysing the data either using the clog-log link or log-rank "O-E" and V statistics facilitating interpretation on the HR scale.

In conclusion, summarising the findings from Chapter 3 and 4, the results obtained indicate that TTE data should be ideally analysed accounting for their natural properties, as it is possible for important discrepancies to be observed and conclusions from the meta-analysis to be altered. I identified that dichotomising TTE outcomes may be adequate for low event probabilities but not for high event probabilities. In meta-analyses where only binary data are available, the clog-log link may be a useful alternative when analysing TTE outcomes as binary, however the exact conditions need further exploration. These findings provide guidance on the appropriate methodology that should be used when conducting such meta-analyses.

Parts of this Chapter were presented as an oral presentation at the 42nd conference of International Society of Clinical Biostatistics and at the 2021 annual meeting of the Society for Research Synthesis Methodology. The results of the chapter were published in BMC Medical Research methodology, doi:10.1186/s12874-022-01541-9.

5. Analysing Time-to-Event Outcomes as Binary in Meta-analysis using Individual Participant Data

5.1 Chapter Overview

Using IPD with TTE outcomes, providing sufficient information on censoring and follow-up time, this chapter makes a comparison between "gold-standard" approaches such as Cox proportional hazards model and log-rank test and less appropriate methods analysing the data as binary on the HR (via the clog-log link) or the OR (via the logit link) scale. In this chapter, I am learning about the magnitudes of discordances in practice while assessing also whether censoring and follow-up time are additional factors affecting any discordances among the results; this is something that I was not able to assess in previous chapters using the CDSR. I am aiming also to confirm previous evidence that method choice does matter and to inform a subsequent simulation study which will provide more definitive evidence on the most appropriate method for handling this data type.

5.2 Introduction

In previous chapters I mentioned that an individual participant data (IPD) meta-analysis directly obtains evidence from researchers responsible for individual studies, aiming to re-analyse them simultaneously using appropriate methodology[141]. This approach can be more costly than conventional systematic reviews extracting aggregate data and is regularly characterised as a "gold-standard" since the time element between randomisation and the event is of interest and allows us to re-analyse individual trial data. Cardiovascular diseases and cancer are medical areas in which an IPD meta-analysis is considerably

important since most interventions used in these areas aim for prolongation of survival[141].

Even though IPD meta-analysis is the ideal approach for TTE outcomes , it can be time consuming to collect and analyse the data[16]. Therefore, systematic reviewers and meta-analysts may instead decide to obtain:

- a HR alongside its confidence interval via the log-rank test using the “O-E” and “V” statistics by extracting aggregate data from trial reports, or
- a HR alongside its confidence interval from trial reports and include them directly in a two-stage meta-analysis model, or
- the number of events and participants per arm and analyse the data as binary giving rise to an OR or a RR

Other approaches, less familiar to systematic reviewers and meta-analysts, can be adopted if only aggregate data are available, such as a normal approximation to binomial likelihood with a clog-log link. However, analysing the CDSR (Chapters 3, 4) did not reveal any circumstances under which undesirable properties of analysing TTE outcomes as binary on an OR scale can be mitigated by using a clog-log link. In the presence of IPD, a Cox proportional hazards model can also be applied with interpretation on a HR scale. One-stage meta-analysis models have also been developed allowing for more accurate inferences on the results[142].

Using IPD of TTE outcomes, I compare the results from the “gold-standard” approaches (Cox PH and log-rank test) on a HR scale to less appropriate methods treating data as binary on a HR (via the clog-log link) and OR (via the logit link) scale. In such a way, I can assess the roles of length of follow-up and censoring in a MA of TTE outcomes that I could not investigate via exploratory work performed using the CDSR[143]. I try to confirm previous evidence indicating that differences between scales arise mainly when event probability is high and may occur via differences in between-study heterogeneity or via increased within-study standard error in the OR relative to the HR analyses. The analyses performed use both two- and one-stage models.

The rest of this chapter is set out as follows: In section 5.3, I describe the IPD and the statistical models I used. In sections 5.4, I present descriptive statistics and in sections 5.5-5.6, I present the results obtained from the models performed.

These results are followed by a discussion exploring the strengths and limitations of my findings, together with conclusions and further work (sections 5.7, 5.8).

5.3 Methods

5.3.1 Data

The Meta-analysis group from the MRC CTU provided data from the IPD on "Neoadjuvant chemotherapy in invasive bladder cancer: a systematic review and meta-analysis"[144]. The IPD MA consists of 11 trials. For 2 trials, investigators did not allow for data analyses related to methodological purposes; therefore data from 9 trials were accessed (Martinez-Piniero[145], Raghavan[146], Malmstrom[147], Wallace[146], Cortesi (unpublished), MRC/EORTC[148], Sherif[147], Sengelov[149], Grossman[150]). A priori, I was interested in exploring different TTE outcomes within IPD since they provide diversity in the lengths of follow-up time and percentage censoring per individual trial. All trials were examining the use of platinum-based combination chemotherapy prior to local treatment in comparison to local therapy only.

5.3.2 Descriptive Statistics

I obtained descriptive statistics such as the number of patients allocated per arm, sex, age group and T category (i.e. size and extent of the main tumour) per trial. I also calculated the median TTE and 95% confidence interval for the outcomes of interest, and median follow-up time and IQR per individual trial (based on the Kaplan-Meier method applied to the censored times reversing the roles of event status and censored). Kaplan-Meier plots were also produced per outcome per trial to examine whether the proportional hazards assumption holds. Finally, I tried to obtain descriptive statistics on the percentage of random and fixed censoring. Random censoring is referring to those observations who were lost to follow-up before the end of the study whereas fixed censoring is referring to those observations that were censored at the end of the follow-up time (i.e., no information obtained on whether participants experienced the event or not).

5.3.3 Methods for Individual Participant Data Meta-Analysis

Initially, a log-rank test and a Cox proportional hazards model were applied to each trial, as described in 5.3.3.1, in order to obtain "O-E" and "V" statistics and

a HR and its standard error; these were entered in a two-stage meta-analysis model as described in 5.3.3.2.

5.3.3.1 Testing Survival Curve Differences & Cox Proportional Hazards Model for Individual Trial Data

I calculated the HR with its associated standard error and "O-E" and "V" statistics per outcome per trial assessing whether the proportionality assumption holds; this would allow me to apply suitable MA models. Calculations were implemented via the "coxph" and "survdiff" command from "survival" package in R. Below I provide the formulas used to calculate these statistics.

A Cox proportional hazards regression model stratified by trial was given by the following equation

$$\lambda_{ij}(t) = \lambda_{i0}(t)exp\ (\beta_i x_{ij}) \quad (5.1)$$

for the jth patient in the ith study with treatment indicator variable $\mathrm{x_{ij}}$, λ_{i0} the baseline hazard function in the ith study and β_i the linear predictor. A Cox proportional hazards model does not make any assumptions on the baseline hazard function but verification of the proportional hazards assumption is necessary (i.e. the effect should be independent of time). No patient-level characteristics were entered into the linear predictor of the model.

Survival curve differences stratified by trial were calculated as follows: For each individual failure time t a $2x2$ table was constructed and the number of participants at risk in each treatment arm (n_{kt}) was recorded alongside the associated number of deaths in each group (d_{kt}) where $k = 0, 1$ denote the treatment group ($k = 0$ indicates the control and $k = 1$ indicates the treatment group). Additionally, $\mathrm{n_t}$ is defined as $\mathrm{n_t = n_{0t} + n_{1t}}$ and $d_t = d_{0t} + d_{1t}$. Under the assumption of no association between the groups and the event, the expected number of deaths in group 1 is

$$e_{1t} = n_{1t} * \frac{d_t}{n_t} \quad (5.2)$$

with variance

$$\mathrm{v_{0t}} = \frac{\mathrm{n_{0t} * n_{1t} * d_t(n_t - d_t)}}{\mathrm{n_t^2 * (n_t - 1)}} \quad (5.3)$$

The log-rank test compares the total number of deaths in one of the treatment groups

$$O_1 = \sum_t d_{1t} \text{ and variance } V_1 = \sum_t v_{1t}$$

with the expected number of deaths in that group under the null hypothesis

$$E_1 = \sum_t e_{1t}$$

Under the null hypothesis of no difference between the groups I can obtain a X^2 test as follows:

$$\frac{(O_1 - E_1)^2}{V_1} \sim X_1^2 \quad (5.4)$$

The test provides a p-value, does not give an estimate of the size of the difference, and does not allow for inclusion of additional covariates in the analysis. It is worth mentioning at this point that often systematic reviewers and meta-analysts do not use the appropriate referencing (i.e. Yusuf et al.[19]) for the use of the log-rank test and they are probably influenced by incorrect citations in previous research publications. The appropriate referencing for the use of the log-rank test is given by Harrington et al.[151] and this is the reference used in the main R documentation.

5.3.3.2 Model Description

Two-stage IPD MA models

First, a Cox proportional hazards model was applied and a HR alongside its standard error was obtained for each outcome in each trial accounting for censoring and the time element. The HR and standard error data were entered in a two-stage meta-analysis model. Second, information on the “O-E” and “V” statistics were obtained when I performed testing of the survival curve differences. The “O-E” and “V” statistics were entered in a two-stage MA model. In the third approach, I modelled the “binary” data obtained from IPD using a normal approximation to binomial likelihood with a clog-log link on a HR scale ignoring censoring and follow-up times. Fourth, I applied a model for the same data, assuming a binomial likelihood and a logit link[133], on an OR scale ignoring the same information as in the aforementioned model.

One-stage IPD MA models

I also applied one-stage IPD MA models as these models use the exact binomial likelihood and may therefore be more accurate, especially with sparse data[130]. The three models described below were applied.

Initially, a one-stage random-effects Cox proportional hazards model was applied with a lognormal frailty term for the intervention effect estimated via penalized partial likelihood. This accounts for censoring, follow-up time, and the within- and between-trial intervention effects are estimated simultaneously aiming to provide a more complete understanding of the data[79, 111]. The median hazard ratio (MHR) is used to evaluate the meaning of the frailty and is defined as "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"[79]. The MHR is referred to as HR for the rest of this chapter to avoid potential confusion between MHR and HR in the results, as they will be treated in the same way for the comparison to the OR scale.

Second, a generalised linear mixed model using a normal approximation to binomial likelihood with a clog-log link was used based on aggregate data; interpretation was on a HR scale. Finally, a generalised linear mixed model using a binomial likelihood with a logit link was used based on binary summaries; interpretation was conducted on an OR scale. More details describing the latter two models were presented in Chapter 3.

5.3.3.3 Fitting Random-Effects Models

The estimated log HRs and log ORs for some of these models apart from the Cox proportional hazards model were presented in detail in Chapters 3 and 4, so are not repeated here. I indicate the equations to refer to below.

Fitting two-stage random-effects models

The estimated log HRs and log ORs for individual studies were given by:

$$y_i = \begin{cases} \log \mathrm{HR} \quad \text{for HRs obtained from Equation (5.1)} \\ \text{Equation (4.1) for HRs using "O} - \text{E" and "V" statistics (Chapter 4)} \\ \text{Equation (3.2) for HRs using the clog} - \text{log link (Chapter 3)} \\ \text{Equation (3.1) for ORs using the logit link (Chapter 3)} \end{cases}$$

The corresponding sampling variances for these estimates were given by:

$$s_i^2 = \begin{cases} \text{Square of standard error obtained from Equation (5.1)} \\ \text{Equation (4.2) for HRs using "O} - \text{E" and "V" statistics (Chapter 4)} \\ \text{Equation (3.4) for HRs using the clog} - \text{log link (Chapter 3)} \\ \text{Equation (3.3) for ORs using the logit link (Chapter 3)} \end{cases}$$

For the Cox proportional hazards model the standard error of the log HR was obtained instead of the variance. Using these estimates and sampling variances I fitted two-stage random-effects models incorporating between-study heterogeneity variance. I also obtained the I^2 statistic from the fitted models as shown in Equation (3.5) from Chapter 3. The models were implemented via the "rma.uni" command from "metafor" package in R (Appendix D.2).

Fitting one-stage random-effects models

A one-stage Cox proportional hazards model was initially fitted by modelling the distribution of the baseline hazard via frailty terms (i.e. random intercept) and accounting for clustering as follows:

$$\lambda_{ij}(t) = \lambda_0(t)\eta_{ij}exp(\beta_i x_{ij}) \quad (5.5)$$

for the jth patient in the ith study with treatment indicator variable x_{ij} where $log(\eta_{ij}) \sim Normal(0, \tau^2)$ represents the log-frailty. The frailty term follows a specific distribution allowing for differences in baseline rate between participants in the groups. This model accounts for these differences between studies on unmeasured covariates, assuming that the baseline hazard within each trial has different magnitude but the same shape[79].

Additionally, two generalised linear mixed models were applied in one stage using binary summaries: one allowing interpretation on a HR scale and the second on an OR scale. These models have been described in detail in Section 3.3.3.2 in Chapter 3. The models applied used: a) the "coxme" command from "coxme" package for one-stage HRs obtained from Cox model, b) the "rma.glmm" command from "metafor" package to calculate the one-stage ORs and c) the "glmer" command from "lme4" package was used for the HR estimates obtained from the clog-log link. Estimation of between-study heterogeneity (I^2) for the one-stage HR models using binary data was performed as described in Section 3.3.3.2 and Appendix B.4. Estimation of between-study heterogeneity (I^2) for the one-stage Cox proportional hazards model was conducted using the detailed description provided by the paper by De Jong et al[79]. Model implementation was conducted in R (version 4.1.1) and is presented in the Appendix D.2.

5.3.3.4 Model Comparison

For each outcome I prepared forest plots and identified any discrepancies observed from the application of the two-stage models. For one-stage models I tried to confirm whether the discrepancies observed from two-stage models are still observed using this methodology which is considered more accurate especially with sparse data[136].

5.4 Descriptive Statistics & Preliminary Calculations

The IPD include 9 trials and 7 clinical outcomes; I focused on 4 outcomes with different lengths of follow-up time and percentage censoring, because these were considered potential important factors that might impact our results. These 4 outcomes were also used in the main IPD meta-analysis publication published in 2003[144]. The median TTE per outcome per trial (and IQR) and follow-up time per trial (and IQR) is presented in Table 5.1. All trials have long follow-up times and the median TTE ranges across trials per outcome from short (e.g., Australia) to long (e.g., SWOG).

Additional descriptive characteristics per individual trial were obtained such as the number of patients included in each arm of the trial together with summaries of sex, age group, and cancer stage (T category) (Table 5.1). As shown in Table 5.1, UK[146], GUONE and SWOG[150] trials did not provide information on local recurrence free survival and metastasis-free survival; additionally to this, UK trial did not provide information on event free survival. The largest trial was the BA06[148] followed by Nordic2[147] and SWOG[150]; the majority of the patients included in this dataset were males, aged over 60 years and were at the third stage of cancer. For GUONE trial, I could not distinguish between stages T3 and T4 of tumours and they were merged.

For percentage random and fixed censoring, I did not have data to distinguish between them (e.g. dates were not provided to me to avoid identifiability issues) and therefore I had to look at the original trial publications. Across trials I was able to identify minimum follow-up time only for 4 out of 9 trials (i.e. Spain, Australia, UK and DAVECA); among those 1 trial (i.e. Spain) had provided additional data and longer follow-up to the IPD meta-analysis, after the original trial publication. Therefore, I would wrongly calculate random censoring if I based it on the initial

follow-up time reported in the trial publication. Nordic 1 and Nordic 2 trials had no patients lost to follow-up and therefore they provided only fixed censored observations. One trial was unpublished (i.e. GUONE) and I could not identify minimum follow-up time, and two trials (BA06 and SWOG) provided median follow-up time and interquartile range per treatment group without any indication on the minimum follow-up time. For the rest of this chapter censoring is discussed as total censoring (i.e. including fixed and random censoring).

	Spain Martinez-Piniero[145]	Australia Raghavan [146]	Nordic1 Malmstrom[147]	UK Wallace[146]	GUONE Cortesi	BA06 MRC/EORTC[148]	Nordic2 Sherif[147]	DAVECA Sengelov[149]	SWOG Grossman[150]
Median TTE (95% CIs)									
Event free survival	3.2 (1.2,4.4)	0.8 (0.7,1.2)	5.9 (3.9, 8.1)	-	1.6 (1.10,2.5)	1.5 (1.3,1.8)	3.2 (2.3,4.9)	0.9 (0.8, 1.1)	2.5 (1.9,4.9)
Local recurrence free survival	3.3 (1.3,4.6)	0.9 (0.8,1.3)	6.8 (4.5, 9.1)	-	-	1.8 (1.6,2.2)	3.7 (2.6,5.1)	1.0 (0.8, 1.2)	-
Metastasis-free survival	3.2 (1.5,4.8)	1.5 (1.2,3.6)	6.8 (4.5, 9.0)	-	-	2.8 (2.1,3.7)	4.1 (2.7,5.7)	1.1 (0.9, 1.6)	-
Overall Survival	3.4 (1.7,5.0)	1.8 (1.3,3.6)	6.9 (4.7, 9.2)	2.0 (1.6,2.4)	2.6 (2.0,4.0)	3.7 (2.9,4.6)	4.9 (3.3,6.6)	1.6 (1.3, 2.0)	5.0 (3.8,6.6)
Median Follow-up Time in years (IQR)	8.8 (6.3,11.2)	7.0 (6.0, 7.7)	6.4 (5.8, 7.0)	4.9 (3.6, 5.6)	10.3 (0.0, 11.8)	7.8 (6.2, 9.7)	5.5 (5.0, 7.5)	7.8 (6.2, 8.6)	10.8 (8.8,12.8)
Age									
<55	21 (17%)	7 (7%)	36 (12%)	13 (8%)	25 (16%)	167 (17%)	49 (16%)	19 (12%)	60 (19%)
55-64	60 (50%)	32 (33%)	119 (38%)	54 (34%)	65 (43%)	366 (38%)	80 (25%)	69 (45%)	117 (37%)
≥65	40 (33%)	57(59%)	156 (50%)	92 (58%)	63 (41%)	443 (45%)	188 (59%)	65 (43%)	140 (44%)

Treatment Arm									
Neoadjuvant Chemotherapy	62 (51%)	41 (43%)	151 (49%)	83 (52%)	82 (54%)	491 (50%)	158 (50%)	78 (51%)	158 (50%)
No Chemotherapy	59 (49%)	55 (57%)	160 (51%)	76 (48%)	71 (46%)	485 (50%)	159 (50%)	75 (49%)	159 (50%)
Sex									
Male	105 (87%)	77 (80%)	246 (79%)	125 (79%)	140 (92%)	863 (88%)	254 (80%)	124 (81%)	258 (81%)
Female	16 (13%)	19 (20%)	65 (21%)	34 (21%)	13 (8%)	113 (12%)	63 (20%)	29 (19%)	59 (19%)
Stage									
T0-T1	0 (0%)	0 (0%)	53 (17%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	2 (1%)	0 (0%)
T2	33 (27%)	16 (17%)	116 (37%)	50 (31%)	57 (37%)	334 (34%)	130 (41%)	24 (16%)	124 (39%)
T3	78 (64%)	34 (35%)	124 (40%)	85 (53%)	79 (52%)	567 (58%)	157 (50%)	85 (56%)	193 (61%)*
T4	10 (8%)	12 (13%)	13 (4%)	24 (15%)	16 (10%)	75 (8%)	23 (7%)	41 (27%)	-
Unknown	0 (0%)	34 (35%)	5 (2%)	0 (0%)	1 (1%)	0 (0%)	7 (25)	1 (1%)	0 (0%)

*T3 and T4 categories are merged for SWOG trial.

Table 5.1: Descriptive characteristics per individual trial.

I calculated the “O-E” and “V” statistics and log HR and its standard error for each outcome in each individual trial as described in 5.3.3. Kaplan-Meier plots were used to assess the proportionality hazards assumption for each outcome. In Figure 5.1, I present an example of Kaplan-Meier plots for the outcome of overall survival indicating that most trials provide no evidence of non-proportional hazards. Situations under which the curve declined more rapidly than other trials were observed (e.g., UK versus SWOG trial). Kaplan-Meier plots for other IPD outcomes are presented in the Appendix D.1 and provide similar interpretation with regards to the proportionality assumption.

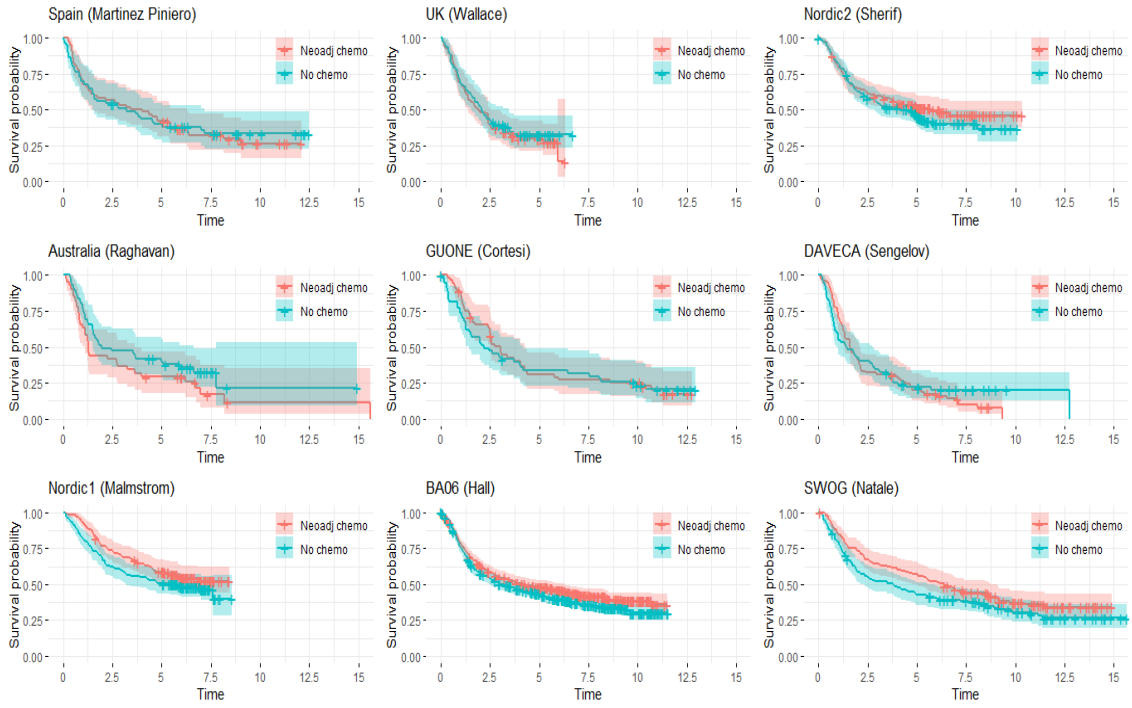


Figure 5.1: Kaplan-Meier plot for overall survival outcome.

5.5 Results from two-stage IPD meta-analysis

In Table 5.3 and Figure 5.2, I present the pooled effect estimates obtained from the use of two-stage IPD MA models for all outcomes analysed, alongside the estimates of between-study heterogeneity (τ) and I^2 estimates.

Model	Log-estimates (95% CIs)	SE	τ	I^2	Median % Total Censoring (IQR)	Median Event Probability (IQR)
Event Free Survival						
OR REML (2-stage)	-0.250 (-0.423, -0.077)	0.088	0.000	0	32% (26%,42%)	0.64 (0.56, 0.76)
HR O-E & V (2-stage)	-0.175 (-0.327, -0.022)	0.078	0.022	49		
HR Cox PH (2-stage)	-0.173 (-0.324, -0.023)	0.077	0.021	49		
HR clog-log (2-stage)	-0.085 (-0.262, 0.092)	0.090	0.034	57		
Local Recurrence Free Survival						
OR REML (2-stage)	-0.173 (-0.364, 0.018)	0.097	0.000	0	32% (18%,41%)	0.68 (0.55, 0.85)
HR O-E & V (2-stage)	-0.122 (-0.232, -0.012)	0.056	0.000	0		
HR Cox PH (2-stage)	-0.121 (-0.231, -0.011)	0.056	0.000	0		
HR clog-log (2-stage)	-0.020 (-0.232, 0.192)	0.108	0.038	59		
Metastasis-Free Survival						
OR REML (2-stage)	-0.027 (-0.409, 0.355)	0.195	0.133	66	35% (27%,43%)	0.65 (0.52, 0.80)
HR O-E & V (2-stage)	-0.125 (-0.281, 0.030)	0.079	0.013	35		
HR Cox PH (2-stage)	-0.122 (-0.280, 0.035)	0.080	0.014	36		
HR clog-log (2-stage)	-0.012 (-0.250, 0.227)	0.122	0.056	68		
Overall Survival						
OR REML (2-stage)	-0.147 (-0.308, 0.013)	0.082	0.000	0	35% (31%,45%)	0.62 (0.52, 0.71)

HR O-E & V (2-stage)	-0.114 (-0.214, -0.015)	0.051	0.000	0
HR Cox PH (2-stage)	-0.114 (-0.213, -0.015)	0.051	0.000	0
HR clog-log (2-stage)	-0.037 (-0.179, 0.105)	0.072	0.016	37

Table 5.2: Pooled effect estimates across different two-stage IPD meta-analysis models.

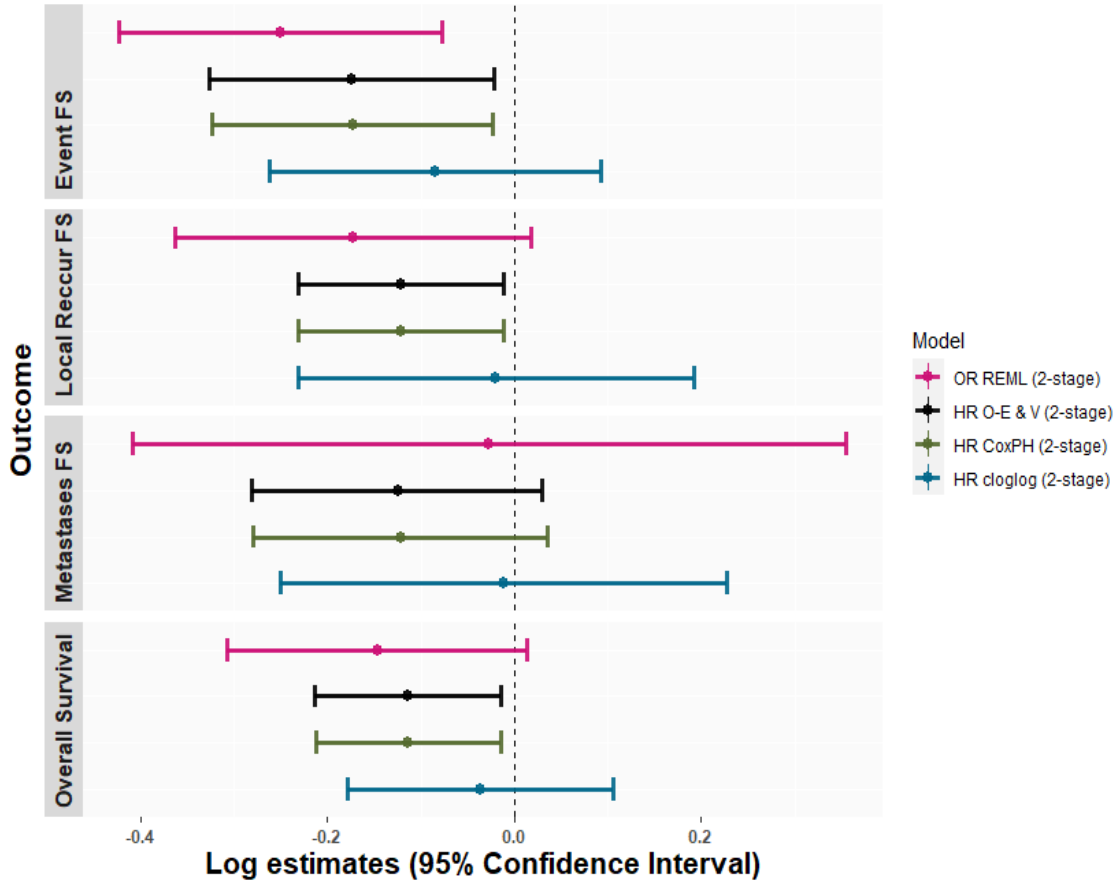


Figure 5.2: Overall meta-analytic estimates across different IPD two-stage meta-analysis models.

5.5.1 Event Free Survival

Event Free Survival was defined as "the time of randomisation until the first recurrence or progression or death, whichever happened first"[144]. There were 2444 participants within 8 trials contributing to meta-analysis of this outcome and 1617 events. The median event probability was 0.64 IQR (0.56, 0.64) (Table 5.3); DAVECA[149] study provided the highest event probability (0.92), Australia study provided the highest baseline risk (0.84), and Nordic1[152] the lowest event probability and baseline risk (0.47 and 0.54 respectively) (Figure 5.3). DAVECA trial reported the smallest percentage total censoring (13%) while Nordic1 reported the highest (49%).

The individual study estimates were almost identical for HR log-rank and HR Cox regression models. In larger trials (BA06, Nordic2[153] , Nordic1[147], and SWOG[154]), the OR estimates give better approximations to the gold-standard approaches. However, when the number of participants included in the studies decreases (e.g.

Spain[145], Australia, DAVECA) and the event probability increases, larger discrepancies can be observed, resulting also in reversal in the direction of the individual study estimates in the Spain and DAVECA trials, although all intervals are overlapping (Figure 5.3).

Weighting allocation for individual studies in the OR analysis is different to that in the HR analyses, with the HR clog-log approach providing a more similar allocation of study weights to the HR log-rank and HR Cox approaches (e.g. for Spain, Australia, BA06, DAVECA) than the OR analysis. The OR analysis did not detect heterogeneity as shown in Figure 5.3. The 95% confidence intervals for individual studies are wider in the OR analysis indicating that the intervals have much more overlap (even though the point estimates are just as heterogeneous or more heterogeneous than under the other scales), making the study results more in agreement with a heterogeneity estimate of 0 (Figure 5.3).

Even though small percent random censoring and short follow-up time are factors theoretically associated with smaller differences between the OR and HR scales in a TTE meta-analysis[35, 37, 39], I was not able to capture a specific pattern for this in this data set. The overall meta-analytic estimate of treatment effect was in the same direction across all MA models and similar between outcome scales, however the clog-log analysis failed to demonstrate statistical significance for the comparison of neoadjuvant chemotherapy versus no chemotherapy prior to local therapy (Table 5.3, Figure 5.3).

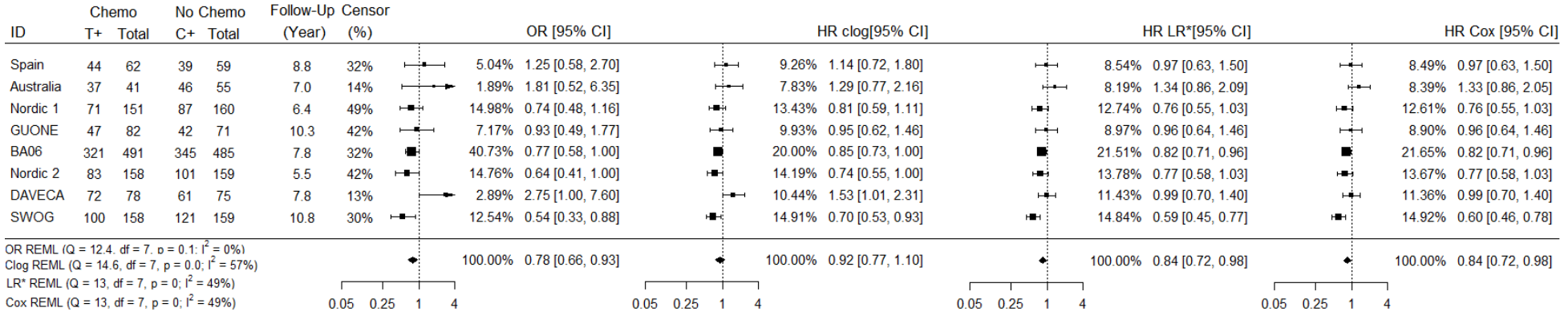


Figure 5.3: Forest plot of two-stage IPD meta-analysis for Event Free Survival.

5.5.2 Local Recurrence Free Survival

Local Recurrence Free Survival was defined as “the time from randomisation to first local recurrence or progression (after randomisation) or death”[144]. There were 6 trials contributing to meta-analysis of this outcome including 1974 participants and 1284 events. The median event probability was 0.68 IQR (0.55, 0.85); similarly, DAVECA study provided the highest event probability (0.92) and baseline risk (0.81) and Nordic1 the lowest event probabilities (0.46 and 0.54 respectively) (Figure 5.4). Furthermore, DAVECA and Nordic1 studies also reported the lowest and highest percentage total censoring (i.e. 13% and 50% respectively) (Figure 5.4).

The individual study estimates under HR log-rank and HR Cox models were almost identical. In larger trials, the OR and HR clog-log approaches provided better approximations to the individual trial estimates across all MA models (e.g. BA06). However, for smaller studies when the participants’ number decreases and the probability of event increases larger discrepancies in the individual trial estimates may appear (e.g. Australia, DAVECA), resulting also in particular circumstances in complete reversal of the results (i.e. Spain), although all intervals are overlapping (Figure 5.4).

The log-rank and Cox models have produced similar study weights in the meta-analysis compared to the OR and HR clog-log models. However, OR study weights for this outcome seem closer to trial weights from the gold standard approaches than do those from the HR clog-log model. This is because the between-study heterogeneity estimate ($\tau = 0.038$) obtained from the model in the HR clog-log analysis is not in agreement with the estimates obtained from the gold-standard approaches, even though it is still quite low and close to the estimates from other models (Table 5.2). As a consequence, this has affected both the individual study weights and the I^2 estimate (I^2=59%); however, the individual study estimates in the HR clog-log analysis are closer than estimates from the OR analysis to the corresponding estimates from the gold-standard approaches (Figure 5.4).

The 95% confidence intervals in the OR model were systematically wider than those obtained from the HR analyses, indicating that the standard error in the OR analyses relative to the HR analyses is substantially larger. Small percent random censoring and short follow-up time are usually associated with smaller

differences between the OR and HR scales in a TTE meta-analysis. However, in this dataset I was not able to detect such a pattern. The overall meta-analytic estimate is in the same direction across all MA models; however, the OR and HR clog-log analysis fail to capture statistical significance with regards to the effectiveness of platinum-based combination chemotherapy versus no chemotherapy (Table 5.3, Figure 5.4).

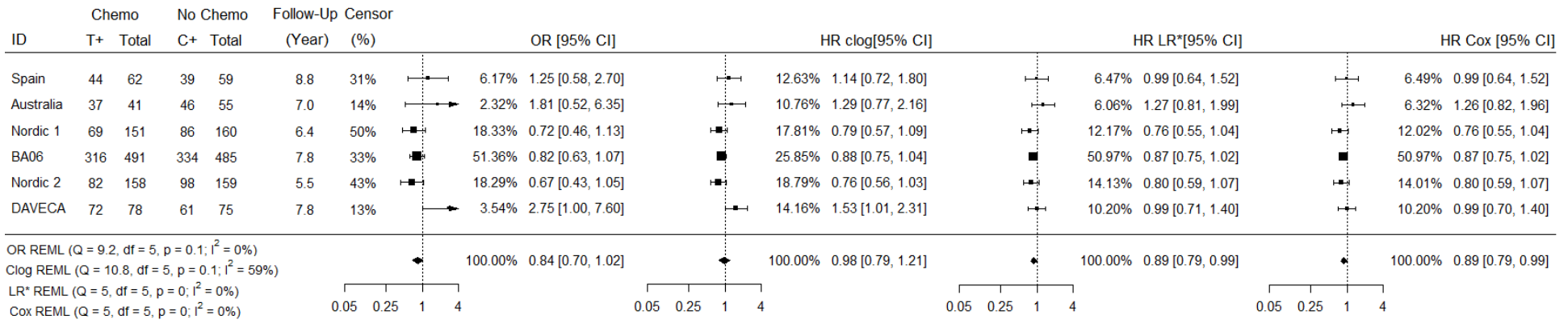


Figure 5.4: Forest plot of two-stage IPD meta-analysis for Local Recurrence Free Survival.

5.5.3 Metastasis-Free Survival

Metastasis-Free Survival was defined as “the time from randomisation to first metastasis (after randomisation) or death”[144]; meta-analysis of this outcome included 1974 participants from 6 trials and 1221 events. The median event probability was 0.65 IQR (0.52, 0.80) (Table 5.2). Similarly to other outcomes, DAVECA and Nordic1 studies had the highest and lowest event probability, and lowest baseline risk (i.e. 0.90 and 0.80 vs 0.46 and 0.53); Those studies reported also the lowest and highest percentage total censoring (i.e. 15% & 50% respectively) (Figure 5.5).

Similar to previous outcomes, the individual study estimates from HR log-rank and HR Cox models were almost identical. Larger studies (e.g. BA06) provide more stable individual trial estimates across all models. On the other hand, when the sample size reduces and the event probability increases in the trials, larger discordances with regards to the calculation of the individual study estimates across models are observed (e.g. Australia, Spain, DAVECA, Figure 5.5).

Furthermore, with respect to allocation of study weights, log-rank and Cox models have similar study weights compared to OR and HR clog-log models. The between-study heterogeneity estimates in the OR and HR clog-log analyses were larger than those from the gold-standard approaches and were $\tau = 0.133 \text{ and } \tau = 0.056$ respectively (Table 5.2). As a consequence, this has affected both the individual study weights and the I^2 estimates ($I^2 = 66\%, I^2 = 68\%$); the individual study estimates though in the HR clog-log analysis are closer than those in the OR analysis to the corresponding estimates from the gold-standard approaches (Figure 5.5).

The standard error and therefore the 95% confidence intervals in the OR model are systematically wider than the corresponding intervals from HR analyses. Similarly, censoring and follow-up times are characteristics for which I could not observe a particular pattern in their impact on the pooled effect estimates in this dataset. Finally, the overall meta-analytic estimate was in the same direction across all MA models favouring the same treatment arm (Figure 5.5).

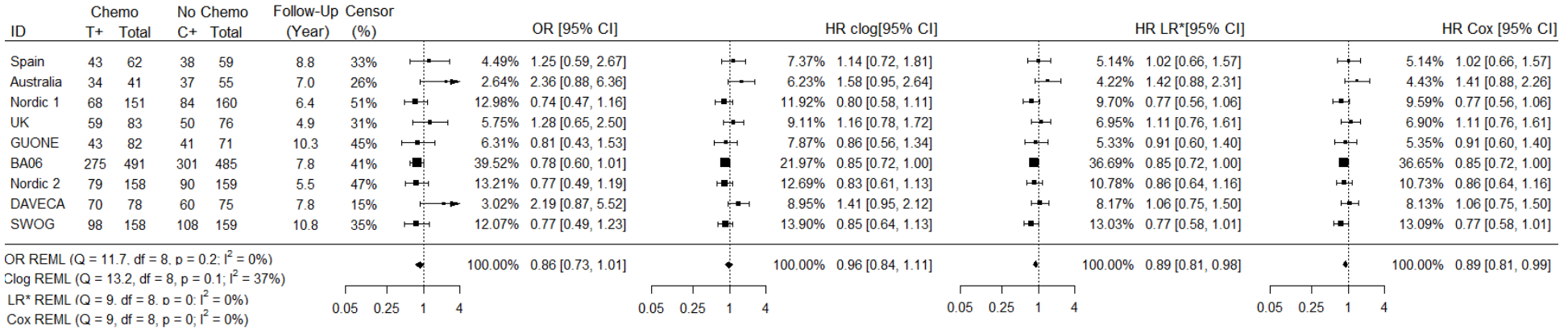


Figure 5.5: Forest plot of two-stage IPD meta-analysis for metastasis-free survival.

5.5.4 Overall Survival

Finally, overall survival, was defined as “the time from randomisation until death”[144]. Nine trials contributed to meta-analysis of this outcome with 2603 patients and 1617 events. The median event probability was 0.62 IQR (0.52, 0.71) (Table 5.3). DAVECA and Nordic1 studies had the 1) highest and lowest event probability (0.90 vs. 0.45), 2) highest and lowest baseline risk (0.80 vs. 0.53) and 3) lowest and highest percentage censoring respectively (15% vs. 51%) (Figure 5.6).

The individual study estimates of HR log-rank and HR Cox models were almost identical. In larger trials such as BA06, Nordic2[153] and SWOG[154], the HR clog-log and OR approaches provide better approximations to the gold-standard approaches for the individual trial estimates across all MA models. In contrast, when the sample size reduces and the event probability increases, larger discordances were observed with regards to the calculation of the individual study estimates between OR and HR “gold standard” MA models (e.g. Australia, Spain, DAVECA, Figure 5.6).

Log-rank and Cox models have identical individual study weights compared to OR and HR clog-log models. OR study weights seemed closer to the trial weights from the gold standard approaches than those from the HR clog-log model. The between-study heterogeneity estimate ($\tau = 0.016$) obtained from the model in the HR clog-log analysis is not in agreement with the estimates obtained from the gold-standard approaches, although it is still quite low and close to the estimates from other models (Table 5.3). This has affected both the individual study weights and the I^2 estimate (I^2 =37%); however, the individual study estimates in the HR clog-log analysis were closer than those from the OR analysis to the corresponding estimates from the gold-standard approaches.

The standard errors (and 95% CIs) in the OR analysis were systematically wider than the corresponding standard errors (and CIs) from HR analyses. For % total censoring and follow-up time I could not observe a particular pattern in their effect on the final pooled effect estimates across the models applied in the dataset. Finally, the overall meta-analytic estimate is in the same direction across all MA models, however the OR and clog-log analysis fail to capture statistical significance for the comparison of platinum-based combination chemotherapy versus no chemotherapy prior to local therapy (Table 5.3, Figure 5.6).

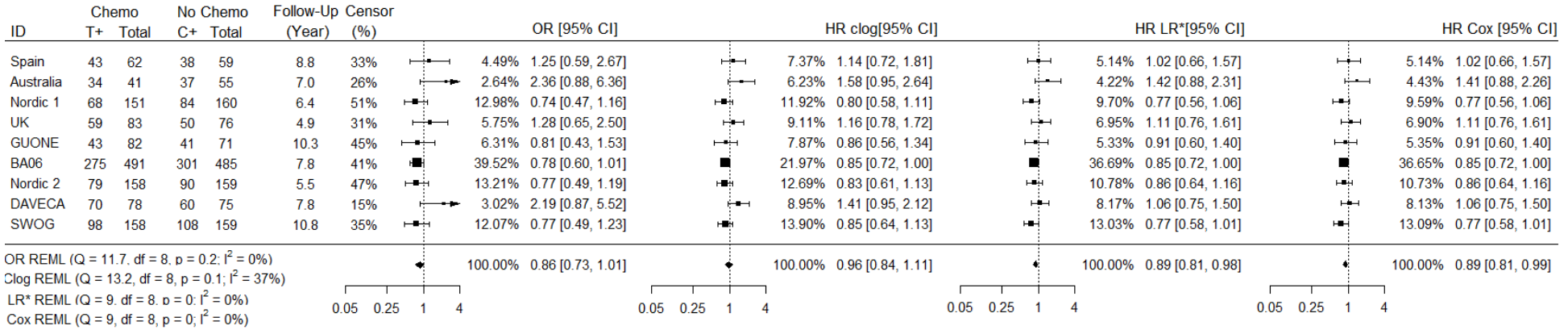


Figure 5.6: Forest plots comparing two-stage models in IPD meta-analysis for overall survival outcome.

5.6 Results from one-stage IPD meta-analysis

Similarly to two-stage models, Figure 5.7 and Table 5.3 present the pooled effect estimates obtained from the use of one-stage IPD meta-analysis models alongside the estimates of between-study heterogeneity (τ) and I^2 estimates. Assuming that one-stage Cox proportional hazard model is the "gold standard" approach, one-stage OR and one-stage clog-log models were compared to identify whether patterns in the results are similar to those observed for two-stage models.

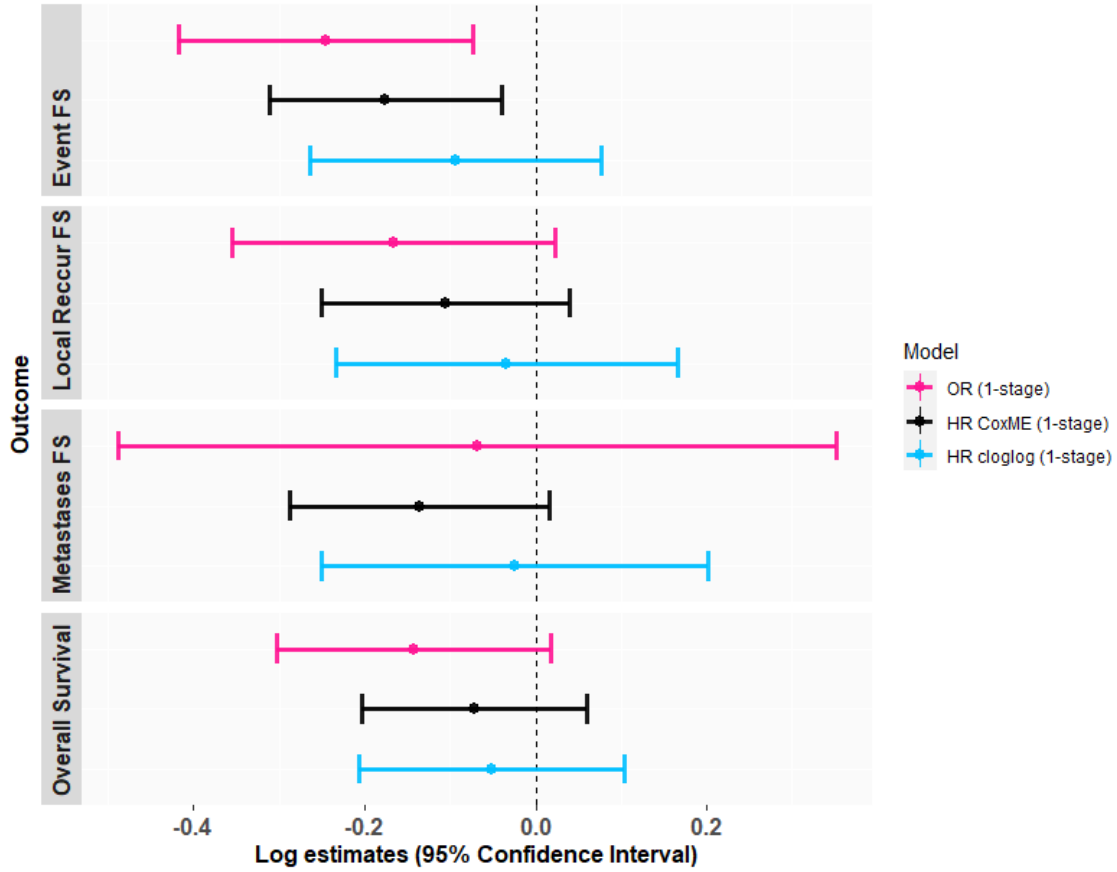


Figure 5.7: Overall meta-analytic estimates across different IPD one-stage meta-analysis models.

Across all models and outcomes, the overall pooled treatment effect estimate remained stable, favouring the same treatment arm. For all models, statistical significance for the treatment effect was only demonstrated in one-stage OR and HR Cox models for the event free survival outcome. The magnitude of the treatment effects was larger in the OR compared to the HR analyses with the exception of metastasis-free survival. In relation to two-stage models, the results

present a similar pattern for comparisons between different modelling approaches, however in one-stage models' statistical significance was not present in the "gold-standard" approaches for local recurrence free survival and overall survival outcomes.

An increased standard error was observed in the OR relative to the HR analyses for all outcomes apart from local recurrence free survival; I assumed that the individual study standard errors were wider in the OR analyses affecting also the overall standard error, even though I was not able to observe them individually as I did with two-stage models. With regards to two-stage models, a similar pattern was apparent.

The between-study heterogeneity estimates (τ) were similar across models except for metastasis-free survival where the estimate is much lower for the one-stage Cox model. The I^2 estimates were identical (i.e. 0%) between OR and HR clog-log analyses for all outcomes apart from metastasis-free survival and differ from those in one-stage HR Cox meta-analysis. Due to the fact that there was no straightforward way of calculating the I^2 estimates for the one-stage Cox model, the method I used might be overestimating the true extent of between-study heterogeneity.

Model	Log-estimates (95% CIs)	SE	τ	I^2
Event Free Survival				
OR (1-stage)	-0.245 (-0.417, -0.074)	0.088	0.000	0
HR Cox ME (1-stage)	-0.176 (-0.312, -0.040)	0.069	0.018	48
HR clog-log (1-stage)	-0.094 (-0.265, 0.077)	0.087	0.023	0
Local Recurrence Free Survival				
OR (1-stage)	-0.166 (-0.356, 0.023)	0.097	0.000	0
HR Cox ME (1-stage)	-0.105 (-0.250, 0.040)	0.074	0.014	43
HR clog-log (1-stage)	-0.034 (-0.234, 0.166)	0.102	0.022	0
Metastasis-Free Survival				
OR (1-stage)	-0.069 (-0.489, 0.351)	0.214	0.074	52
HR Cox ME (1-stage)	-0.236 (-0.287, 0.015)	0.077	0.015	44
HR clog-log (1-stage)	-0.025 (-0.251, 0.202)	0.115	0.038	74
Overall Survival				
OR (1-stage)	-0.143 (-0.303, 0.017)	0.082	0.000	0
HR Cox ME (1-stage)	-0.072 (-0.204, 0.059)	0.067	0.017	44
HR clog-log (1-stage)	-0.052 (-0.206, 0.103)	0.079	0.008	0

Table 5.3: Pooled effect estimates across different one-stage IPD meta-analysis models.

5.7 Discussion

Using IPD, I investigated whether important properties of TTE data such as percentage total censoring and follow-up times could additionally affect the results obtained from a MA when data are analysed using “gold-standard” approaches (such as Cox proportional hazards model and the log-rank test) as opposed to analysing the data as binary using the clog-log or the logit link where interpretation is conducted on a HR or an OR scale respectively.

I confirmed previous findings obtained from the CDSR that the method choice does matter[143]. Cox proportional hazards model (if the proportionality assumption holds) and the log-tank test (if “O-E” and “V” statistics can be obtained) are known to be suitable models for an IPD meta-analysis of TTE outcomes. However,

analysing these data as binary on an OR scale could be inappropriate especially when event probability is high. My analyses indicated discordancy both in the individual and pooled effect estimates. Smaller trials provided consistently different individual trial effect estimates in the OR relative to the HR analyses, with consequent impact on the pooled effect estimates. The confidence intervals for individual study results were systematically wider in the OR compared to HR analyses since they provided an increased within-study standard error, and consequently confidence intervals for pooled effect estimates were also wider. Compared to "gold-standard" approaches, a mixed pattern was observed for between-study heterogeneity and I^2 estimates in the OR and HR clog-log analyses. I observed TTE outcomes where the intervals have much more overlap making the study results more in agreement or τ estimates that were quite low and close to the estimates from "gold-standard" methods affecting though substantially the study weights and I^2 estimates.

The Cox proportional hazards model I used is considered the "gold-standard" approach for analysing IPD of TTE outcomes. It does not make any assumptions on the baseline hazard rate but requires the proportional hazards assumption meaning that the effect should be independent of time[155]. If this assumption is violated other suitable methods can be used such as Poisson regression models[156], restricted mean survival times (RMST)[157], and percentile ratios[158]. The use of the log-rank approach via the "O-E" and "V" statistics is the most popular approach for most IPD meta-analyses, perhaps due to the fact there is a lack of expertise and readily available software to fit the Cox proportional hazards model. Previous research indicated that the log-rank approach may give biased estimates for both treatment effect and heterogeneity estimates, however it is the easiest method to implement[159].

I was not able to identify situations where a model using the complementary log-log link is a more suitable approach than a model treating TTE as binary in a meta-analysis. For most outcomes, the individual study estimates in the HR clog-log analysis were closer to the corresponding estimates of the gold-standard approaches; however, this was not the case for the estimates of between-study heterogeneity and I^2. Real world evidence cannot facilitate providing a definitive answer because the true underlying model parameters are unknown. To overcome this limitation and to provide details on the preferable method a

satisfactory answer could not be purely based on empirical investigations and therefore a comprehensive simulation study will be carried out as further work.

It was not possible to explain whether censoring and follow-up time were distinct factors affecting the discordance among the MA estimates for this dataset since a) high event probability was a strong factor affecting the results as observed in previous chapters and b) I could not distinguish between random and fixed censoring given the data I had. Small % random censoring and short follow-up times are theoretically associated with smaller differences between the OR and HR scales in a TTE meta-analysis and therefore a combination of the factors obtained from Chapters 3-5 will be examined separately in a simulation study.

I did not implement the RMST method[64] in addition to the gold-standard approaches. A previous study conducted by Wei et al.[157] performing MA using RMST using the same IPD, indicated that degrees of departure from non-proportional hazards do not seem to be large in this dataset when a Grambsch and Therneau's (G-T) test[160] was performed. The results were dominated by trials in which the proportional hazards assumption was not violated and the results between RMST and a log-rank approach were similar. Therefore, I did not proceed with further application of RMST since the scope of this thesis was to examine differences between methods which account and do not account for the important properties of TTE data with interpretation on the HR and OR scale and not to compare which one of the methodologies using a HR approach was the best.

To my knowledge, there is very limited research performed (for example by Michiels et al.[161]) using IPD to assess the differences between analysing the data using the "gold standard" approaches (Cox PH model, log-rank test) on a HR scale compared to analysing the data the logit link on the OR scale as binary and no research specifically into using the clog-log link on a HR scale. I was able to demonstrate the impact of re-analysing IPD via various meta-analytic models interpreting the results on a different scale, confirming previous evidence and identifying a combination of characteristics that could influence the final pooled meta-analytic estimates. Careful consideration on the most appropriate approach depending on data availability is necessary.

5.8. Conclusion

In conclusion, my findings indicated that choice of method does matter and that smaller trials provided consistently different individual trial effect estimates in the OR compared to the HR analyses, with consequent impact on the pooled effect estimates. The confidence intervals for individual effect estimates and pooled effect estimates were wider in the OR analyses. There was no consistent pattern across methods for heterogeneity estimates. Findings from this chapter suggest that the influence of trial size, event probability and heterogeneity on differences between methods should be explored further in the planned simulation study. For the clog-log link approach, I observed a mixed pattern regarding whether it falls in between the “gold-standard” approaches and the binary model with a logit link. A comprehensive simulation study is necessary to examine and compare separately the factors affecting the results in a TTE meta-analysis.

Parts of this Chapter were presented as an oral presentation at the 42nd conference of International Society of Clinical Biostatistics and at the 2021 annual meeting of the Society for Research Synthesis Methodology.

6. A Simulation Study Comparing Methods for Meta-Analysis of Time-to-Event Outcomes

6.1 Chapter Overview

This Chapter is a simulation-based comparison of the methods applied in a time-to-event meta-analysis. In the previous chapters, I was able to identify that the correct choice of method for handling this type of data does matter, however, I was not able to distinguish specific patterns among the multiple factors observed affecting the magnitudes of discordances in the results between scales. Therefore, this chapter provides more definitive evidence on the most appropriate method for handling these data and assesses whether undesirable properties from treating data as binary can be reduced by the use of alternative methods such as the use of the clog-log link facilitating interpretation on a HR scale.

6.2 Introduction

The previous meta-epidemiological study[143], using real-world aggregate data from the CDSR of 2008 (Chapters 3, 4) and an IPD meta-analysis performed using data from the MRC CTU (Chapter 5), identified that dichotomising time-to-event outcomes may be adequate for low event probabilities but not for high event probabilities. Differences between scales arose mainly when event probability was high and could occur via differences in between-study heterogeneity or increased within-study standard error in the OR relative to the HR analyses. Additionally, the combination of censoring and follow-up times could affect the results, however, I was not able to identify to what extent these factors affect these differences. Details on the exact conditions under which the various methods provide a satisfactory answer could not be based purely on empirical studies. Hence, I performed a comprehensive simulation study allowing separate

examination of the factors appearing to affect the results obtained from the previous chapters.

Using simulation-based datasets, I performed a simultaneous comparison of the "gold standard" approaches (Cox and log-rank method) to the approximate methods (using the clog-log or logit link functions) for using aggregate data to conduct a TTE MA. I did not aim to compare directly approaches that provide an OR estimate to approaches providing a HR estimate but was interested to assess how well the method behaves as an approximation to the HR. The conditions under which I simulated the datasets were informed by the findings from previous chapters, in addition to information from the recent literature (Appendix E.2). Specifically, using this simulation study I tried to answer the following questions:

- If we analyse time-to-event outcomes as binary how much bias do we observe in the pooled $\hat{\theta}_{OR}$ compared to the pooled $\hat{\theta}_{HR}$?
- In which situations do we observe most bias focusing particularly on the role of event probability and random censoring?
- If bias exists, can we minimise it via the use of the clog-log link as an alternative method? Might we be willing to accept the bias observed in exchange for other good properties of the method?
- What is the relative precision of $\hat{\theta}_{OR}$ compared to $\hat{\theta}_{HR}$ ($i.e. Var(\hat{\theta}_{OR})/Var(\hat{\theta}_{HR})$)?
- How does the coverage compare among the methods?

The rest of the chapter is set out as follows. In the methods section (6.3), I describe the data generating mechanisms used in the study and the statistical models I applied for both IPD and aggregate data. In the results (Section 6.4), I perform a simultaneous comparison of the methods used, presenting important characteristics of the performance measures. These results are followed by a discussion (Section 6.5) exploring the conclusions and implications of my findings (Section 6.6).

6.3 Methods

The following section describes the data generating mechanisms used to create the simulated datasets and the meta-analysis methods I applied.

6.3.1 Data generating mechanisms

An outline of my simulation approach is presented below. In total, I generated 28 distinct scenarios for TTE meta-analysis covering a wide range of realistic scenarios informed by the literature[21, 107-109, 111], the empirical work conducted using the CDSR (Chapters 3, 4) and results from analysing an IPD obtained from the MRC CTU (Chapter 5). The initial 20 scenarios are described in 6.3.1.1 and another 8 were created under specific conditions described in 6.3.1.2.

6.3.1.1 Initial simulation scenarios

Table 6.1 presents the initial values chosen for the simulation parameters; each simulation scenario is a combination of the options in the table selected based on what I think may drive the differences in method performance and also for generalisability. Twenty scenarios were initially created as follows. The number of studies per meta-analysis is set at 2 levels (5 and 20) to represent small and large meta-analyses. The number of participants per trial is set at 3 levels (with a mean of 100, 400, and 1000 and a standard deviation of 15, 40, and 100) to represent small, medium and large study sizes. The log HR is set at 3 values (0, -0.3, -0.8) representing zero, medium and large treatment effect. The between-study variance is set at 4 values (0, 0.001, 0.05, 0.1 respectively) representing zero, near-zero, medium and large heterogeneity between the studies. Follow-up time and percentage random censoring are set simultaneously at 3 levels (1 year with 0% censoring, 3 years with 25% censoring, 5 years with 40% censoring) to represent small, medium and large proportions of participants censored and follow-up times within trials.

#	**Parameters**	**Values**
1	Number of Studies per Meta-analysis (K)	5, 20
2	Study sample size (N)	~100, ~400, ~1000
3	Log HR	0, -0.3, -0.8
4	Between-Study variability (τ^2) of log HR	0, 0.001, 0.05, 0.1
5	Follow-up time (St)	~1 year, ~3 years, ~5 years
6	Percentage (%) random censoring (C)	0%, 25%, 40%

Table 6.1: Initial simulation parameters selected for the study.

Event times are constructed to have proportional hazards by simulating from a Weibull distribution for the majority of simulation scenarios ($\lambda_e = 0.1, \lambda_c = 0.05, \gamma_e = \gamma_c = 2$); Weibull distribution is considered as one of the most appropriate distributions to create survival times[162, 163]. Scenarios were examined by varying characteristics one by one from a baseline setting (N=400, Log HR=-0.3, $\tau^2 = 0.05$, St=3, C=25%) apart from percentage random censoring and follow-up time which changed simultaneously since it is expected that larger follow-up times cause larger percentage of random censoring. Scenario under no effect size (i.e. Log HR=0) was designed using the baseline setting above, whereas scenario 0 was designed as a more extreme scenario as follows (N=1000, Log HR=0, $\tau^2 = 0$, St=5, C=0%) including 5 or 20 studies.

6.3.1.2 Additional simulation scenarios

Since event probability and percentage random censoring are likely to affect method performance, I created another four scenarios involving 5 and 20 studies with ($\lambda_e = 0.05, \lambda_c = 0.04, \gamma_e = \gamma_c = 2$); for two of them I only changed the event probability from the baseline setting described above and for the other two I changed event probability and increased the percentage of random censoring. Two scenarios were specifically designed produce 80% power for a random-effects meta-analysis under the Cox proportional hazards model, including 5 or 20 studies ($N = 400, Log\ HR = -0.3, St = 3, C = 25\%, \tau^2 = 0.027$ for 5 studies, $\tau^2 = 0.2$ for 20 studies). Finally, another two scenarios including 5 or 20 studies were designed aiming to distinguish between the effect of censoring and follow-up times favouring the use of the clog-log link as follows: (N=400, Log HR=-0.3, $\tau^2 = 0.05$, St=5, C=0%). Table 6.2 presents all the simulation scenarios used in this chapter.

Simulation Scenarios	Values of simulation parameters under 5 and 20 trials per meta-analysis
Initial Simulation Scenarios	
Scenario 0	N=1000, Log HR=0, $\tau^2 = 0$, St=5, C=0%
Base Case	N=400, Log HR=-0.3, $\tau^2 = 0.05$, St=3, C=25%
Short F-up*	N=400, Log HR=-0.3, $\tau^2 = 0.05$, St=1, C=0%
Long F-up*	N=400, Log HR=-0.3, $\tau^2 = 0.05$, St=5, C=40%
Large heterogeneity	N=400, Log HR=-0.3, $\tau^2 = 0.1$, St=3, C=25%
Small heterogeneity	N=400, Log HR=-0.3, $\tau^2 = 0.001$, St=3, C=25%
Large effect size	N=400, Log HR=-0.8, $\tau^2 = 0.05$, St=3, C=25%
No effect size	N=400, Log HR=0, $\tau^2 = 0.05$, St=3, C=25%
Small sample size	N=100, Log HR=-0.3, $\tau^2 = 0.05$, St=3, C=25%
Large sample size	N=1000, Log HR=-0.3, $\tau^2 = 0.05$, St=3, C=25%
Additional Simulation Scenarios	
Small P(Event)	N=400, Log HR=-0.3, $\tau^2 = 0.05$, St=3, C=25%, $\lambda_e = 0.05, \lambda_c = 0.04, \gamma_e = \gamma_c = 2$
Large % R_cens+Small P(Event)[†]	N=400, Log HR=-0.3, $\tau^2 = 0.05$, St=5, C=40%, $\lambda_e = 0.05, \lambda_c = 0.04, \gamma_e = \gamma_c = 2$
80% Power	5 trials per MA: N=400, Log HR=-0.3, $\tau^2 = 0.027$, St=3, C=25% 20 trials per MA: N=400, Log HR=-0.3, $\tau^2 = 0.2$, St=3, C=25%
Long Follow-up+0% R_cens	N=400, Log HR=-0.3, $\tau^2 = 0.05$, St=5, C=0%

*F-up: Follow-up time; [†]R_cens=Random censoring; P(Event)=Probability of event

Table 6.2: Exact simulation scenarios applied of the chapter.

I generated 1000 random meta-analyses for each scenario. The exact tables are presented in Appendix E.2, together with additional information collected from the literature that helped to inform the choice of the parameters.

6.3.2 R software use for simulations

Independent simulation datasets were generated for each of the 28 scenarios using the R statistical software (Version 4.1.1). I used a starting seed (seed=2109990) which remained fixed in order to allow potential future replication of the simulation datasets. The simulation datasets were used, and various time-

to-event meta-analysis models were applied. The reliability of the simulation was confirmed as all our simulation scenarios did run without producing any convergence problems[164]. The R code used for this simulation study is presented in Appendix E.1.

6.3.3 Estimands

The estimand is the log HR βx for the treatment effect, whose true value is -0.3. I evaluated the performance of log OR approach as if the estimand would be the log HR, in order to assess how well this method behaves as an approximation to the HR.

6.3.4 Two-stage meta-analysis models for IPD

6.3.4.1 Model description for Cox proportional hazards model and log-rank approach

In Chapter 5 (sub-sections 5.3.3.1-2), I explained how I obtained a HR and its standard error and “O-E” and “V” statistics via the use of a Cox proportional hazards model and a log-rank test respectively. For each trial per simulation dataset, I applied the same methodology in this chapter, using both methods which account for censoring and follow-up times. Then, the HR and standard error data were entered in a two-stage meta-analysis model. Additionally, information on the “O-E” and “V” statistics were obtained when I performed testing of the survival curve differences (see 5.3.3.1). The “O-E” and “V” statistics were entered in a two-stage MA model.

6.3.4.2 Model Fitting for Cox proportional hazards model and log-rank approach

The estimated log hazard ratios for individual studies were given by:

$$y_i = \begin{cases} \log \text{HR obtained from Cox PH model for HRs (Chapter 5)} \\ \text{Equation 4.1 for HRs using "O} - \text{E" \& "V" statistics (Chapter 4)} \end{cases}$$

The corresponding sampling variances for the Cox and log-rank test were given as follows. For the Cox proportional hazards model the standard error of the log HR was squared to give the variance. For the log-rank test, I obtained the variance using the calculations presented in Chapter 4 from Equation 4.2 using the “O-E” and “V” statistics. Using these estimates and sampling variances I fitted two-stage random-effects models incorporating between-study heterogeneity variance. The models were implemented via the “rma.uni” command from “metafor” package in R.

6.3.5 Two-stage meta-analysis models using aggregate data

6.3.5.1 Model Description using aggregate data

From the IPD data, I was able to obtain binary summaries for each trial per simulation dataset. I modelled the "binary" data obtained using a normal approximation to binomial likelihood with a clog-log link on a HR scale ignoring censoring and follow-up times. Then I applied a model for the same data, assuming a binomial likelihood and a logit link[133], on an OR scale ignoring the same information as the aforementioned model.

6.3.5.2 Model Fitting using aggregate data

The estimated log hazard ratios and log odds ratios were given by:

$$y_i = \begin{cases} \text{Equation 3.2 for HRs using the clog} - \text{log link (Chapter 3)} \\ \text{Equation 3.1 for ORs using the logit link (Chapter 3)} \end{cases}$$

The corresponding variances were given by:

$$s_i^2 = \begin{cases} \text{Equation 3.4 for HRs using the clog} - \text{log link (Chapter 3)} \\ \text{Equation 3.3 for ORs using the logit link (Chapter 3)} \end{cases}$$

I estimated the study-specific log odds ratios or log hazard ratios, y_i and their within-study variances s_i^2 as shown above and fitted a standard two-stage random-effects model to these. The models were implemented via the "rma.uni" command from "metafor" package in R. One-stage meta-analysis models were not applied to avoid the additional complexities of including these in a simulation study.

6.3.6 Performance Measures

The performance of the methods was quantified via calculation of bias (also known as systematic error), relative precision, root mean squared error, model based standard error, coverage and power. Each of these performance measures was stored for the 28 simulation scenarios of 1000 iterations. In Table 6.3, I define the performance metrics I used in this study.

Metric	Formula	Description
Bias	$\sum_{i=1}^{1000} \frac{\widehat{log\theta}_i - log\theta}{1000}$	Average difference between the true simulated HR and its estimate across 1000 simulation replicates in a simulation scenario. Desirable to be near zero.
Relative precision	$100((\frac{\widehat{EmpSE}_a}{\widehat{EmpSE}_b})^2 - 1)$, where $EmpSE = \sqrt{Var(\widehat{\theta}_i)}$	The relative increase (or decrease) in precision when using one method (b) (i.e. "O-E" &"V" statistics (HR), logit link (OR), clog-log link (HR)) relative to another (a) (i.e. Cox PH (HR)).
Root mean squared error	$\sum_{1}^{1000} \frac{(\widehat{log\theta}_i - log\theta)^2}{1000}$	The squared average difference between the true simulated HR and its estimate across 1000 simulation replicates in a simulation scenario. Desirable to be near zero.
Model based standard error	$\sqrt{\frac{1}{1000}\sum_{i=1}^{1000} \widehat{Var}(\widehat{\theta}_i)}$	The square root of the summary of the simulated variance of the HR across 1000 simulation replicates in a simulation scenario. Desirable to be equal to the empirical SE.
Coverage	$\frac{1}{1000}\sum_{i=1}^{1000} 1(\hat{\theta}_{low,i} \leq \theta \leq \hat{\theta}_{upp,i})$	The proportion of times the two-sided 95% CI of the estimated summary HR contains the true HR. Desirable to be near 95%.
Power	$\frac{1}{1000}\sum_{i=1}^{1000} 1\begin{pmatrix}(\hat{\theta}_{low,i} > 0) \\ or\ (\hat{\theta}_{upp,i} < 0)\end{pmatrix}$	The proportion of times the two-sided 95% confidence interval of the estimated summary HR does not contain 0.

Table 6.3: Description of performance measures used in simulation analysis.

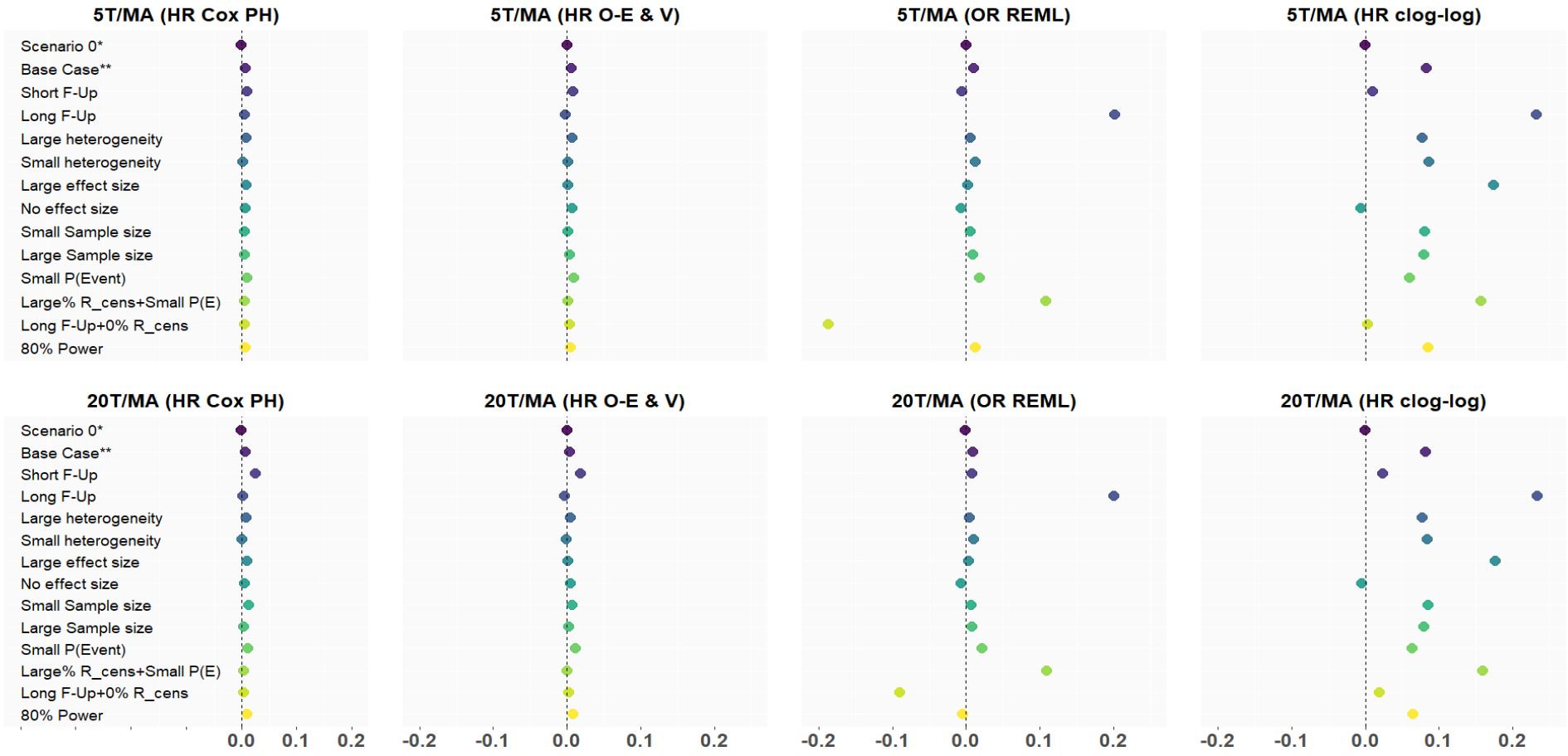
6.4 Results

6.4.1 Bias

Bias was obtained per simulation scenario for each method applied indicating how much the average estimate exceeds the true estimate; the reference method was the Cox proportional hazards model. Figure 6.1 indicates the results obtained using various meta-analysis methods (i.e. columns) across different simulation scenarios (i.e. rows).

Initially I observed that bias was similar between the scenarios including 5 or 20 trials. With regards to the Cox proportional hazards model and the log-rank test using the “O-E” and “V” statistics, no bias was observed across all simulation scenarios (i.e. columns 1-2, Figure 6.1). In those scenarios where the effect size was zero, the analysis of the data as binary did not demonstrate bias compared to analysing the data accounting for their natural properties (i.e. Scenario 0). Bias was low for the majority of scenarios analysing TTE data as binary using the logit link (i.e. OR REML), except in those where large % random censoring (~40%) and long follow-up time (~5 years) were present. Across most simulation scenarios, I identified more bias when the data were analysed as binary under the clog-log link (column 4, Figure 6.1) on the HR scale. Even though there was a theoretical assumption that the clog-log link function can be used as a useful alternative to analysing the data as binary on the HR scale, the bias observed in the results was much larger than the bias obtained from the logit link (column 4, Figure 6.1).

Additionally, in the scenario with small % random censoring and short follow-up time, the bias across all the methods was similar (row 3, Figure 6.1). In the presence of medium follow up time regardless of the amount of heterogeneity and the sample size treating the data as binary using the logit link (and not the clog-log link) might be acceptable because the bias was very low; therefore, heterogeneity and sample size did not seem to affect bias (rows 5-6, 9-10, Figure 6.1). For larger percentage random censoring and therefore length of follow-up time more bias was observed for both link functions that treat the data as binary (row 4, Figure 6.1). Finally, for those scenarios designed with large effect size large bias was observed when data were treated as binary using the clog-log link (row 7, Figure 6.1). The tables including the exact numbers obtained for bias are presented in Appendix E.2.



Note: 5T/MA, 20T/MA: 5 or 20 trials per meta-analysis; Each row within a panel is a different data generating mechanism (DGM); Upper and lower rows of panels are DGMs with K = 5 and 20; Columns of panels are different analysis methods; Scenario 0: Large sample size, No effect size, No heterogeneity, Long F-Up; Base Case**: Medium sample size, medium effect size, medium heterogeneity, medium follow up; Other scenarios change from base case; R_cens=Random censoring; P(E)=P(Event)

Figure 6.1: Bias observed per simulation scenario across different meta-analysis models.

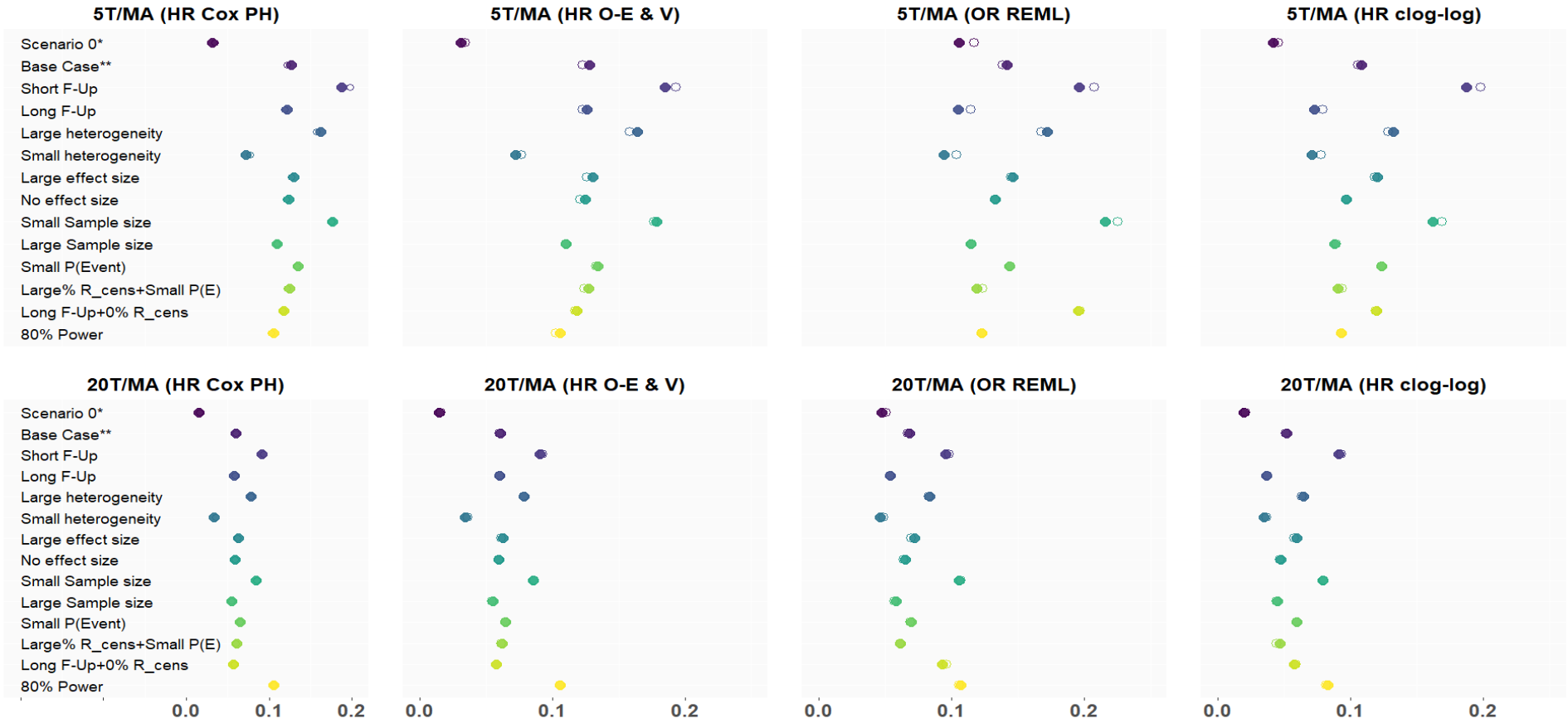
6.4.2 Empirical & Model Based Standard Errors

The empirical and model based standard errors were obtained per simulation scenario for each method. The empirical standard error measures the spread of the estimator, and the model based standard error should be equal to the empirical one. Figure 6.2 shows only how close the empirical (i.e., filled circle) and model based standard errors (i.e., hollow circles) are without aiming to compare empirical standard errors between the methods; this is covered in section 6.4.3.

The empirical and model based standard errors were closer to each other in the scenarios including 20 trials per meta-analysis than in those including 5 trials per meta-analysis due to the amount of information involved in the meta-analysis.

For those scenarios designed with 20 trials per meta-analysis, I observed similar results across different methods for MA of TTE data for the majority of situations. In the presence of zero or small heterogeneity or short follow-up times the models are overestimating the standard error; the overestimation/underestimation of standard error is relatively small (<2%) for most scenarios. For the rest of the simulation scenarios the differences between the empirical and model based standard errors seem negligible (Figure 6.2).

On the other hand, for the scenarios designed with 5 trials per meta-analysis a slightly different pattern was observed apart from the cases with small or zero heterogeneity where the models seem to overestimate the standard error as in 20 trials per meta-analysis. However, there were a lot of scenarios (e.g., short follow-up time, large follow-up, small sample size) under which the standard error is overestimated to a similar extent by the use of the complementary log-log link or the logit link; the same pattern was also true for the Cox proportional hazards model and the log-rank approach in the scenario where short follow-up time was present (Figure 6.2).



Note: 5T/MA, 20T/MA: 5 or 20 trials per meta-analysis; Each row within a panel is a different data generating mechanism (DGM); Upper and lower rows of panels are DGMs with K = 5 and 20; Columns of panels are different analysis methods; Scenario 0: Large sample size, No effect size, No heterogeneity, Long F-Up; Base Case**: Medium sample size, medium effect size, medium heterogeneity, medium follow up; Other scenarios change from base case; R_cens=Random censoring; P(E)=P(Event)

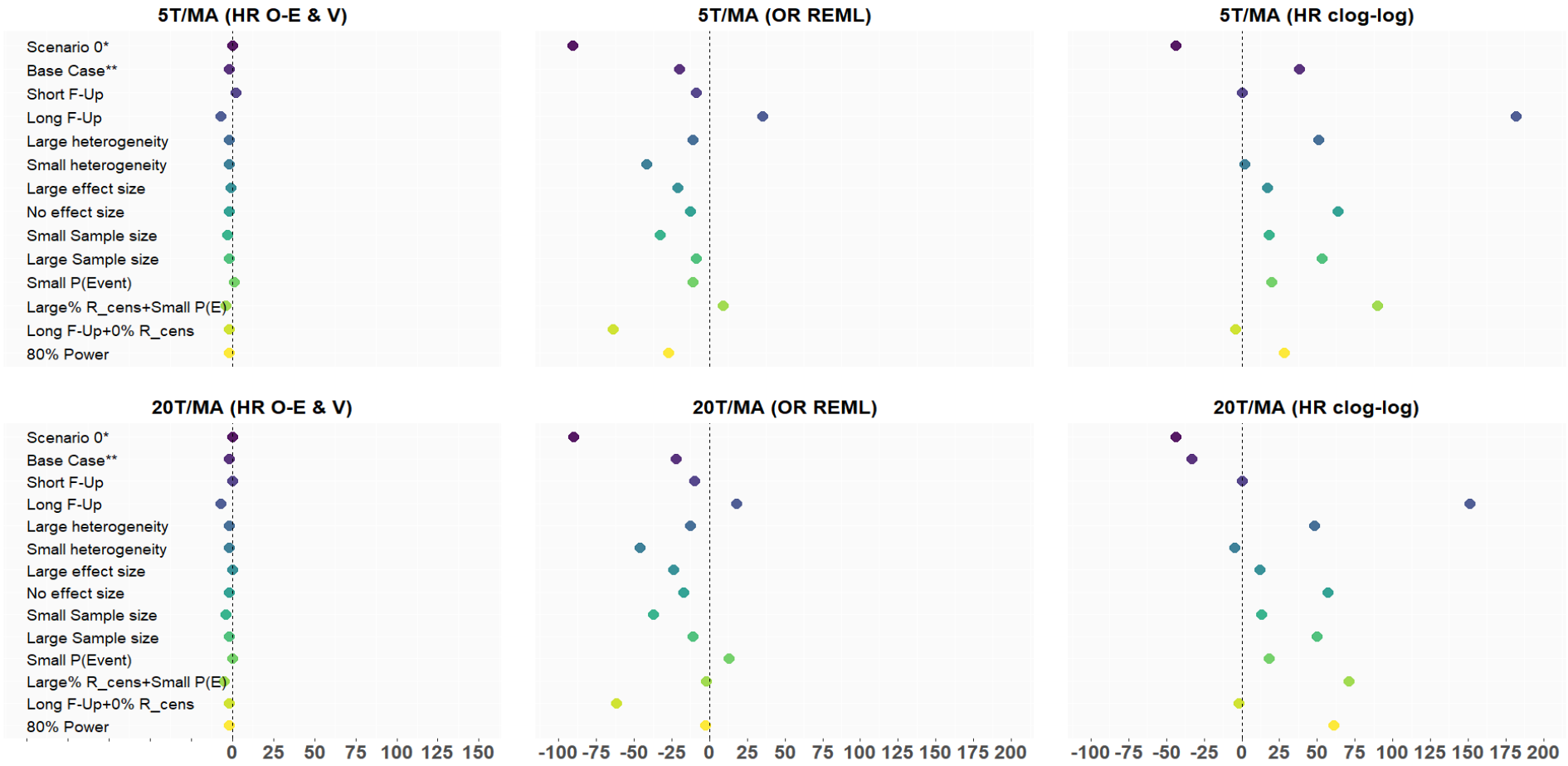
Figure 6.2 Empirical and model-based standard errors obtained per simulation scenario across different meta-analysis models.

6.4.3 Relative % increase in precision

The percent increase (or decrease) in precision relative to the Cox proportional hazards model per simulation scenario for each method applied was obtained. A similar pattern was observed in the simulation results when 5 or 20 trials were involved in the meta-analysis. Initially, as expected, I observed that the log-rank approach provided quite precise estimates relative to the Cox proportional hazards model (Figure 6.3).

In most scenarios, analysing the data on the HR scale using the clog-log link shows an increase in relative precision, while analysing data on an OR scale using the logit link shows a reduction in precision. Specifically, analysing the data as binary using the logit link was much less precise than the Cox proportional hazards model since binary analyses are throwing away data (Figure 6.3).

On the other hand, the increased precision observed for the clog-log link should be cautiously interpreted since according to the literature, a method that has increased bias towards the null may have small empirical standard error as a result of the bias[165]. Therefore, analysing the data as binary using the clog-log link in the presence of the bias towards the null discussed in 6.4.1 causes the meta-analytic estimates to appear more precise (Figure 6.3). The corresponding tables including the exact numbers obtained for the relative increase/decrease in precision are presented in Appendix E.2.



Note: 5T/MA, 20T/MA: 5 or 20 trials per meta-analysis; Each row within a panel is a different DGM; Upper and lower rows of panels are DGMs with K = 5 and 20; Columns of panels are different analysis methods; Scenario 0: Large sample size, No effect size, No heterogeneity, Long F-Up; Base Case**: Medium sample size, medium effect size, medium heterogeneity, medium follow up; Other scenarios change from base case; R_cens=Random censoring; P(E)=P(Event)

Figure 6.3: Relative percent (%) increase in precision per simulation scenario across different meta-analysis models.

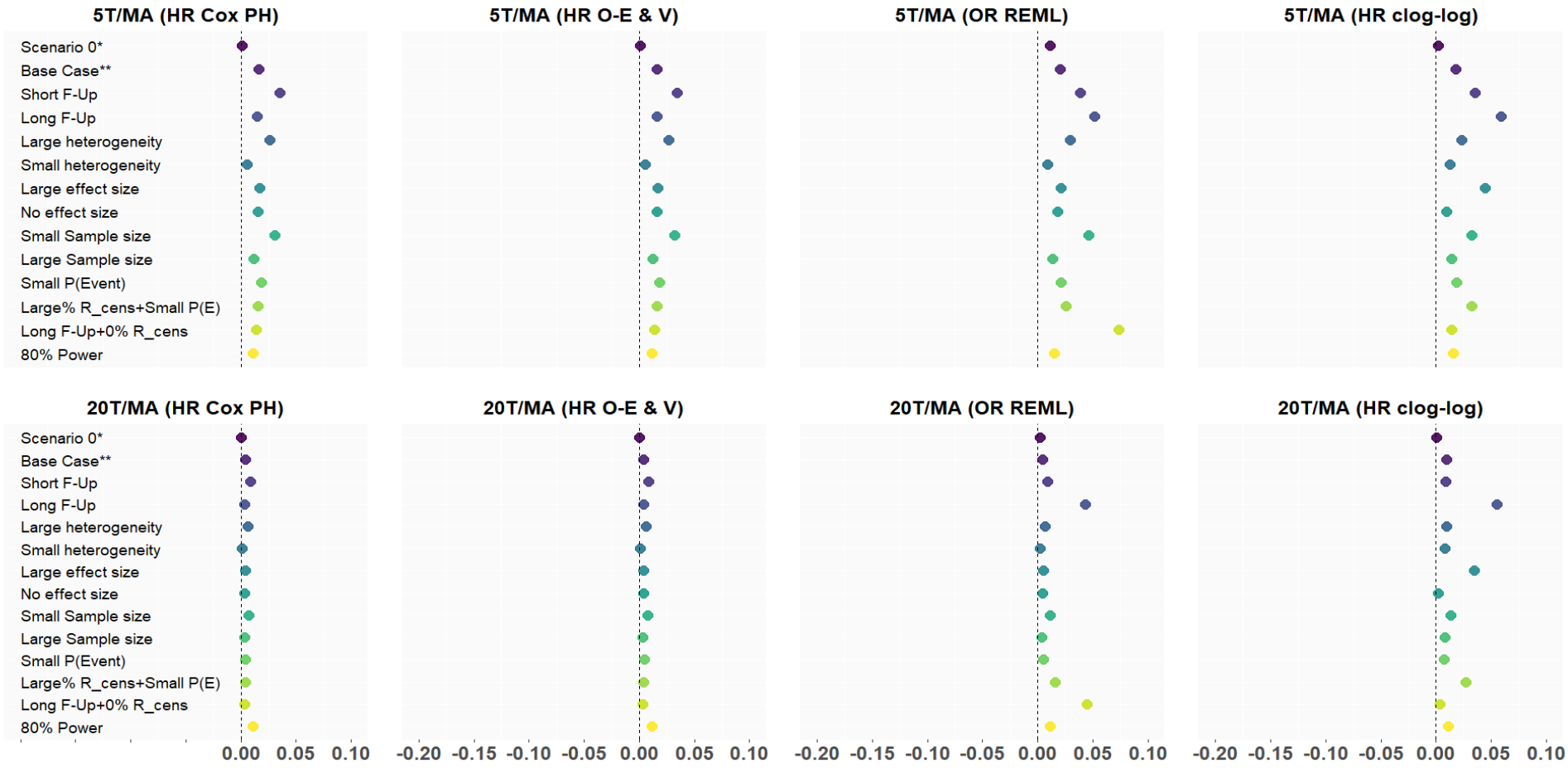
6.4.4 Mean squared error

The mean squared error was also obtained per simulation scenario as shown in Figures 6.4 and 6.5. Figure 6.5 was created to facilitate comparison of the methods applied and can be considered as another representation of the same results. The mean squared error indicated the overall performance of an estimator since it integrates both bias and variance; it was desirable for this to be close to 0.

A slightly different pattern was observed in the scenarios involving 20 trials per meta-analysis compared to those involving 5 trials per meta-analysis due to the amount of information involved. For most simulation scenarios, the mean squared error was closer to 0 for trials involving 20 trials per meta-analysis compared to those involving 5 trials per meta-analysis. I also observed that the Cox proportional hazards model and the log-rank approach provide similar mean squared error estimates (Figure 6.4).

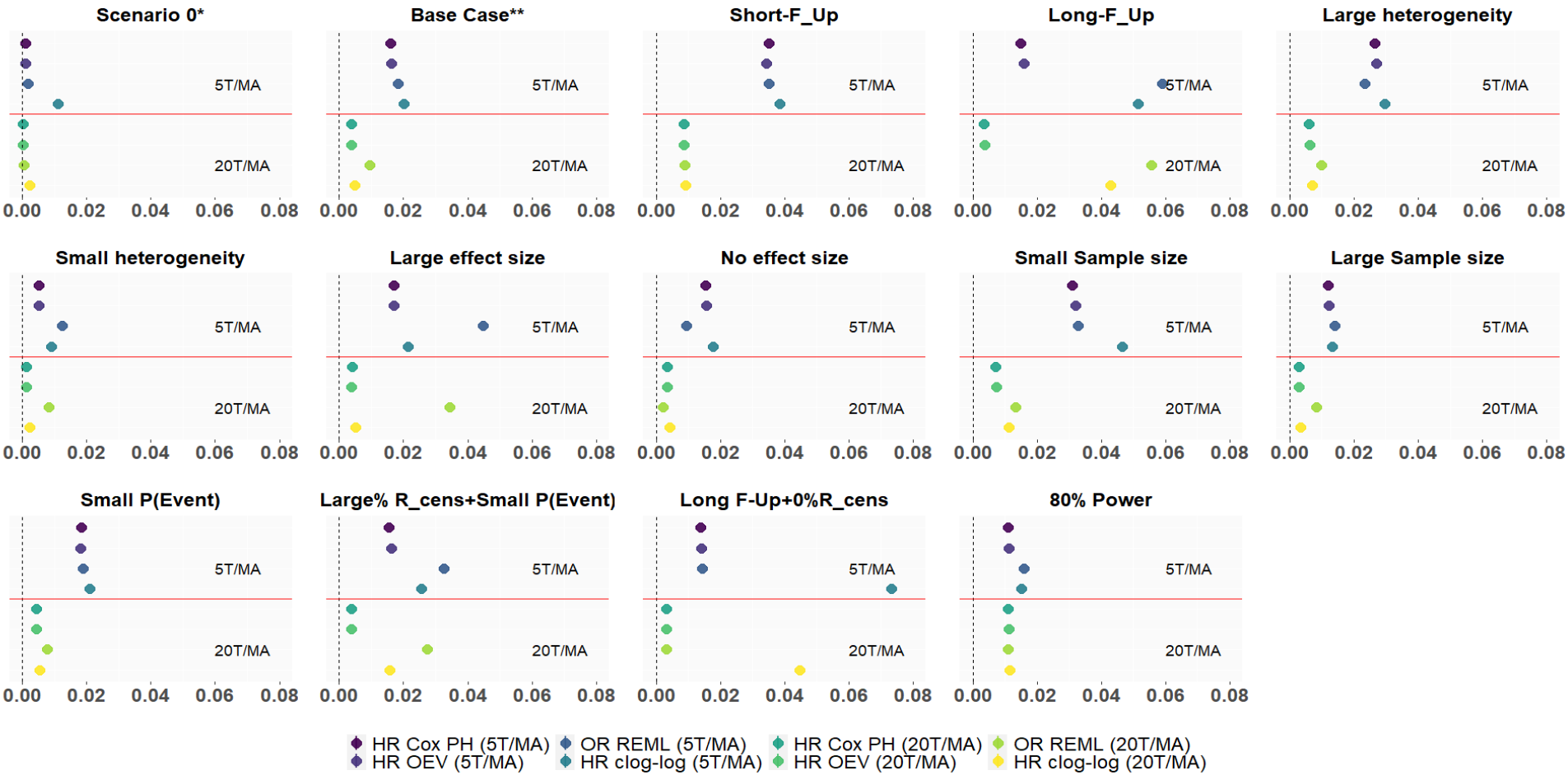
Specifically, for scenarios including 20 trials per meta-analysis, a small mean squared error was observed (i.e. MSE<0.02, actual numbers are presented in Appendix E.2). In the scenarios with longer follow-up time, large % random censoring and/or small event probability, the mean squared error is larger for both methods treating the data as binary (i.e. using both the clog-log and logit link). In the presence of large effect size, particularly the clog-log link approach (and not the logit approach) is performing badly since it is biased with a large model based standard error (Figure 6.4).

In the scenarios with long follow-up time, large % random censoring and/or small event probability, the mean squared error was even larger for both methods treating the data as binary (i.e. using both the clog-log and logit link). For the scenario created under small sample size, both the clog-log link and the logit link provided larger mean squared errors; the former mainly driven from the larger bias observed (Figure 6.1) and the latter one mainly driven from decreased precision (Figure 6.3). Finally, in the presence of large effect size , particularly the clog-log link approach (and not the logit approach) is performing badly, similarly to the 20 trials per meta-analysis scenario, since it is biased with a large model based standard error (Figure 6.4).



Note: 5T/MA, 20T/MA: 5 or 20 trials per meta-analysis; Each row within a panel is a different data generating mechanism (DGM); Upper and lower rows of panels are DGMs with K = 5 and 20; Columns of panels are different analysis methods; Scenario 0: Large sample size, No effect size, No heterogeneity, Long F-Up; Base Case**: Medium sample size, medium effect size, medium heterogeneity, medium follow up; Other scenarios change from base case; R_cens=Random censoring; P(E)=P(Event)

Figure 6.4: Mean squared error obtained per simulation scenario across different meta-analysis models.



Note: 5T/MA, 20T/MA: 5 or 20 trials per meta-analysis; Each row within a panel is a different meta-analysis method applied to meta-analyses containing either 5 or 20 trials; Panels are different DGMs; Scenario 0: Large sample size, No effect size, No heterogeneity, Long F-Up; Base Case**: Medium sample size, medium effect size, medium heterogeneity, medium follow up; R_cens=Random censoring; P(E)=P(Event)

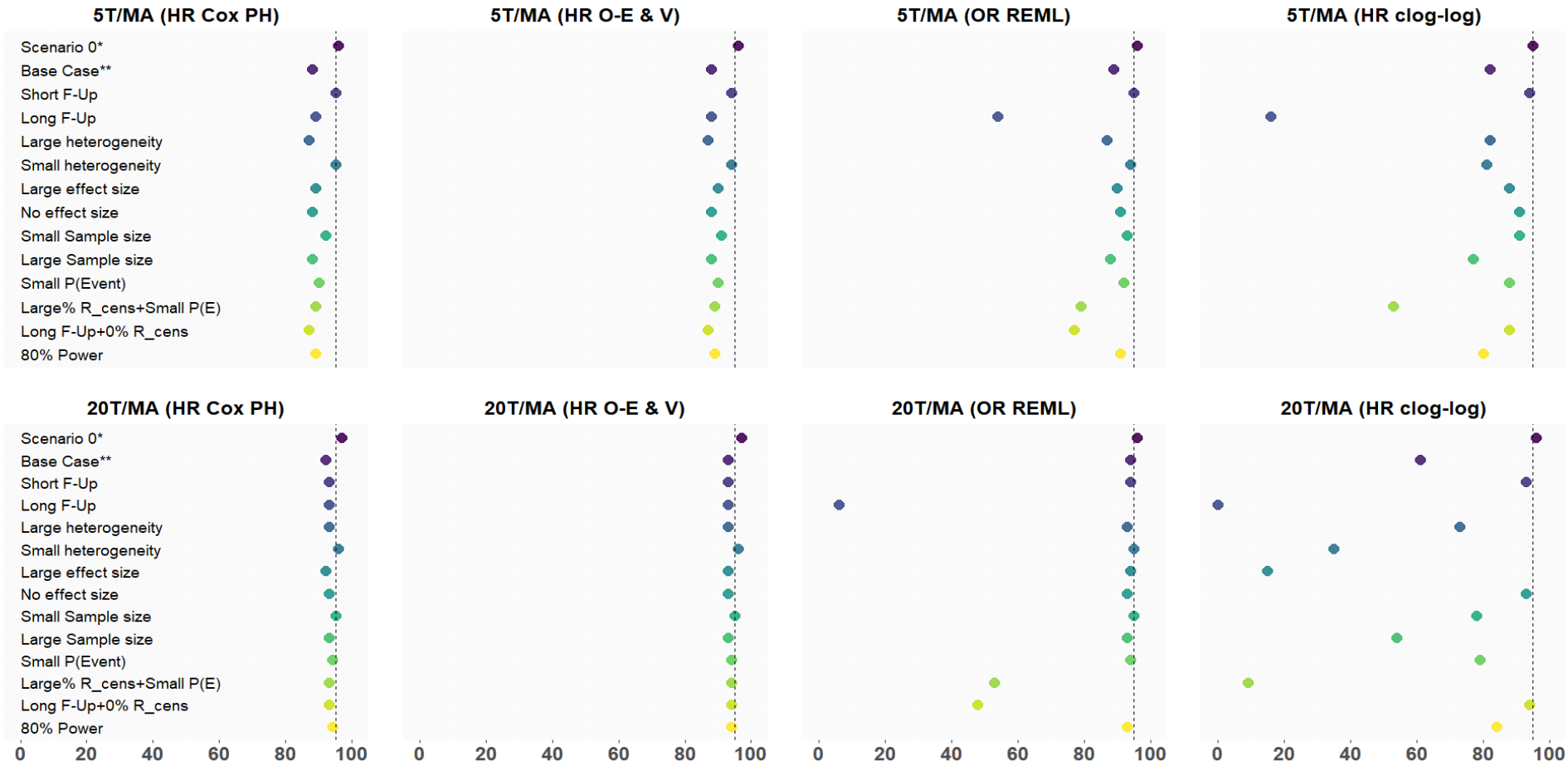
Figure 6.5: Another representation of the mean squared error obtained per simulation scenario across different meta-analysis models.

6.4.5 Coverage

Coverage for each simulation scenario per method was obtained (Figure 6.6) and it was desirable to be close to 95%. For those scenarios consisting of 20 trials per meta-analysis I initially observed that coverage is close to 95% for the majority of scenarios in three out of the four methods applied. Specifically, the Cox model, log-rank approach and treating the data as binary using the logit link performed equally well in terms of coverage for the majority of simulation scenarios. Some exceptions were observed when the logit link was applied in the presence of long follow-up times, or large % random censoring in combination with small event probability; low coverage in these situations appeared to be driven by the bias observed (Figure 6.1). The coverage under the use of the clog-log link was quite poor for the majority of the simulation scenarios compared to the other three methods (column 4, Figure 6.6).

In the presence of 5 trials per meta-analysis, there were situations where the gold-standard approaches provided a lower coverage than the target of 95%. This is known to occur when the methods are not allowing for uncertainty in estimating heterogeneity[127, 132, 166]. As a consequence, the scenarios that behaved well in terms of coverage for the gold-standard approaches were those where heterogeneity was low or close to 0 (Figure 6.6). Other methods could be used to calculate alternative confidence intervals such as the Knapp-Hartung method, bootstrap confidence intervals, however calculation was conducted using the Wald-type confidence intervals since they are more widely used. A detailed comparison of these alternative methods was presented by Veroniki et al[132].

Lower coverage than the target of 95% was observed when the data were treated as binary using the logit link with 5 trials per meta-analysis especially when long follow-up times, large % random censoring and small event probability, large heterogeneity was present; however, it was close to the coverage obtained from the gold-standard approaches. The clog-log link provided poor coverage especially in the presence of long follow-up times and large % random censoring however not to the extent of the poor coverage observed with 20 trials per meta-analysis; the clog-log link performed slightly better in terms of coverage with 5 trials per meta-analysis, but it was still low (Figure 6.6). On this situation, poor coverage appeared to be driven by bias which becomes more important as the amount of information increases (i.e. standard error decreases).



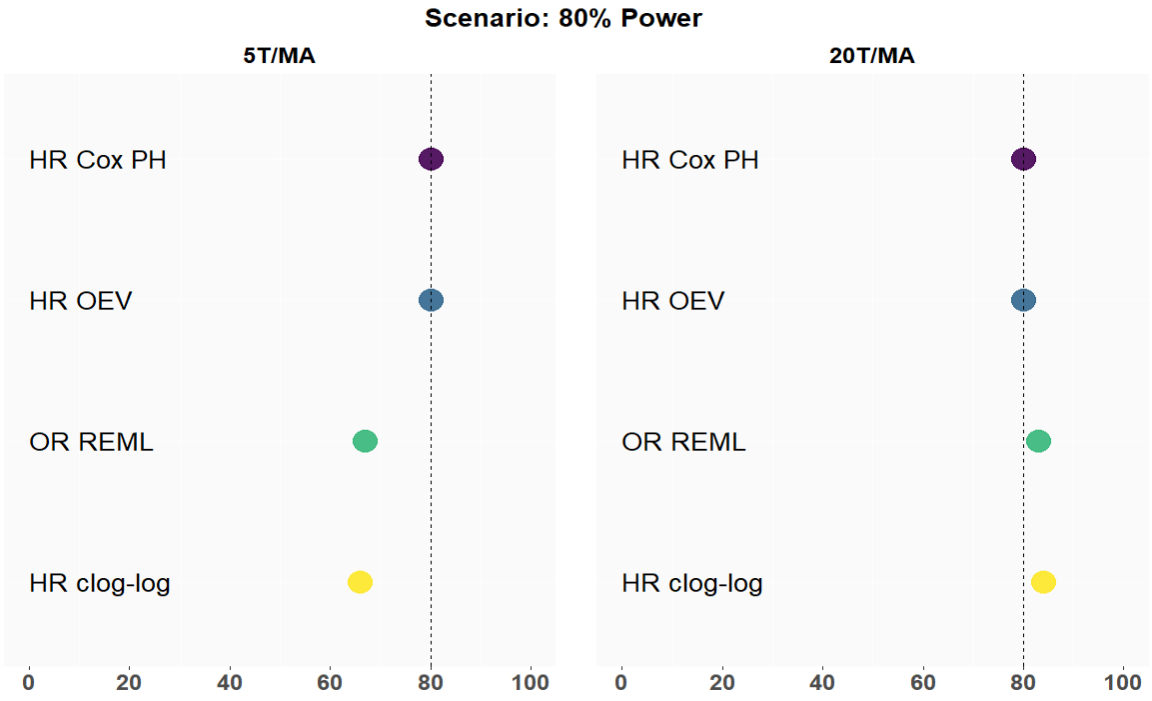
Note: 5T/MA, 20T/MA: 5 or 20 trials per meta-analysis; Each row within a panel is a different data generating mechanism (DGM); Upper and lower rows of panels are DGMs with K = 5 and 20; Columns of panels are different analysis methods; Scenario 0: Large sample size, No effect size, No heterogeneity, Long F-Up; Base Case**: Medium sample size, medium effect size, medium heterogeneity, medium follow up; Other scenarios change from base case; R_cens=Random censoring; P(E)=P(Event)

Figure 6.6: Coverage obtained per simulation scenario across different meta-analysis models.

6.4.6 Power

Two scenarios were designed to provide 80% power under the Cox proportional hazards model as described in 6.3.1. For the scenario involving 5 trials per meta-analysis providing 80% power, analysing the data as binary using the clog-log and the logit link functions caused loss of power (Figure 6.7); the loss of power on these methods was due to the fact that the former method (i.e., clog-log link) was more biased towards the null as presented in Figure 6.1 and the latter method (i.e., logit link) was less precise (Figure 6.3).

On the other hand, I observed a different pattern for the scenario including 20 trials per meta-analysis. Specifically, analysing the data as binary under both link functions improved the power to be slightly higher than 80%. For the clog-log link this was observed since the method provided a much more precise estimate compared to the 5 trials scenario and was slightly less biased; hence this caused better power. For the logit link, the estimate was unbiased and quite precise relative to the Cox model and therefore the power was improved. Power would be meaningless without type I error being controlled, but as observed from Figure 6.6 that this was not a problem here (i.e. Scenario under no effect size).



*Note: 5T/MA, 20T/MA: 5 or 20 trials per meta-analysis; Rows of panels are different analysis methods; One simulation scenario designed specifically with 80% power under the Cox proportional hazards model.

Figure 6.7: Power obtained designing a scenario with 80% power under Cox proportional hazards model.

6.5 Discussion

Using a simulation study and following specific pre-determined data generating mechanisms, I performed a simultaneous comparison of the "gold standard" approaches (Cox and log-rank method) to the approximate methods (using the clog-log or a logit link) using aggregate data to conduct time-to-event meta-analysis. The data generating mechanisms used were representative of situations observed in reality and informed also by previous simulation studies[21, 107-109, 111]; some extreme scenarios were also created for clarification purposes and to ease some results' interpretation. The estimands that were used were different, but researchers sometimes use the log HR and log OR interchangeably to address the same question. The performance of each method was assessed in terms of bias, empirical standard error, model based standard error, mean squared error, coverage and the power each method provided.

The simulation indicated that analysing time-to-event outcomes as binary using the logit link performed well with respect to bias and coverage in many simulation scenarios apart from those where large percentage random censoring (~40%) and long follow-up time (~5 years) was present. However, for the majority of the simulation scenarios the method lacks precision, particularly for small meta-analyses.

On the other hand, analysing the data as binary using the clog-log link consistently produced more bias, low coverage and low power. However, the method provided much more precise estimates compared to the logit link. Therefore, although the clog-log link allowing interpretation on the HR scale was considered a possible alternative to analysing the TTE outcomes as binary, and we observed precise results in the previous empirical analyses presented in Chapters 3-5, based on this simulation I identified that use of this method adversely affected the results. It is worth mentioning though that the method performed well under situations ideally suited to it, which meta-analysts will probably never face in practice (i.e. Scenario 0, Long follow-up time and 0% random censoring). Finally, between-study heterogeneity and study sample size did not affect the levels of bias.

In my simulation study, I was able to identify important factors associated with the results' discordance between the different scales. However, a limitation of this study was that characteristics were varied one by one from a baseline setting

instead of examining every combination of factors simultaneously, additional scenarios might have been missed under which analysing time-to-event data as binary appears acceptable.

To date, there were very limited simulation studies conducted by Tudur-Smith and Williamson[107] and Simmonds et al.[21] examining similar methodologies to those used in this chapter. However, the former authors did not include any comparisons related to the use of the clog-log link function whereas the latter restricted their attention to estimation of the log hazard or log odds ratio and its variance for a single trial.

Tudur-Smith and Williamson[107], compared the stratified log-rank analysis, stratified Cox regression and inverse variance weighted average of estimates. The authors indicated circumstances where the models produced similar estimates of the pooled log HR and its variance (when the underlying treatment effect was close to zero and the degree of heterogeneity across trials was minimal). The stratified log-rank analysis biased the results for larger treatment effects.

Simmonds et al.[21] showed that bias is present when the hazards or the odds are not proportional; this was not the focus of our simulation which did not compare methods in scenarios where the proportionality assumption does not hold. The authors of this paper also stated the potential implications for a meta-analysis setting highlighting specifically the extra complications that are introduced in the presence of heterogeneity included in a random-effects meta-analysis but they did not explicitly explore that. My study added this extra complexity that Simmonds et al[21] recommended in their paper to assess all the factors affecting a time-to-event meta-analysis simultaneously.

6.6 Conclusion

A time-to-event IPD meta-analysis is the gold standard and should be analysed on a HR scale. In the absence of IPD, alternative methodology exists allowing researchers to perform a TTE meta-analysis on a HR scale by extracting suitable information from trial reports and applying a log-rank test. The simulation study indicated that small differences were observed between the gold-standard approaches and therefore there was no reason to recommend the one over the other. The logit link performed well in many simulation scenarios with some exceptions where large percentage random censoring (~40%) and long follow-up time (~5 years) were present; the method though lacked precision in the majority of scenarios. On the other hand, the complementary log-log link was not suitable to analyse the data as binary on a HR scale since a lot of bias was observed, the coverage was low, and the method provided also low power. If a HR estimate cannot be obtained per trial to perform a meta-analysis of TTE data, a meta-analysis using the OR scale (using the logit link) could be conducted but with awareness that this would provide less precise estimates in the analysis. Investigators should avoid performing meta-analyses on the OR scale in the presence of large percentage random censoring (~ 40%) and long follow-up times (~5 years) of the trials included in the meta-analysis.

Parts of this Chapter were presented as an oral presentation at the 43rd conference of International Society of Clinical Biostatistics, 2022.

7. Discussion

7.1 Motivation and Thesis Aims

As outlined in the previous chapters, the overall objective of this thesis was to provide guidance to systematic reviewers and meta-analysts about the implications of analysing TTE outcomes as binary, how the implications vary according to MA characteristics and in which circumstances analysing the outcome as binary may be adequate. The research questions I aimed to answer were as follows:

- What are the implications of analysing TTE outcomes as binary in MA and how do the implications vary according to MA characteristics?
- How are TTE outcomes analysed within the biggest database publishing systematic reviews and MA, the CDSR? Are they analysed as binary or are they analysed as HR, taking into account the full properties of the data?
- Which medical areas within the database analyse the data under which scale?
- What are the assumptions made when different meta-analytic models are applied and what are the advantages and disadvantages of each one of them?
- Is there any other method that could allow us to mitigate the undesirable properties from treating the data as binary?

This chapter summarises the work discussed throughout this thesis. The rest of this chapter is set out as follows: In Section 7.2, I provide a summary of the key findings of each chapter and in Section 7.3, I indicate the strengths and limitations of my research. In Section 7.4, I discuss the generalisability of my results and additional research opportunities (Section 7.5) following this piece of work. I finally provide a conclusion of my thesis in Section 7.6.

7.2 Summary of key findings

In Chapter 2, I carried out a methodology review outlining all guidance that exists for performing a MA of TTE outcomes and also any discussions presented on analysing TTE data as binary in a MA. I used Medline (Ovid version), Scopus and Web of Science for my search and according to prespecified criteria, I identified 75 methodological publications until December 2021.

I categorised the publications into seven categories: Models for aggregate data (11 publications), methods for reconstruction of TTE data (5 publications), models for IPD (16 publications), methods for NMA (12 publications), multivariate MA (7 publications), method comparison via real life conditions and/or simulations (16 publications) and finally papers including discussions, critiques and other suggestions for MA of TTE outcomes (6 publications). Publications could overlap among multiple categories. The methodology review identified the research that exists in the literature to support systematic reviewers and meta-analysts to perform MA of TTE outcomes. It has also described more complex methodologies with regards to different modelling techniques that are primarily aimed at statisticians and not necessarily aimed to be applied by systematic reviewers and meta-analysts. The review identified that most publications in the past were focusing mainly on models for aggregate data, whereas recent publications are focusing mainly on meta-analysis of IPD or NMA. The use of Bayesian techniques in recent years has explored.

The review identified limited publications focusing on the issue of analysing TTE outcomes as binary such Michiels et al[161]. I was able also to extract information from some research publications on the significance of the use of different effect measures. I described various methodologies for MA of TTE data, however, according to past reviews their application to date was still quite limited[45, 46].

Finally, I indicated that further research is needed in order to understand the impact of analysing TTE outcomes as binary rather than using specific methods developed for MA of TTE outcomes, within different MA datasets having various characteristics.

In Chapter 3, I used TTE data from the CDSR (Issue 1, 2008) analysed originally as binary and explored the differences that occur when data are analysed as

binary on an OR scale as opposed to analysing the data using the complementary log-log link where interpretation is conducted on a HR scale.

My analysis showed that there are important reasons associated with discordance among the results, indicating that the correct choice of the method does matter and may affect the interpretation and conclusions drawn from the results. I highlighted that those differences between the scales arise mainly when event probability is high and may occur via differences in between-study heterogeneity or via increased within-study standard error in the OR relative to the HR analyses. There were also situations where there was no clear single factor driving the differences, since there was a combination of reasons affecting the individual study estimates and corresponding weights. All my analyses were conducted both under two- and one-stage random effects models in R.

In Chapter 4, using an additional subset of meta-analysis data from the CDSR ("OEV" data), I re-investigated the impact of analysing TTE outcomes as binary within meta-analysis. I identified the differences that occur when these data are analysed as binary as opposed to analysing the data using the complementary log-log link or using the "O-E" and "V" statistics where interpretation is conducted on a HR scale.

As in Chapter 3, my analysis confirmed that the correct choice of method for a MA of TTE data does matter; high event probability, changes to between and within-study variation appeared to be important factors producing differences in the results in this subset of meta-analyses.

However, in this subset there were more occasions under which there was no clear indication of one single factor driving these differences and a combination of reasons affected the discordance among the results. Therefore, regarding method selection, based on the "OEV" data I identified that a mixed pattern was observed and there was no clear indication of the exact conditions under which the clog-log link outperforms logit link on an OR scale or vice versa. In this subset, my analyses were conducted only under two-stage random effects models using R software.

Summarising the findings from Chapters 3 and 4, I indicated that TTE data should ideally be analysed accounting for their natural properties, as it is possible for

important discrepancies to be observed and conclusions from the MA to be altered. I identified that dichotomising TTE outcomes may be adequate for low event probabilities but not for high event probabilities. In meta-analyses where only binary data are available, the complementary log-log link may be a useful alternative when analysing TTE outcomes as binary, however the exact conditions under which this would be acceptable needed further exploration.

In Chapter 5, using IPD, I investigated whether important properties of TTE data such as percentage total censoring and follow-up times could additionally affect the results obtained from a MA when data are analysed using "gold-standard" approaches (such as Cox proportional hazards model and the log-rank test) as opposed to analysing the data as binary using the clog-log or the logit link where interpretation is conducted on a HR or an OR scale respectively.

Compared to the "gold-standard" methods, my analyses conducted on an OR scale indicated discordancy both in the individual and pooled effect estimates when the event probability was high. Smaller trials provided consistently different individual trial and pooled effect estimates in the OR relative to the HR analyses. The confidence intervals for individual study results were systematically wider in the OR compared to HR analyses since they provided an increased within-study standard error. I also identified a mixed pattern in between-study heterogeneity and I^2 estimates in the OR and HR clog-log analyses. For some TTE outcomes, the between-study heterogeneity estimate obtained from the model in the HR clog-log analysis was not in agreement with the estimates obtained from the gold-standard approaches, although it was still quite low and close to the estimates from other models. This has affected both the individual study weights and the I^2 estimates; however, the individual study estimates in the HR clog-log analysis were closer than those from the OR analysis to the corresponding estimates from the gold-standard approaches.

From my analyses in Chapter 5, I was not able to explain the situations where a model using the complementary log-log link would be a more suitable approach than a model treating TTE as binary in a meta-analysis since a mixed pattern was observed regarding whether or not the results fall in between the "gold-standard" approaches and the binary model with a logit link. I could not explain whether censoring and follow-up time were distinct factors affecting the discordance among the MA estimates since a) high event probability was a strong factor

affecting the results as observed in previous chapters and b) I could not distinguish between random and fixed censoring given the data I had.

Finally, I suggested a comprehensive simulation study that would examine which meta-analysis characteristics affect differences between MA results obtained on the HR or OR scales.

In Chapter 6, using simulation-based datasets, I performed a simultaneous comparison of the “gold standard” approaches (Cox and log-rank method) to the approximate methods (assuming a normal approximation to binomial likelihood with a clog-log link or a binomial likelihood with a logit link) for using aggregate data to conduct TTE MA.

I generated 28 simulation scenarios defined by: number of trials per meta-analysis, trial sample size, log HR, between-study variability, follow-up time, and percentage random and fixed censoring. I compared “gold standard” approaches to analysis on the HR scale (Cox and log-rank method) with analysis as binary using either a logit link on the OR scale or a clog-log link on the HR scale.

The simulation indicated that analysing TTE outcomes as binary using the logit link performed well with respect to bias and coverage in many simulation scenarios apart from those where large percentage random censoring (~40%) and when long follow-up time (~5 years) was present. However, for the majority of the simulation scenarios the method lacked precision particularly for small meta-analyses. On the other hand, analysing the data as binary using the clog-log link consistently produced more bias, low coverage and low power. This method though provided much more precise estimates compared to the logit link.

I concluded that, if a HR estimate cannot be obtained per trial to perform a meta-analysis of TTE data, a meta-analysis using the OR scale (using the logit link) could be conducted but with awareness that this would provide less precise estimates in the analysis. Investigators should avoid performing meta-analyses on the OR scale in the presence of large percentage random censoring (~ 40%) and long follow-up times (~5 years) of the trials included in the meta-analysis.

Chapter & Objectives	Key findings	Limitations	Further work
Chapter 2: Identify the guidance that exists in the literature for MA of TTE outcomes and any discussions raised for analysing these data as binary.	• Identified 75 methodological publications divided into 7 categories. • Many methodologies have been proposed but their application is limited. • No publications discussing the analysis of TTE outcomes as binary.	• It is not a systematic review, although I followed a systematic approach to searching and screening to identify the necessary evidence. • Excluded publications reported in languages other than English.	• Understand how these methodologies perform comparatively when applied to different MA datasets having various characteristics, using effect measures such as the HR and OR.
Chapter 3: Exploring the differences between TTE MA analysed originally as binary on the OR scale in the CDSR with MA results from analyses performed	• High event probability was an important factor associated with discordant effect estimates. Changes to between and within-study variation were mechanisms producing differences in the results. • Combination of reasons affecting the individual study estimates and corresponding weights.	• Not able to distinguish MAs with short follow-up which may have been appropriately analysed as binary. • Results might be different for other TTE outcomes and results might have changed in reviews after 2008.	• Additional research is needed in order to examine whether meta-analysts have improved the way they are performing MA of TTE outcomes after 2008.

using the clog-log link on the HR scale.	• One-stage MA models were also used and demonstrated a similar pattern to the two-stage models for comparisons between different modelling approaches.	• Lack of information on censoring and follow-up times.	
Chapter 4: Comparing a subset of TTE data initially analysed using “O-E” and “V” statistics in the CDSR on the HR scale to results from analysing these data as binary on the OR (via the logit link) or HR (via the clog-log link) scale.	• TTE data should be ideally analysed accounting for their natural properties. • Dichotomising TTE outcomes may be adequate for low event probabilities but not for high event probabilities. • The clog-log link may be a useful alternative when analysing TTE outcomes as binary, however the exact conditions need further exploration.	• The comparison of OR/HR scale in the “OEV” data was slightly different; the number of events and non-events were used for the OR & HR clog-log calculation (as in Chapter 3) and calculated a HR based on “O-E” & “V” statistics. For some cases the two data sets entered by Cochrane reviewers may not completely correspond to each other. • Not able to make comparisons using one-stage models in the “OEV” data due to IPD unavailability.	• The exact conditions under which the clog-log link might be a useful alternative to analysing TTE data as binary on an OR scale need further exploration.

Chapter 5: Comparison between "gold-standard" approaches to analysing the data as binary on the HR (via the clog-log link) or the OR (via the logit link) scale using IPD.	• Confirmed previous findings obtained from the CDSR that the method choice does matter. • Discordancy between OR and HR analyses both in the individual and pooled effect estimates in presence of high event probability. • Smaller trials provided consistently different individual trial and pooled effect estimates in the OR relative to the HR analyses. • Increased within-study standard error in the OR relative to the HR analyses. • Mixed pattern was observed for between-study heterogeneity and I^2 estimates between the OR and HR clog-log analyses. • Careful consideration on the most appropriate method for a TTE MA	• Not able to explain the situations where a model using the clog-log link is more suitable than analysing TTE data as binary in a MA • Could not explain whether censoring and follow-up time were distinct factors affecting the discordance among the MA estimates since: **a)** high event probability was a strong factor affecting the results as observed in previous chapters **b)** I could not distinguish between random and fixed censoring given the data I had.	• A comprehensive simulation study is necessary since real world evidence only cannot explain the situations where a model using the clog-log link is more suitable than analysing TTE data as binary in a MA.

	depending on data availability is necessary.	• RMST method was not used since the scope was not to compare methods in the presence of non-proportional hazards.	
Chapter 6: Comprehensive simulation study examining which meta-analysis characteristics affect differences between results obtained on the HR and OR scales.	• Analysing TTE data as binary using the logit link performed well in many scenarios with some exceptions where large % random censoring (~40%) and long follow-up time (~5 years) were present; the method lacked precision in the majority of scenarios. • Analysing TTE data using the clog-log link was not suitable to analyse the data as binary on a HR scale since a lot of bias was observed, the coverage was low, and the method provided also low power.	• Scenarios were examined by varying characteristics one by one from a baseline setting rather than examining every possible combination of parameters.	• Implications of analysing TTE data as binary in other settings (e.g. NMA, multivariate MA, inclusion of interaction terms).

	• Between-study heterogeneity and study sample size did not affect the levels of bias

Table 7.1: Summary of objectives, key findings, limitations and future work per individual chapter.

7.3 Strengths and limitations

In this thesis, I gave an adequate description of all related guidance for meta-analyses of TTE outcomes. Several methods have been discussed (Chapter 2) and have been applied (Chapters 3-6).

There has been limited research on assessing the impact of analysing these outcomes as binary in a meta-analysis setting. For example, Michiels et al.[161] found that both median survival times and OR methods could result in an important loss of statistical power and under- or overestimation of treatment effects. In the presence of lower event rates, median survival time methods provided more biased results. Although there was limited evidence identified at a MA level, Green and Symons[33] on an individual study level indicated that proportional hazards models provide relatively stable coefficients and decreased SE with increasing follow-up time, which is not the case for logistic models where SEs of the estimates generally increase. These authors also mentioned that the two models produce similar estimates in the presence of rare incidence of a disease and short follow-up time.

My analyses using the CDSR of 2008 was a very large empirical study of the implications of different methods of analysis within real meta-analyses and my analyses using IPD allowed a more thorough exploration of the differences within several real meta-analyses using more detailed study data. They both provided useful information on the potential factors affecting the differences between analysing the data as binary and accounting for their natural properties. These analyses informed my subsequent simulation study which provided the most accurate evidence as the truth was known and gave a more definitive answer about the circumstances under which analysing TTE data as binary could be acceptable. Via my simulation study, I found that analysing TTE data as binary using the clog-log link is not a suitable approach.

The models I used to obtain the results for all my analyses were the most suitable according to the literature and the advantages and disadvantages of each one of them have been discussed extensively in earlier chapters. Finally, even though I focused mainly on outcomes such as overall survival and all-cause mortality in the empirical research, I considered a range of event probabilities and censoring rates in my simulation study and therefore similar findings could be expected for other TTE outcomes.

A limitation associated with the analyses using the CDSR (Chapters 3,4) was that the database provided meta-analyses up to 2008. However, since 2008 it has no longer been possible to export the Cochrane database in the same form (i.e., in the form of an Access database as originally extracted in 2008), and therefore obtaining an updated database requires scraping of HTML files and would involve a lot of additional work.

Additionally, my analyses have focused on TTE meta-analyses where the proportionality assumption holds. I did not examine how the results would differ if other methodology accounting for non-proportional hazards such as RMST or Poisson regression models could affect the results. Little discussion (and no additional method implementation) has also been provided related to the use of Kaplan-Meier plots and its importance in a TTE meta-analysis.

In my simulation study (Chapter 6), I was able to identify important factors associated with the results being discordant between the different scales. However, since I varied characteristics one by one from a baseline setting instead of examining every combination of factors simultaneously, I might have missed additional situations under which analysing TTE data as binary might have been acceptable.

7.4 Related research

As identified in previous chapters, limited research exists to date assessing the impact of analysing TTE outcomes as binary in a meta-analysis. For example, at a single study level, T.V. Perneger[54] proposed the use of the relative log survival and complementary log-log link for binary TTE analyses when the traditional two-by-two table is a fair summary of results and therefore duration of follow-up is the same for all individuals. The author suggested the use of Kaplan-Meier curves in case follow-up time varies among individuals.

Simmonds et al.[21] restricted their attention to estimation of the log hazard or log odds ratio and its variance for a single trial. Simmonds et al.[21] specifically showed that bias is present when the hazards or the odds are not proportional. The authors of this paper also stated the potential implications for a meta-analysis setting highlighting specifically the extra complications that are introduced in the presence of heterogeneity included in a random-effects meta-analysis but they did not explicitly explore that.

On a meta-analysis level, Michiels et al.[65] compared results obtained from MAs when median survival times were used as an alternative to HRs, or ORs of survival rates. Authors found that both median survival times and OR methods could result in an important loss of statistical power and under- or overestimation of treatment effects. In the presence of lower event rates, the median survival time method provided more biased results.

Additionally, Tudur-Smith and Williamson[107], in 2007, compared three methods for fixed-effect IPD MA using TTE outcomes: the stratified log-rank analysis, stratified Cox regression and inverse variance weighted average of estimates. The authors indicated circumstances under which the models could produce similar estimates of the pooled log HR and its variance (when the underlying treatment effect was close to zero and the degree of heterogeneity across trials was minimal).

In relation to the research above my study added this extra complexity that Simmonds et al[21] recommended in their paper to assess all the factors affecting a time-to-event meta-analysis simultaneously and tried to assess whether the use of the clog-log link presented by T.V. Perneger[54] could also be observed at a meta-analysis level.

7.5 Generalisability

The results obtained from the CDSR of 2008 in Chapters 3 and 4 include meta-analyses of clinical trials. The conclusions drawn from these chapters could be generalised to non-Cochrane reviews; projects that are still within the scope of evidence synthesis, beyond the requirements set by Cochrane and different in terms of reporting quality[167]. In comparison to Cochrane reviews, non-Cochrane reviews report usually larger effect sizes with lower precision and provide systematically larger methodological differences that can generate different interpretations of the interventions under question[168]. The improvement of reporting and transparency of non-Cochrane reviews has been discussed multiple times in the past[169, 170].

7.6 Opportunities for further research

A number of possible extensions could be conducted for the present project. My work in Chapter 2 allowed me to obtain an in-depth summary of relevant

published literature for a MA of TTE data, identify any discussions raised for analysing these data as binary and inform the subsequent research presented in later chapters. However, it would be interesting to understand how many of these novel methodologies proposed would perform in different meta-analysis settings since to date most systematic reviewers and meta-analysts are focusing mostly on the more conventional methodology to perform MA of TTE outcomes.

The analyses performed in Chapters 3 and 4 involve meta-analyses from the CDSR up to 2008. Given the fact that methodology for meta-analyses of TTE outcomes has been improved over the last decade, examination on whether systematic reviewers and meta-analysts have improved the way they are performing these analyses would be interesting. There are plans to obtain an updated CDSR database with classifications of outcomes and intervention types, and there could be a possibility to use this to explore whether analysis choices have changed.

Additionally, I considered differences between analysing using the effect sizes of OR and HR solely within a pairwise meta-analysis framework. Other types of meta-analyses exist such as network meta-analysis and multivariate meta-analysis, having their own assumptions that were not considered in this thesis. For example, a NMA allows for simultaneous comparison of multiple interventions by combining the direct and indirect evidence in a network. Direct evidence is obtained from a specific pairwise comparison whereas indirect evidence is derived from studies that do not include that specific comparison. Therefore, an extension of this project could be the implications of analysing TTE outcomes as binary in NMA framework and how this would affect the results between the scales, and how this could affect treatment rankings. Additionally, multivariate meta-analysis allows for simultaneous analysis of multiple outcomes which allows us to incorporate the correlation that might be present across them and also facilitates more studies to contribute towards each outcome and treatment comparison[171]. Another extension of the current project could be the implications of analysing multiple correlated TTE outcomes as binary in multivariate meta-analysis.

7.7 Conclusion & Recommendations

The findings presented on this thesis provide an adequate and comprehensive exploration of the implications of analysing TTE outcomes as binary in meta-analysis. No research exists examining the specific comparisons I performed in this thesis. However, there is a limited body of work that did look at related comparisons such as combining published survival curves using effect measures of HR and RR or weighted log RR[102], or using median survival times as an alternative to HRs and ORs, stratified Cox models and ORs[161]. Those comparisons have been discussed extensively in Chapter 2, Section 2.4.6.

A time-to-event IPD meta-analysis is the gold standard and should be analysed on a HR scale since it allows systematic reviewers and meta-analysts to overcome limitations of already published data, avoids data quality issues and usually includes more mature data. However, if the data available in publications (which can also include data obtained from Cox proportional hazards model) are sufficient then an aggregate data approach would also be appropriate and less time consuming. In the absence of IPD, alternative methodology exists allowing researchers to perform a TTE meta-analysis on a HR scale either by extracting suitable information from trial reports and applying a log-rank test or by using the Kaplan-Meier plots which could be used to collect an approximate HR rather than a direct one; this latter method would also be preferable to applying a TTE MA on an OR scale.

The complementary log-log link is not a suitable approach to analyse the data as binary on a HR scale a lot of bias is observed, the coverage is low, and the method provides also low power. If a HR estimate cannot be obtained per trial to perform a meta-analysis of TTE data, a meta-analysis using the OR scale (using the logit link) could be conducted but with awareness that this would provide less precise estimates in the analysis.

It is advised that systematic reviewers and meta-analysts should think carefully about the circumstances before analysing time-to-event data as binary because this may produce different conclusions than the correct time-to-event analysis. Investigators should avoid performing meta-analyses on the OR scale in the presence of high event probability, large percentage random censoring (~40%) and therefore longer follow-up times (~5 years) assuming of large event rates (>70%) of the trials included in the meta-analysis. Investigators should also be

cautious about performing meta-analyses on the OR scale if events are not likely to occur early on time, events are not rare, and the lengths of follow-up are not similar between the patients. It is worth reminding researchers also that interpretations on an OR scale should be interpreted in the context of a particular time point for a TTE outcome. Finally, on occasions where some MAs are providing HR estimates and standard errors or “O-E” and “V” statistics and others OR estimates, investigators should either consider excluding the studies involving ORs from the meta-analysis or transforming the HR estimates into OR estimates and performing a MA on an OR scale. The decision should take into account whether the events are occurring earlier on, events are rare and lengths of follow-up are similar between the patients.

References

1. Altman DG, Bland JM. Time to event (survival) data. BMJ. 1998;317(7156):468-9.
2. Collett D. Modelling survival data in medical research: Chapman and Hall/CRC; 2015.
3. Cox DR. Regression models and life-tables. Journal of the Royal Statistical Society: Series B (Methodological). 1972;34(2):187-202.
4. Bewick V, Cheek L, Ball J. Statistics review 12: survival analysis. Critical care. 2004;8(5):389.
5. Bland JM, Altman DG. Survival probabilities (the Kaplan-Meier method). BMJ. 1998;317(7172):1572-80.
6. Armitage P, Berry G, Matthews JNS. Statistical methods in medical research: John Wiley & Sons; 2008.
7. Rodrıguez G. Parametric survival models. Rapport technique, Princeton: Princeton University. 2010.
8. Lagakos SW. General right censoring and its impact on the analysis of survival data. Biometrics. 1979:139-56.
9. Ranganathan P, Pramesh C. Censoring in survival analysis: potential for bias. Perspect Clin Res. 2012;3(1):40.
10. Shih WJ. Problems in dealing with missing data and informative censoring in clinical trials. Current controlled trials in cardiovascular medicine. 2002;3(1):4.
11. Horton R. Encyclopaedic companion to medical statistics. 2nd Edition ed: John Wiley & Sons; 2011.
12. Haidich A-B. Meta-analysis in medical research. Hippokratia. 2010;14(Suppl 1):29.
13. Cochrane L. Cochrane Library 2020 [Available from: https://www.cochranelibrary.com/.
14. Sutton AJ, Abrams KR, Jones DR, Jones DR, Sheldon TA, Song F. Methods for meta-analysis in medical research: Wiley Chichester; 2000.

15. Fleiss J. Review papers: The statistical basis of meta-analysis. Statistical methods in medical research. 1993;2(2):121-45.
16. Higgins JP, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, et al. Cochrane handbook for systematic reviews of interventions: John Wiley & Sons; 2019.
17. Cochrane L. Cochrane Library of Systematic Reviews 2021 [cited 2021 02/11]. Available from: https://www.cochranelibrary.com/cdsr/reviews.
18. Higgins PTJ, Green S. Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0: The Cochrane Collaboration; 2011.
19. Yusuf S, Peto R, Lewis J, Collins R, Sleight P. Beta blockade during and after myocardial infarction: an overview of the randomized trials. Progress in cardiovascular diseases. 1985;27(5):335-71.
20. Greenland S, Salvan A. Bias in the one-step method for pooling study results. Statistics in medicine. 1990;9(3):247-52.
21. Simmonds MC, Tierney J, Bowden J, Higgins JP. Meta-analysis of time-to-event data: a comparison of two-stage methods. Research synthesis methods. 2011;2(3):139-49.
22. Parmar MK, Torri V, Stewart L. Extracting summary statistics to perform meta-analyses of the published literature for survival endpoints. Statistics in medicine. 1998;17(24):2815-34.
23. Williamson PR, Smith CT, Hutton JL, Marson AG. Aggregate data meta-analysis with time-to-event outcomes. Statistics in medicine. 2002;21(22):3337-51.
24. Clarke MJ, Stewart LA. Systematic Reviews: Obtaining data from randomised controlled trials: how much do we need for reliable and informative meta-analyses? BMJ. 1994;309(6960):1007-10.
25. Hunink MG, Wong JB. Meta-analysis of failure-time data with adjustment for covariates. Medical Decision Making. 1994;14(1):59-70.
26. StataCorp L. Stata data analysis and statistical Software. Special Edition Release. 2007;10:733.
27. Wicklin R. Statistical programming with SAS/IML software: SAS Institute; 2010.
28. Crawley MJ. The R book: John Wiley & Sons; 2012.
29. Van Rossum G, Drake Jr FL. Python tutorial: Centrum voor Wiskunde en Informatica Amsterdam; 1995.

30. Lunn DJ, Thomas A, Best N, Spiegelhalter D. WinBUGS-a Bayesian modelling framework: concepts, structure, and extensibility. Statistics and computing. 2000;10(4):325-37.
31. RevMan R. Review Manager (RevMan)(Version 5.3). The Nordic Cochrane Centre, The Cochrane Collaboration Copenhagen, Denmark; 2014.
32. Stevens A, Abrams K, Brazier J, Fitzpatrick R, Lilford R. The advanced handbook of methods in evidence based healthcare: Sage; 2001.
33. Green MS, Symons MJ. A comparison of the logistic risk function and the proportional hazards model in prospective epidemiologic studies. Journal of chronic diseases. 1983;36(10):715-23.
34. Peduzzi P, Holford T, Detre K, Chan Y-K. Comparison of the logistic and Cox regression models when outcome is determined in all patients after a fixed period of time. Journal of chronic diseases. 1987;40(8):761-7.
35. Annesi I, Moreau T, Lellouch J. Efficiency of the logistic regression and Cox proportional hazards models in longitudinal studies. Statistics in Medicine. 1989;8(12):1515-21.
36. Cuzick J. The efficiency of the proportions test and the logrank test for censored survival data. Biometrics. 1982:1033-9.
37. Ingram DD, Kleinman JC. Empirical comparisons of proportional hazards and logistic regression models. Statistics in medicine. 1989;8(5):525-38.
38. Doksum KA, Gasko M. On a correspondence between models in binary regression analysis and in survival analysis. International Statistical Review/Revue Internationale de Statistique. 1990:243-52.
39. Callas PW, Pastides H, Hosmer DW. Empirical comparisons of proportional hazards, poisson, and logistic regression modeling of occupational cohort data. American journal of industrial medicine. 1998;33(1):33-47.
40. Symons M, Moore D. Hazard rate ratio and prospective epidemiological studies. Journal of clinical epidemiology. 2002;55(9):893-9.
41. Stare J, Maucort-Boulch D. Odds ratio, hazard ratio and relative risk. Metodoloski zvezki. 2016;13(1):59.
42. Tierney J, Rydzewska L. Improving the quality of the analysis of time-to-event outcomes in Cochrane reviews. [Unpublished]. In press 2008.
43. Parmar MK, Torri V, Stewart L. Extracting summary statistics to perform meta-analyses of the published literature for survival endpoints. Statistics in Medicine. 1998;17(24):2815-34.

44. Williamson PR, Smith CT, Hutton JL, Marson AG. Aggregate data meta-analysis with time-to-event outcomes. Statistics in Medicine. 2002;21(22):3337-51.

45. Otwombe KN, Petzold M, Martinson N, Chirwa T. A review of the study designs and statistical methods used in the determination of predictors of all-cause mortality in HIV-infected cohorts: 2002-2011. PLoS ONE [Electronic Resource]. 2014;9(2):e87356.

46. Batson S, Greenall G, Hudson P. Review of the Reporting of Survival Analyses within Randomised Controlled Trials and the Implications for Meta-Analysis. PLoS ONE [Electronic Resource]. 2016;11(5):e0154870.

47. Tierney JF, Stewart LA, Ghersi D, Burdett S, Sydes MR. Practical methods for incorporating summary time-to-event data into meta-analysis. Trials [Electronic Resource]. 2007;8:16.

48. Whitehead A, Whitehead J. A general parametric approach to the meta-analysis of randomized clinical trials. Statistics in Medicine. 1991;10(11):1665-77.

49. Mantel N, Haenszel W. Statistical aspects of the analysis of data from retrospective studies of disease. Journal of the national cancer institute. 1959;22(4):719-48.

50. Vale CL, Tierney JF, Stewart LA. Effects of adjusting for censoring on meta-analyses of time-to-event outcomes. International Journal of Epidemiology. 2002;31(1):107-11.

51. Moodie PF, Nelson NA, Koch GG. A non-parametric procedure for evaluating treatment effect in the meta-analysis of survival data. Statistics in Medicine. 2004;23(7):1075-93.

52. Yuan X, Anderson SJ. Meta-analysis methodology for combining treatment effects from Cox proportional hazard models with different covariate adjustments. Biometrical Journal. 2010;52(4):519-37.

53. Combescure C, Courvoisier DS, Haller G, Perneger TV. Meta-analysis of binary outcomes from two-by-two tables when the length of follow-up varies and hazards are proportional. Statistical Methods in Medical Research. 2011;20(5):531-40.

54. Perneger TV. Estimating the relative hazard by the ratio of logarithms of event-free proportions. Contemporary clinical trials. 2008;29(5):762-6.

55. Combescure C, Courvoisier DS, Haller G, Perneger TV. Meta-analysis of two-arm studies: Modeling the intervention effect from survival probabilities. Statistical Methods in Medical Research. 2012;25(2):857-71.
56. Bonofiglio F, Beyersmann J, Schumacher M, Koller M, Schwarzer G. Meta-analysis for aggregated survival data with competing risks: a parametric approach using cumulative incidence functions. Research Synthesis Methods. 2016;7(3):282-93.
57. Holzhauer B. Meta-analysis of aggregate data on medical events. Statistics in Medicine. 2016;36(5):723-37.
58. Irvine AF, Waise S, Green EW, Stuart B. A non-linear optimisation method to extract summary statistics from Kaplan-Meier survival plots using the published P value. BMC medical research methodology. 2020;20(1):1-17.
59. Messori A, Trippoli S, Vaiani M, Cattel F. Survival meta-analysis of individual patient data and survival meta-analysis of published (aggregate) data: Is there an intermediate approach between these two opposite options? Clinical Drug Investigation. 2000;20(5):309-16.
60. Rubin DB. An Alternative to Pooling Kaplan-Meier Curves in Time-to-Event Meta-Analysis. International Journal of Biostatistics. 2011;7(1).
61. Xie J, Liu C. Adjusted Kaplan–Meier estimator and log-rank test with inverse probability of treatment weighting for survival data. Statistics in medicine. 2005;24(20):3089-110.
62. Guyot P, Ades AE, Ouwens M, Welton N. Enhanced secondary analysis of survival data: Reconstructing the data from published Kaplan-Meier survival curves. BMC Medical Research Methodology. 2012;12.
63. Ghanbari S, Zare N, Shayan Z. A practival method based on functional data analysis and signle exponential smoothing to combine survival curves in Meta-analysis. A simulation study. Advances and Applications in Statistics. 2018;53(2):179-97.
64. Royston P, Parmar MK. Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. Statistics in medicine. 2002;21(15):2175-97.
65. Michiels S, Baujat B, Mahe C, Sargent DJ, Pignon JP. Random effects survival models gave a better understanding of heterogeneity in individual patient data meta-analyses. Journal of Clinical Epidemiology. 2005;58(3):238-45.

66. Tudur-Smith C, Williamson PR, Marson AG. Investigating heterogeneity in an individual patient data meta-analysis of time to event outcomes. Statistics in Medicine. 2005;24(9):1307-19.
67. Massonnet G, Janssen P, Burzykowski T. Fitting conditional survival models to meta-analytic data by using a transformation toward mixed-effects models. Biometrics. 2008;64(3):834-42.
68. 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. Statistics in Medicine. 2008;27(11):1894-910.
69. Vaida F, Xu R. Proportional hazards model with random effects. Statistics in medicine. 2000;19(24):3309-24.
70. Thompson S, Kaptoge S, White I, Wood A, Perry P, Danesh J, et al. Statistical methods for the time-to-event analysis of individual participant data from multiple epidemiological studies. International Journal of Epidemiology. 2010;39(5):1345-59.
71. Siannis F, Barrett JK, Farewell VT, Tierney JF. One-stage parametric meta-analysis of time-to-event outcomes. Statistics in Medicine. 2010;29(29):3030-45.
72. Barrett JK, Farewell VT, Siannis F, Tierney J, Higgins JP. Two-stage meta-analysis of survival data from individual participants using percentile ratios. Statistics in Medicine. 2012;31(30):4296-308.
73. 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 Medical Research Methodology. 2012;12:34.
74. Simmonds MC, Higgins JP, Stewart LA. Random-effects meta-analysis of time-to-event data using the expectation-maximisation algorithm and shrinkage estimators. Research Synthesis Methods. 2013;4(2):144-55.
75. 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. Statistics in Medicine. 2014;33(22):3844-58.
76. Rondeau V, Pignon JP, Michiels S. A joint model for the dependence between clustered times to tumour progression and deaths: A meta-analysis of

chemotherapy in head and neck cancer. Statistical Methods in Medical Research. 2015;24(6):711-29.

77. Wang XV, Cole B, Bonetti M, Gelber RD. Meta-STEPP: subpopulation treatment effect pattern plot for individual patient data meta-analysis. Statistics in Medicine. 2016;35(21):3704-16.

78. Wang XV, Cole B, Bonetti M, Gelber RD. Meta-STEPP with random effects. Research Synthesis Methods. 2018;9(2):312-7.

79. de Jong VM, Moons KG, Riley RD, Tudur Smith C, Marson AG, Eijkemans MJ, et al. Individual participant data meta-analysis of intervention studies with time-to-event outcomes: A review of the methodology and an applied example. Research synthesis methods. 2020;11(2):148-68.

80. Welton NJ, Willis SR, Ades AE. Synthesis of survival and disease progression outcomes for health technology assessment of cancer therapies. Research Synthesis Methods. 2010;1(3-4):239-57.

81. Woods BS, Hawkins N, Scott DA. Network meta-analysis on the log-hazard scale, combining count and hazard ratio statistics accounting for multi-arm trials: A tutorial. BMC Medical Research Methodology. 2010;10.

82. Ouwens MJ, Philips Z, Jansen JP. Network meta-analysis of parametric survival curves. Research synthesis methods. 2010;1(3-4):258-71.

83. Arends LR, Hunink MGM, Stijnen T. Meta-analysis of summary survival curve data. Statistics in Medicine. 2008;27(22):4381-96.

84. Jansen JP. Network meta-analysis of survival data with fractional polynomials. BMC Medical Research Methodology. 2011;11(1):61.

85. Jansen JP, Cope S. Meta-regression models to address heterogeneity and inconsistency in network meta-analysis of survival outcomes. BMC Medical Research Methodology. 2012;12:152.

86. Saramago P, Chuang L-H, Soares MO. Network meta-analysis of (individual patient) time to event data alongside (aggregate) count data. BMC Medical Research Methodology. 2014;14(1):105.

87. Soares MO, Dumville JC, Ades A, Welton NJ. Treatment comparisons for decision making: facing the problems of sparse and few data. Journal of the Royal Statistical Society: Series A (Statistics in Society). 2014;177(1):259-79.

88. Freeman SC, Carpenter JR. Bayesian one-step IPD network meta-analysis of time-to-event data using Royston-Parmar models. Research Synthesis Methods. 2017;8(4):451-64.

89. Watkins C, Bennett I. A simple method for combining binomial counts or proportions with hazard ratios for evidence synthesis of time-to-event data. Research Synthesis Methods. 2018;9(3):352-60.
90. Cope S, Chan K, Jansen JP. Multivariate network meta-analysis of survival function parameters. Research synthesis methods. 2020;11(3):443-56.
91. Wiksten A, Hawkins N, Piepho H-P, Gsteiger S. Nonproportional Hazards in Network Meta-Analysis: Efficient Strategies for Model Building and Analysis. Value in Health. 2020;23(7):918-27.
92. Ollier E, Blanchard P, Teuff GL, Michiels S. Penalized Poisson model for network meta-analysis of individual patient time-to-event data. arXiv preprint arXiv:210300069. 2021.
93. Tang X, Trinquart L. Bayesian multivariate network meta-analysis model for the difference in restricted mean survival times. Statistics in medicine. 2021.
94. Weir IR, Tian L, Trinquart L. Multivariate meta-analysis model for the difference in restricted mean survival times. Biostatistics. 2019;22(1):82-96.
95. Daly CH, Maconachie R, Ades A, Welton NJ. A non-parametric approach for jointly combining evidence on progression free and overall survival time in network meta-analysis. Research Synthesis Methods. 2021.
96. Wei Y, Royston P, Tierney JF, Parmar MK. Meta-analysis of time-to-event outcomes from randomized trials using restricted mean survival time: application to individual participant data. Statistics in Medicine. 2015;34(21):2881-98.
97. Dear KB. Iterative generalized least squares for meta-analysis of survival data at multiple times. Biometrics. 1994;50(4):989-1002.
98. Fiocco M, Putter H, Van Houwelingen J. A new serially correlated gamma-frailty process for longitudinal count data. Biostatistics. 2009;10(2):245-57.
99. Fiocco M, Putter H, Van Houwelingen J. Meta-analysis of pairs of survival curves under heterogeneity: a Poisson correlated gamma-frailty approach. Statistics in medicine. 2009;28(30):3782-97.
100. Jackson D, Rollins K, Coughlin P. A multivariate model for the meta-analysis of study level survival data at multiple times. Research Synthesis Methods. 2014;5(3):264-72.
101. Riley RD, Price MJ, Jackson D, Wardle M, Gueyffier F, Wang J, et al. Multivariate meta-analysis using individual participant data. Research Synthesis Methods. 2015;6(2):157-74.

102. Earle CC, Ba'Pham, Wells GA. An assessment of methods to combine published survival curves. Medical Decision Making. 2000;20(1):104-11.
103. Duchateau L, Pignon JP, Bijnens L, Bertin S, Bourhis J, Sylvester R. Individual patient-versus literature-based meta-analysis of survival data: time to event and event rate at a particular time can make a difference, an example based on head and neck cancer. Controlled Clinical Trials. 2001;22(5):538-47.
104. Tudur C, Williamson PR, Khan S, Best LY. The value of the aggregate data approach in meta-analysis with time-to-event outcomes. Journal of the Royal Statistical Society Series a-Statistics in Society. 2001;164:357-70.
105. 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. Journal of Evaluation in Clinical Practice. 2005;11(5):468-78.
106. Lambert PC, Sutton AJ, Abrams KR, Jones DR. A comparison of summary patient-level covariates in meta-regression with individual patient data meta-analysis. Journal of clinical epidemiology. 2002;55(1):86-94.
107. Tudur-Smith C, Williamson PR. A comparison of methods for fixed effects meta-analysis of individual patient data with time to event outcomes. Clinical Trials. 2007;4(6):621-30.
108. Katsahian S, Latouche A, Mary JY, Chevret S, Porcher R. Practical methodology of meta-analysis of individual patient data using a survival outcome. Contemporary Clinical Trials. 2008;29(2):220-30.
109. Hirooka T, Hamada C, Yoshimura I. A note on estimating treatment effect for time-to-event data in a literature-based meta-analysis. Methods of Information in Medicine. 2009;48(2):104-12.
110. Fisher D, Copas A, Tierney J, Parmar M. A critical review of methods for the assessment of patient-level interactions in individual participant data meta-analysis of randomized trials, and guidance for practitioners. Journal of clinical epidemiology. 2011;64(9):949-67.
111. 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. Research Synthesis Methods. 2011;2(3):150-62.

112. Simmonds MC, Tierney J, Bowden J, Higgins JP. Meta-analysis of time-to-event data: a comparison of two-stage methods. Research Synthesis Methods. 2011;2(3):139-49.
113. Fiocco M, Stijnen T, Putter H. Meta-analysis of time-to-event outcomes using a hazard-based approach: Comparison with other models, robustness and meta-regression. Computational Statistics & Data Analysis. 2012;56(5):1028-37.
114. Bennett MM, Crowe BJ, Price KL, Stamey JD, Seaman JW, Jr. Comparison of Bayesian and frequentist meta-analytical approaches for analyzing time to event data. Journal of Biopharmaceutical Statistics. 2013;23(1):129-45.
115. Heinze G, Schemper M. A solution to the problem of monotone likelihood in Cox regression. Biometrics. 2001;57(1):114-9.
116. 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 Medical Research Methodology. 2016;16(1).
117. van Beekhuizen S, Ouwens MJ, Postma MJ, Heeg B. Network Meta-analyses on Survival data: A comparison and guidance for different methodologies. Value in Health. 2018;21:S397-S8.
118. Abel UR, Edler L. A pitfall in the meta-analysis of hazard ratios. Controlled clinical trials. 1988;9(2):149-51.
119. Keene ON. Alternatives to the hazard ratio in summarizing efficacy in time-to-event studies: an example from influenza trials. Statistics in Medicine. 2002;21(23):3687-700.
120. Bennett DA. Review of analytical methods for prospective cohort studies using time to event data: single studies and implications for meta-analysis. Statistical Methods in Medical Research. 2003;12(4):297-319.
121. Simmonds MC, Higgins JP, Stewart LA, Tierney JF, Clarke MJ, Thompson SG. Meta-analysis of individual patient data from randomized trials: a review of methods used in practice. Clinical Trials. 2005;2(3):209-17.
122. Cope S, Jansen JP. Quantitative summaries of treatment effect estimates obtained with network meta-analysis of survival curves to inform decision-making. BMC Medical Research Methodology. 2013;13:147.
123. Buyse M, Ryan LM. Issues of efficiency in combining proportions of deaths from several clinical trials. Statistics in medicine. 1987;6(5):565-76.

124. Puljak L. Methodological studies evaluating evidence are not systematic reviews. Journal of clinical epidemiology. 2019;110:98.
125. Singer JD, Willett JB, Willett JB. Applied longitudinal data analysis: Modeling change and event occurrence: Oxford university press; 2003.
126. Nevitt S. Data sharing and transparency: the impact on evidence synthesis: University of Liverpool; 2017.
127. Smith CT, Williamson PR, Marson AG. Investigating heterogeneity in an individual patient data meta-analysis of time to event outcomes. Statistics in Medicine. 2005;24(9):1307-19.
128. Hedeker D, Siddiqui O, Hu FB. Random-effects regression analysis of correlated grouped-time survival data. Statistical Methods in Medical Research. 2000;9(2):161-79.
129. Davey J, Turner RM, Clarke MJ, Higgins JPT. Characteristics of meta-analyses and their component studies in the Cochrane Database of Systematic Reviews: a cross-sectional, descriptive analysis. BMC Medical Research Methodology. 2011;11(1):160.
130. 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. Statistics in medicine. 2018;37(7):1059-85.
131. J. Sweeting M, J. Sutton A, C. Lambert P. What to add to nothing? Use and avoidance of continuity corrections in meta-analysis of sparse data. Statistics in medicine. 2004;23(9):1351-75.
132. Veroniki AA, Jackson D, Viechtbauer W, Bender R, Bowden J, Knapp G, et al. Methods to estimate the between-study variance and its uncertainty in meta-analysis. Research synthesis methods. 2016;7(1):55-79.
133. Simmonds MC, Higgins JP. A general framework for the use of logistic regression models in meta-analysis. Statistical methods in medical research. 2016;25(6):2858-77.
134. Rhodes KM, Turner RM, Higgins JP. Empirical evidence about inconsistency among studies in a pair-wise meta-analysis. Research synthesis methods. 2016;7(4):346-70.
135. Bland JM, Altman D. Statistical methods for assessing agreement between two methods of clinical measurement. The lancet. 1986;327(8476):307-10.

136. Burke DL, Ensor J, Riley RD. Meta-analysis using individual participant data: one-stage and two-stage approaches, and why they may differ. Statistics in medicine. 2017;36(5):855-75.
137. Holzhauer B. Meta-analysis of aggregate data on medical events. Statistics in medicine. 2017;36(5):723-37.
138. Higgins JPT, Green S. Cochrane Handbook for Systematic Reviews of Interventions The Cochrane Collaboration; 2011. Available from: www.handbook.cochrane.org.
139. Bland JM, Altman DG. The logrank test. BMJ. 2004;328(7447):1073.
140. Davey J, Turner RM, Clarke MJ, Higgins JP. Characteristics of meta-analyses and their component studies in the Cochrane Database of Systematic Reviews: a cross-sectional, descriptive analysis. BMC medical research methodology. 2011;11(1):160.
141. Stewart LA, Tierney JF, Clarke M. Reviews of individual patient data. Cochrane Handbook for Systematic Reviews of Interventions: Cochrane Book Series. 2008:547-58.
142. Kontopantelis E. A comparison of one-stage vs two-stage individual patient data meta-analysis methods: A simulation study. Research synthesis methods. 2018;9(3):417-30.
143. Salika T, Turner RM, Fisher D, Tierney JF, White IR. Implications of analysing time-to-event outcomes as binary in meta-analysis: empirical evidence from the Cochrane Database of Systematic Reviews. BMC medical research methodology. 2022;22(1):1-14.
144. Vale C, Collaboration ABCM-a. Neoadjuvant chemotherapy in invasive bladder cancer: a systematic review and meta-analysis. The Lancet. 2003;361(9373):1927-34.
145. Martinez-Pineiro JA, Martin MG, Arocena F, Flores N, Roncero CR, Portillo JA, et al. bladder cancer: neoadjuvant cisplatin chemotherapy before radical cystectomy in invasive transitional cell carcinoma of the bladder: a prospective randomized phase III study. The Journal of urology. 1995;153(3S):964-73.
146. Wallace D, Raghavan D, Kelly K, Sandeman T, Conn I, Teriana N, et al. Neo-adjuvant (pre-emptive) cisplatin therapy in invasive transitional cell carcinoma of the bladder. British journal of urology. 1991;67(6):608-15.
147. Sherif A, Holmberg L, Rintala E, Mestad O, Nilsson J, Nilsson S, et al. Neoadjuvant cisplatinum based combination chemotherapy in patients with

invasive bladder cancer: a combined analysis of two Nordic studies. European urology. 2004;45(3):297-303.

148. Finnbladder NBCSG, de Tratamiento Oncologico CUE, Group EG-U, Group ABCS, Group NCIoCCT. Neoadjuvant cisplatin, methotrexate, and vinblastine chemotherapy for muscle-invasive bladder cancer: a randomised controlled trial. The Lancet. 1999;354(9178):533-40.

149. Sengeløv L, Maase HVD, Lundbeck F, Barlebo H, Colstrup H, Engelholm SA, et al. Neoadjuvant chemotherapy with cisplatin and methotrexate in patients with muscle-invasive bladder tumours. Acta Oncologica. 2002;41(5):447-56.

150. Grossman HB, Natale RB, Tangen CM, Speights V, Vogelzang NJ, Trump DL, et al. Neoadjuvant chemotherapy plus cystectomy compared with cystectomy alone for locally advanced bladder cancer. New England Journal of Medicine. 2003;349(9):859-66.

151. Harrington DP, Fleming TR. A class of rank test procedures for censored survival data. Biometrika. 1982;69(3):553-66.

152. Malmstrom P-U, Rintala E, Wahlqvist R, Hellstrom P, Hellsten S, Hannisdal E. Five-year followup of a prospective trial of radical cystectomy and neoadjuvant chemotherapy: Nordic Cystectomy Trial 1. The Journal of urology. 1996;155(6):1903-6.

153. Sherif A, Rintala E, Mestad O, Nilsson J, Holmberg L, Nilsson S, et al. Neoadjuvant cisplatin-methotrexate chemotherapy for invasive bladder cancer-Nordic cystectomy trial 2. Scandinavian journal of urology and nephrology. 2002;36(6):419-25.

154. Natale R, editor SWOG 8710 (INT-0080): Randomized phase III trial of neoadjuvant MVAC+ cystectomy versus cytectomy alone in patients with locally advanced bladder cancer. Proc ASCO; 2001.

155. Smith CT, Williamson PR, Marson AG. Investigating heterogeneity in an individual patient data meta-analysis of time to event outcomes. Statistics in medicine. 2005;24(9):1307-19.

156. 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 Medical Research Methodology. 2012;12(1):34.

157. Wei Y, Royston P, Tierney JF, Parmar MK. Meta-analysis of time-to-event outcomes from randomized trials using restricted mean survival time: application to individual participant data. Statistics in medicine. 2015;34(21):2881-98.

158. Siannis F, Barrett J, Farewell V, Tierney J. One-stage parametric meta-analysis of time-to-event outcomes. Statistics in medicine. 2010;29(29):3030-45.
159. 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. Research synthesis methods. 2011;2(3):150-62.
160. Grambsch PM, Therneau TM. Proportional hazards tests and diagnostics based on weighted residuals. Biometrika. 1994;81(3):515-26.
161. Michiels S, Piedbois P, Burdett S, Syz N, Stewart L, Pignon JP. Meta-analysis when only the median survival times are known: a comparison with individual patient data results. International Journal of Technology Assessment in Health Care. 2005;21(1):119-25.
162. Carroll KJ. On the use and utility of the Weibull model in the analysis of survival data. Controlled clinical trials. 2003;24(6):682-701.
163. Bender R, Augustin T, Blettner M. Generating survival times to simulate Cox proportional hazards models. Statistics in medicine. 2005;24(11):1713-23.
164. Burton A, Altman DG, Royston P, Holder RL. The design of simulation studies in medical statistics. Statistics in medicine. 2006;25(24):4279-92.
165. Morris TP, White IR, Crowther MJ. Using simulation studies to evaluate statistical methods. Statistics in medicine. 2019;38(11):2074-102.
166. Ioannidis JP, Patsopoulos NA, Evangelou E. Uncertainty in heterogeneity estimates in meta-analyses. Bmj. 2007;335(7626):914-6.
167. Cochrane. Non-Cochrane Reviews 2022 [28/05/2022]. Available from: https://work.cochrane.org/non-cochrane-projects.
168. Useem J, Brennan A, LaValley M, Vickery M, Ameli O, Reinen N, et al. Systematic differences between Cochrane and non-Cochrane meta-analyses on the same topic: a matched pair analysis. PloS one. 2015;10(12):e0144980.
169. Koensgen N, Rombey T, Allers K, Mathes T, Hoffmann F, Pieper D. Comparison of non-Cochrane systematic reviews and their published protocols: differences occurred frequently but were seldom explained. Journal of clinical epidemiology. 2019;110:34-41.
170. Tricco AC, Tetzlaff J, Brehaut J, Moher D. Non-Cochrane vs. Cochrane reviews were twice as likely to have positive conclusion statements: cross-sectional study. Journal of clinical epidemiology. 2009;62(4):380-6. e1.

171. Jackson D, Riley R, White IR. Multivariate meta-analysis: potential and promise. Statistics in medicine. 2011;30(20):2481-98.
172. Higgins JP, Thompson SG. Quantifying heterogeneity in a meta-analysis. Statistics in medicine. 2002;21(11):1539-58.

Appendices

A - Articles collected for Methodology Review

I identified 75 articles via electronic and hand searching describing the methodology for MA of TTE outcomes. I present the papers included in our review in chronological order in the following table.

Author	Journal	Title	Year
U. R. Abel and L. Edler	Controlled clinical trials	A pitfall in the meta-analysis of hazard ratios	1988
A. Whitehead and J. Whitehead	Statistics in Medicine	A general parametric approach to the meta-analysis of randomized clinical trials	1991
K. B. Dear	Biometrics	Iterative generalized least squares for meta-analysis of survival data at multiple times	1994
M. G. Hunink and J. B. Wong	Medical Decision Making	Meta-analysis of failure-time data with adjustment for covariates	1994
M. K. Parmar, V. Torri and L. Stewart	Statistics in Medicine	Extracting summary statistics to perform meta-analyses of the published literature for survival endpoints	1998
A. Messori, S. Trippoli, M. Vaiani and F. Cattel	Clinical Drug Investigation	Survival meta-analysis of individual patient data and survival meta-analysis of published (aggregate) data: Is	2000

		there an intermediate approach between these two opposite options?	
C. C. Earle, Ba'Pham and G. A. Wells	Medical Decision Making	An assessment of methods to combine published survival curves	2000
L. Duchateau, J. P. Pignon, L. Bijnens, S. Bertin, J. Bourhis and R. Sylvester	Controlled Clinical Trials	Individual patient-versus literature-based meta-analysis of survival data: time to event and event rate at a particular time can make a difference, an example based on head and neck cancer	2001
C. Tudur, P. R. Williamson, S. Khan and L. Y. Best	Journal of the Royal Statistical Society Series a-Statistics in Society	The value of the aggregate data approach in meta-analysis with time-to-event outcomes	2001
C. L. Vale, J. F. Tierney and L. A. Stewart	International Journal of Epidemiology	Effects of adjusting for censoring on meta-analyses of time-to-event outcomes	2002
P. R. Williamson, C. T. Smith, J. L. Hutton and A. G. Marson	Statistics in Medicine	Aggregate data meta-analysis with time-to-event outcomes	2002
P. Royston and M. K. Parmar	Statistics in medicine	Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects	2002

O. N. Keene	Statistics in Medicine	Alternatives to the hazard ratio in summarizing efficacy in time-to-event studies: an example from influenza trials	2002
P. F. Moodie, N. A. Nelson and G. G. Koch	Statistics in Medicine	A non-parametric procedure for evaluating treatment effect in the meta-analysis of survival data	2004
S. Michiels, P. Piedbois, S. Burdett, N. Syz, L. Stewart and J. P. Pignon	International Journal of Technology Assessment in Health Care	Meta-analysis when only the median survival times are known: a comparison with individual patient data results	2005
M. C. Simmonds, J. P. Higgins, L. A. Stewart, J. F. Tierney, M. J. Clarke and S. G. Thompson	Clinical Trials	Meta-analysis of individual patient data from randomized trials: a review of methods used in practice	2005
C. T. Smith, P. R. Williamson and A. G. Marson	Journal of Evaluation in Clinical Practice	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	2005
S. Michiels, B. Baujat, C. Mahe, D. J. Sargent and J. P. Pignon	Journal of Clinical Epidemiology	Random effects survival models gave a better understanding of heterogeneity in individual patient data meta-analyses	2005
C. T. Smith, P. R. Williamson and A. G. Marson	Statistics in Medicine	Investigating heterogeneity in an individual patient data meta-analysis of time to event outcomes	2005

C. T. Smith and P. R. Williamson	Clinical Trials	A comparison of methods for fixed effects meta-analysis of individual patient data with time to event outcomes	2007
J. F. Tierney, L. A. Stewart, D. Ghersi, S. Burdett and M. R. Sydes	Trials [Electronic Resource]	Practical methods for incorporating summary time-to-event data into meta-analysis	2007
L. R. Arends, M. G. M. Hunink and T. Stijnen	Statistics in Medicine	Meta-analysis of summary survival curve data	2008
S. Katsahian, A. Latouche, J. Y. Mary, S. Chevret and R. Porcher	Contemporary Clinical Trials	Practical methodology of meta-analysis of individual patient data using a survival outcome	2008
G. Massonnet, P. Janssen and T. Burzykowski	Biometrics	Fitting conditional survival models to meta-analytic data by using a transformation toward mixed-effects models	2008
V. Rondeau, S. Michiels, B. Liquet and J. P. Pignon	Statistics in Medicine	Investigating trial and treatment heterogeneity in an individual patient data meta-analysis of survival data by means of the penalized maximum likelihood approach	2008
T. Hirooka, C. Hamada and I. Yoshimura	Methods of Information in Medicine	A note on estimating treatment effect for time-to-event data in a literature-based meta-analysis	2009
Fiocco, M., Putter, H. and Van Houwelingen, J.C.	Statistics in Medicine	Meta-analysis of pairs of survival curves under heterogeneity: A poisson correlated gamma-frailty approach	2009

Fiocco, M., Putter, H. and Van Houwelingen, J.C.	Biostatistics	A new serially correlated gamma-frailty process for longitudinal count data	2009
N. J. Welton, S. R. Willis and A. E. Ades	Research Synthesis Methods	Synthesis of survival and disease progression outcomes for health technology assessment of cancer therapies	2010
B. S. Woods, N. Hawkins and D. A. Scott	BMC Medical Research Methodology	Network meta-analysis on the log-hazard scale, combining count and hazard ratio statistics accounting for multi-arm trials: A tutorial	2010
X. Yuan and S. J. Anderson	Biometrical Journal	Meta-analysis methodology for combining treatment effects from Cox proportional hazard models with different covariate adjustments	2010
S. Thompson, S. Kaptoge, I. White, A. Wood, P. Perry, J. Danesh and C. Emerging Risk Factors	International Journal of Epidemiology	Statistical methods for the time-to-event analysis of individual participant data from multiple epidemiological studies	2010
F. Siannis, J. K. Barrett, V. T. Farewell and J. F. Tierney	Statistics in Medicine	One-stage parametric meta-analysis of time-to-event outcomes	2010
Ouwens MJ, Philips Z and Jansen JP.	Research Synthesis Methods	Network meta-analysis of parametric survival curves	2010
D. Fisher, A. Copas, J. Tierney and M. Parmar	Journal of clinical epidemiology	A critical review of methods for the assessment of patient-level interactions in individual participant data	2011

		meta-analysis of randomized trials, and guidance for practitioners	
J. Bowden, J. F. Tierney, M. Simmonds, A. J. Copas and J. P. Higgins	Research Synthesis Methods	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	2011
C. Combescure, D. S. Courvoisier, G. Haller and T. V. Perneger	Statistical Methods in Medical Research	Meta-analysis of binary outcomes from two-by-two tables when the length of follow-up varies and hazards are proportional	2011
M. C. Simmonds, J. Tierney, J. Bowden and J. P. Higgins	Research Synthesis Methods	Meta-analysis of time-to-event data: a comparison of two-stage methods	2011
D. B. Rubin	International Journal of Biostatistics	An Alternative to Pooling Kaplan-Meier Curves in Time-to-Event Meta-Analysis	2011
Jansen J.P.	BMC Medical Research Methodology	Network meta-analysis of survival data with fractional polynomials	2011
M. J. Crowther, R. D. Riley, J. A. Staessen, J. Wang, F. Gueyffier and P. C. Lambert	BMC Medical Research Methodology	Individual patient data meta-analysis of survival data using Poisson regression models	2012
J. P. Jansen and S. Cope	BMC Medical Research Methodology	Meta-regression models to address heterogeneity and inconsistency in network meta-analysis of survival outcomes	2012

J. K. Barrett, V. T. Farewell, F. Siannis, J. Tierney and J. P. Higgins	Statistics in Medicine	Two-stage meta-analysis of survival data from individual participants using percentile ratios	2012
M. Fiocco, T. Stijnen and H. Putter	Computational Statistics & Data Analysis	Meta-analysis of time-to-event outcomes using a hazard-based approach: Comparison with other models, robustness and meta-regression	2012
P. Guyot, A. E. Ades, M. Ouwens and N. Welton	BMC Medical Research Methodology	Enhanced secondary analysis of survival data: Reconstructing the data from published Kaplan-Meier survival curves	2012
C. Combescure, D. S. Courvoisier, G. Haller and T. V. Perneger	Statistical Methods in Medical Research	Meta-analysis of two-arm studies: Modelling the intervention effect from survival probabilities	2012
M. C. Simmonds, J. P. Higgins and L. A. Stewart	Research Synthesis Methods	Random-effects meta-analysis of time-to-event data using the expectation-maximisation algorithm and shrinkage estimators	2013
S. Cope and J. P. Jansen	BMC Medical Research Methodology	Quantitative summaries of treatment effect estimates obtained with network meta-analysis of survival curves to inform decision-making	2013
M. M. Bennett, B. J. Crowe, K. L. Price, J. D. Stamey and J. W. Seaman, Jr.	Journal of Biopharmaceutical Statistics	Comparison of Bayesian and frequentist meta-analytical approaches for analyzing time to event data	2013

M. J. Crowther, M. P. Look and R. D. Riley	Statistics in Medicine	Multilevel mixed effects parametric survival models using adaptive Gauss-Hermite quadrature with application to recurrent events and individual participant data meta-analysis	2014
D. Jackson, K. Rollins and P. Coughlin	Research Synthesis Methods	A multivariate model for the meta-analysis of study level survival data at multiple times	2014
P. Saramago, L. H. Chuang and M. O. Soares	BMC Medical Research Methodology	Network meta-analysis of (individual patient) time to event data alongside (aggregate) count data	2014
Y. Wei, P. Royston, J. F. Tierney and M. K. Parmar	Statistics in Medicine	Meta-analysis of time-to-event outcomes from randomized trials using restricted mean survival time: application to individual participant data	2015
R. D. Riley, M. J. Price, D. Jackson, M. Wardle, F. Gueyffier, J. Wang, J. A. Staessen and I. R. White	Research Synthesis Methods	Multivariate meta-analysis using individual participant data	2015
V. Rondeau, J. P. Pignon and S. Michiels	Statistical Methods in Medical Research	A joint model for the dependence between clustered times to tumour progression and deaths: A meta-analysis of chemotherapy in head and neck cancer	2015

S. Batson, G. Greenall and P. Hudson	PLoS ONE [Electronic Resource]	Review of the Reporting of Survival Analyses within Randomised Controlled Trials and the Implications for Meta-Analysis	2016
F. Bonofiglio, J. Beyersmann, M. Schumacher, M. Koller and G. Schwarzer	Research Synthesis Methods	Meta-analysis for aggregated survival data with competing risks: a parametric approach using cumulative incidence functions	2016
B. Lueza, F. Rotolo, J. Bonastre, J. P. Pignon and S. Michiels	BMC Medical Research Methodology	Bias and precision of methods for estimating the difference in restricted mean survival time from an individual patient data meta-analysis	2016
X. V. Wang, B. Cole, M. Bonetti and R. D. Gelber	Statistics in Medicine	Meta-STEPP: subpopulation treatment effect pattern plot for individual patient data meta-analysis	2016
S. C. Freeman and J. R. Carpenter	Research Synthesis Methods	Bayesian one-step IPD network meta-analysis of time-to-event data using Royston-Parmar models	2017
B. Holzhauer	Statistics in Medicine	Meta-analysis of aggregate data on medical events	2017
C. Watkins and I. Bennett	Research Synthesis Methods	A simple method for combining binomial counts or proportions with hazard ratios for evidence synthesis of time-to-event data	2018
S. Ghanbari, N. Zare and Z. Shayan	Advances and Applications in Statistics	A practical method based on functional data analysis and single exponential smoothing to combine survival curves in meta-analysis: A simulation study	2018

S. van Beekhuizen, M. J. Ouwens, M. J. Postma and B. Heeg	Value in Health	Network Meta-analyses in survival data: A comparison and guidance for different methodologies	2018
X. V. Wang, B. Cole, M. Bonetti and R. D. Gelber	Research Synthesis Methods	Meta-STEPP with random effects	2018
V.M.T. de Jong, K. Moons, R. Riley, C.T. Smith, A.G.G Marson, M.J.C. Eikemans, T.P.A. Debray	Research Synthesis Methods	Individual Participant data meta-analysis of intervention studies with time-to-event outcomes: A review of methodology and an applied example	2019
A.Wiksten, N. Hawkins, H-P. Piepho, S. Gsteiger	Value in Health	Nonproportional Hazards in Network Meta-analysis: Efficient Strategies for Model Building and Analysis	2020
B. Holzhauer	Statistics in Biopharmaceutical Research	Methods for Using Aggregate Historical Control Data in Meta-Analyses of Clinical Trials with Time-to-Event Endpoints	2020
S. Cope, K. Chan, J.P. Jansen	Research Synthesis Methods	Multivariate network meta-analysis of survival function parameters	2020
A.Irvine, S. Waise, E.W. Green, B. Stuart	BMC Medical Research Methodology	A non-linear optimisation method to extract summary statistics from Kaplan-Meier survival plots using the published P value	2020
I.Weier, L. Tian, L. Trinquart	Biostatistics	Multivariate meta-analysis model for the difference in restricted mean survival times	2021

E. Ollier, P. Blanchard, G. Le Tueff, S. Michiels	Statistics in Medicine	Penalized Poisson model for network meta-analysis of individual patient time-to-event data	2021
X.Tang, L. Trinquart	Statistics in Medicine	Bayesian multivariate network meta-analysis model for the difference in restricted mean survival times	2021
Tamási, Bálint ; Crowther, Michael ; Puhan, Milo Alan ; Steyerberg, Ewout W ; Hothorn, Torsten	Biostatistics	Individual participant data meta-analysis with mixed-effects transformation models.	2021
C. H. Daly, R. Maconachie, AE Ades, N.J. Welton	Research Synthesis Methods	A non-parametric approach for jointly combining evidence on progression free and overall survival time in network meta-analysis	2021

Table 2.2: Methodological papers obtained from MEDLINE (Ovid Version), Scopus, and Web of Science.

B – Additional material relating to the binary data meta-analyses analysed in Chapter 3

B.1 – Baseline Graphs

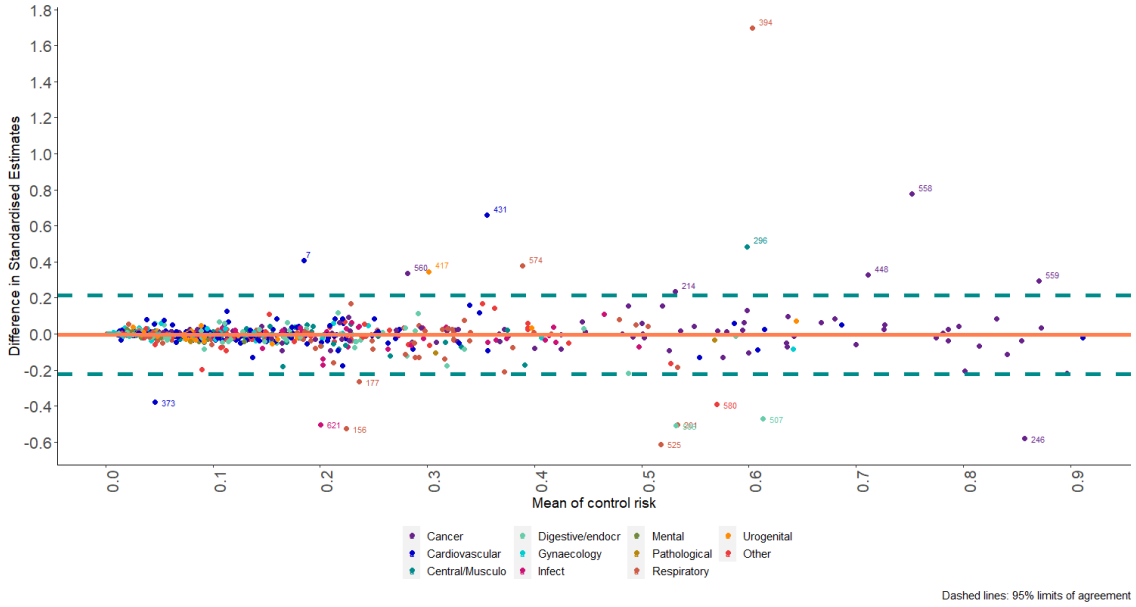


Figure 3.9: Examination of baseline risk for two-stage models

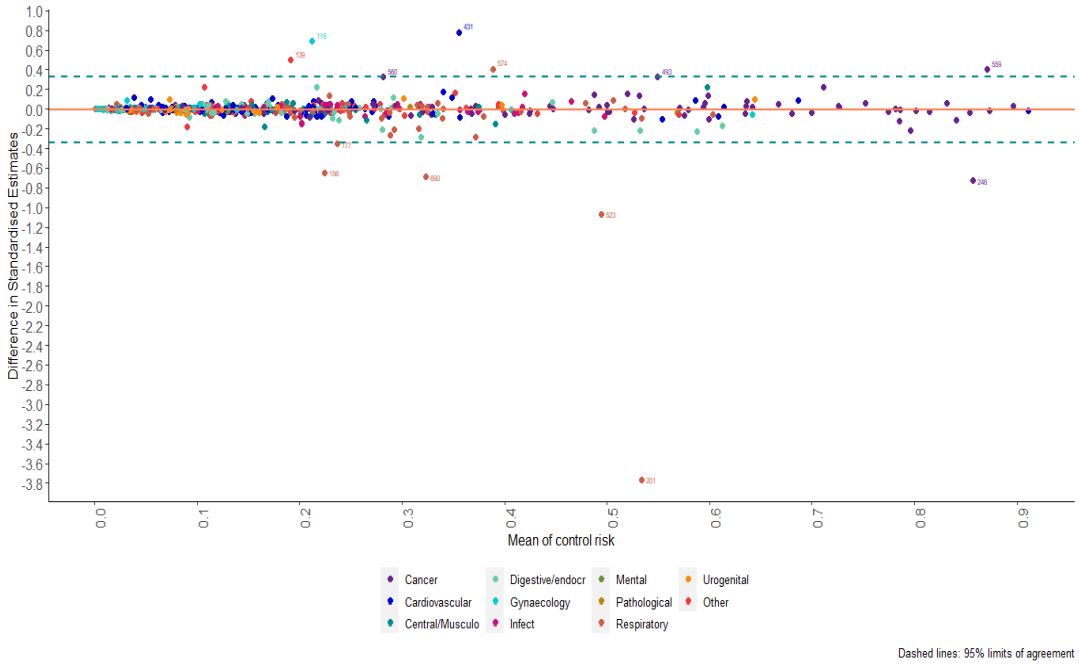


Figure 3.10: Examination of baseline risk for one-stage models.

B.2 – Model Implementation

Two-stage MA for ORs

```
resultsREML=data.frame(matrix(NA, max(CDSR_2008$ma), 8))

colnames(resultsREML)<-c("estimates","SE", "LowerCI", "UpperCI","Tau", "Isq",
"MA", "Med_Area")

for (i in unique(CDSR_2008$ma)) {

    cat(i,"\n")

    try.fit<- try(rma.uni(ai = treat_n, bi = nontreat_n, ci = ctrl_n, di = nonctrl_n,

    data = CDSR_2008[CDSR_2008$ma==i,], measure =
"OR",method="REML",

    control=list(maxiter=500, verbose=TRUE, stepadj=0.5), verbose=TRUE))

    resultsREML[i,7]<-i

    resultsREML[i,8]<-unique(CDSR_2008[CDSR_2008$ma==i,]$scode)

    if (class(try.fit)!="try-error") {

        CDSR.2008[[i]]<- try.fit

        resultsREML[i,1]<-as.numeric(exp(CDSR.2008[[i]]$b))

        resultsREML[i,2]<-as.numeric(CDSR.2008[[i]]$se)

        resultsREML[i,3]<-as.numeric(exp(CDSR.2008[[i]]$ci.lb))

        resultsREML[i,4]<-as.numeric(exp(CDSR.2008[[i]]$ci.ub))

        resultsREML[i,5]<-as.numeric(CDSR.2008[[i]]$tau2)

        resultsREML[i,6]<-as.numeric(CDSR.2008[[i]]$I2)

    } else {

        CDSR.2008[[i]] <- NULL   }}
```

Two-stage MA for HRs

```
resultsREMLHR=data.frame(matrix(NA, max(CDSR_2008$ma), 8))

colnames(resultsREMLHR)<-c("estimates","SE", "LowerCI", "UpperCI","Tau",
"Isq", "MA", "Med_Area")
```

```
for (i in unique(CDSR_2008$ma)) {
    cat(i,"\n")
    try.fit1<- try(rma.uni(yi = logHR, vi = varHR, data =
CDSR_2008[CDSR_2008$ma==i,], method="REML",control=list(maxiter=10e9,
verbose=TRUE, stepadj=0.2), verbose=TRUE))
    resultsREMLHR[i,7]<-i
    resultsREMLHR[i,8]<-unique(CDSR_2008[CDSR_2008$ma==i,]$scode)
    if (class(try.fit1)!="try-error") {
        CDSR.2008HR[[i]]<- try.fit1
        resultsREMLHR[i,1]<-as.numeric(exp(CDSR.2008HR[[i]]$b))
        resultsREMLHR[i,2]<-as.numeric(CDSR.2008HR[[i]]$se)
        resultsREMLHR[i,3]<-as.numeric(exp(CDSR.2008HR[[i]]$ci.lb))
        resultsREMLHR[i,4]<-as.numeric(exp(CDSR.2008HR[[i]]$ci.ub))
        resultsREMLHR[i,5]<-as.numeric(CDSR.2008HR[[i]]$tau2)
        resultsREMLHR[i,6]<-as.numeric(CDSR.2008HR[[i]]$I2)
    } else {
        CDSR.2008HR[[i]] <- NULL }}
resultsREMLHR<-na.omit(resultsREMLHR)
```

One-stage MA for ORs

```
for (i in unique(CDSR_2008$ma)) {
    cat(i,"\n")
    try.fit2<- try(rma.glmm(ai = treat_n, bi = nontreat_n, ci = ctrl_n, di =
nonctrl_n, data = CDSR_2008[CDSR_2008$ma==i,], measure =
"OR",model="UM.FS", drop00=F,nAGQ=7))
    resultsUMFS[i,7]<-i
    resultsUMFS[i,8]<-unique(CDSR_2008[CDSR_2008$ma==i,]$scode)
```

```
    if (class(try.fit2)!="try-error") {
        CDSR.2008stg1[[i]]<- try.fit2
        resultsUMFS[i,1]<-as.numeric(exp(CDSR.2008stg1[[i]]$b))
        resultsUMFS[i,2]<-as.numeric(CDSR.2008stg1[[i]]$se)
        resultsUMFS[i,3]<-as.numeric(exp(CDSR.2008stg1[[i]]$ci.lb))
        resultsUMFS[i,4]<-as.numeric(exp(CDSR.2008stg1[[i]]$ci.ub))
        resultsUMFS[i,5]<-as.numeric(CDSR.2008stg1[[i]]$tau2)
        resultsUMFS[i,6]<-as.numeric(CDSR.2008stg1[[i]]$I2)
    } else {
        CDSR.2008stg1[[i]] <- NULL  }}
```

One-stage MA for HRs

```
for (i in unique(datlong.CDSR_2008$ma.num)) {
    cat(i,"\n")
    try.fit3<-try(glmer(cbind(event,n-event) ~ factor(treat) + factor(study) +
(treat12-1|study), data=datlong.CDSR_2008[datlong.CDSR_2008$ma.num==i,],
family=binomial(link="cloglog"),nAGQ=7))
    resultsUMFSHR[i,6]<-i
    resultsUMFSHR[i,7]<-
unique(datlong.CDSR_2008[datlong.CDSR_2008$ma.num==i,]$medical.area)
    if (class(try.fit3)!="try-error") {
        CDSR.2008long[[i]]<- try.fit3
        CDSR.2008long1.CI[[i]]<-confint.merMod(CDSR.2008long[[i]],
method="Wald")
        resultsUMFSHR[i,1]<-
as.numeric(exp(summary(CDSR.2008long[[i]])$coeff[2,1]))
```

```
        resultsUMFSHR[i,2]<-
as.numeric(summary(CDSR.2008long[[i]])$coeff[2,2])
        resultsUMFSHR[i,3]<-
as.numeric(summary(CDSR.2008long[[i]])$varcor)
        resultsUMFSHR[i,4]<-exp(CDSR.2008long1.CI[[i]][3,1])
        resultsUMFSHR[i,5]<-exp(CDSR.2008long1.CI[[i]][3,2])
    } else {
        CDSR.2008long[[i]] <- NULL   }}
```

B.3 – Clog-log link & HR derivation

A) Information on the Complementary log-log link (clog-log)

Singer et al.[125] and Hedeker et al.[128] are providing all the information including the assumptions and details involved when models are analysed under a logit or a clog-log link. Both papers are describing that another useful link function for the discrete-time hazards models that worth's consideration is the clog-log function. While logit link provides with the logarithm of the odds of event of occurance, clog-log link produces the logarithm of the negative logarithm of the probability of event non-occurrence[125]. Differences between the logit and clog-log transformations invoke the following: 1) the logit transformation is symmetric mapping event probabilities from $[0,1]$ to $(-\infty,\infty)$ (i.e. without lower and upper bound) whereas clog-log transformation is asymmetric meaning that approaches zero at a slower pace than approaches 1, and does not have upper and lower bound. 2) When small hazards are involved (i.e. the probability of the event occurring is small), both transformations produce similar results whereas at higher values of hazard the transformation produce discordant results as shown in Figure 3.8. 3) Logit link has a build-in proportional odds assumption whereas clog-log link a build-in proportional hazards assumption, a direct analogy to survival analysis in which the same assumption is made. Familiar terminology for model specification, comfortability with results interpretation and widely available software for estimation of the results are important advantages of the logit transformation. On the other hand, specifying a model with a clog-log link invokes the following assumptions: a) "for each combination of predictor values, there is a postulated clog-log hazard function; b) each of these clog-log hazard functions has an identical shape; c) and the distance between each of these clog-log hazard functions is identical in every time period"[125].

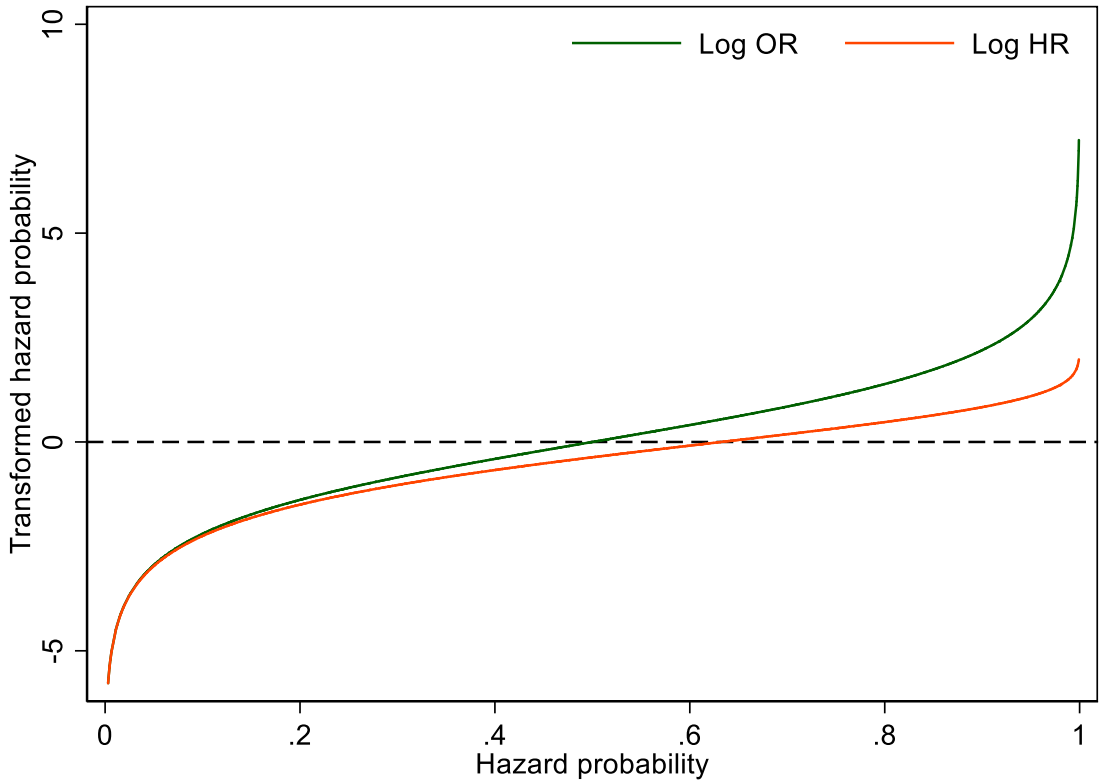


Figure 3.11: Replication of the figure presented on (Singer et al., 2003) paper. Identification of differences when comparing effects of the logit and clog-log transformations.

B) Derivation of the log Hazard Ratio and its variance using the clog-log link for the two-stage MA models.

Assuming $Y = cloglog(x) = \log[-log(1-x)]$, and

$Var(y) \approx \left(\frac{dy}{dx}\right)^2_{x=E(x)} Var(x)$ we have:

If $f(x) = log\,[-log(1-x)]$, then:

$f'(x) = \frac{d}{dx}\{\log[-\log(1-x)]\} = \frac{1}{-[\log(1-x)]}[-\log(1-x)]'$ or

$f'(x) = \frac{1}{-[\log(1-x)]}[\underbrace{(-1)'\log(1-x)}_{0} + (-1)\log(1-x)']$ or

$f'(x) = \frac{1}{-[\log(1-x)]} + (-1)\frac{1}{1-x}(1-x)'$ or

$f'(x) = \frac{1}{-[\log(1-x)]} + (-1)\frac{1}{1-x}[\underbrace{(-1)'}_{0} - (x)']$ or

$f'(x) = \frac{1}{-[\log(1-x)]} + (-1)\frac{1}{1-x}(-1)$ or

$f'(x) = \frac{1}{-[\log(1-x)]}\frac{1}{1-x}$ or

$$f'(x) = \frac{1}{\log(1-x)\,(x-1)}$$

In our occasion:

For the treatment arm: $Y_1 = \log[-\log(1-x_1)]$

For the control arm: $Y_0 = \log[-\log(1-x_0)]$

$$E(Y_1 - Y_0) = \log[-\log(1-x_1)] - \log[-\log(1-x_0)]$$

$$Var(Y_1 - Y_0) = Var\{\log[-\log(1-x_1)]\} + Var\{\log[-\log(1-x_0)]\}$$

Where $x_1 = \frac{A}{A+B}$, and $x_0 = \frac{C}{C+D}$ is the proportion of events for the treatment and control arms obtained from the $2x2$ table.

B.4 – Calculation of I^2

Assuming we have obtained the variance of the logarithm of the HRs and ORs, s_i^2, we can obtain the corresponding within-study precisions as $w_i = \frac{1}{s_i^2}$, and the within-study precisions of power of 2, $w_i^2 = \frac{1}{(s_i^2)^2}$.

Then, according to Higgins et al.[172], we can obtain the within-study variances as:

$$\sigma^2 = \frac{\sum w_i(m-1)}{(\sum w_i)^2 - \sum w_i^2}$$

where m is the number of studies included in meta-analysis.

From the models applied we can obtain τ^2, and hence the between-study heterogeneity using the following formula:

$$I^2 = \frac{\tau^2}{\tau^2 + \sigma^2}$$

B.5 – Table containing the exact results from the two-stage meta-analysis models & additional forest plots considered as outliers from the Bland-Altman plots

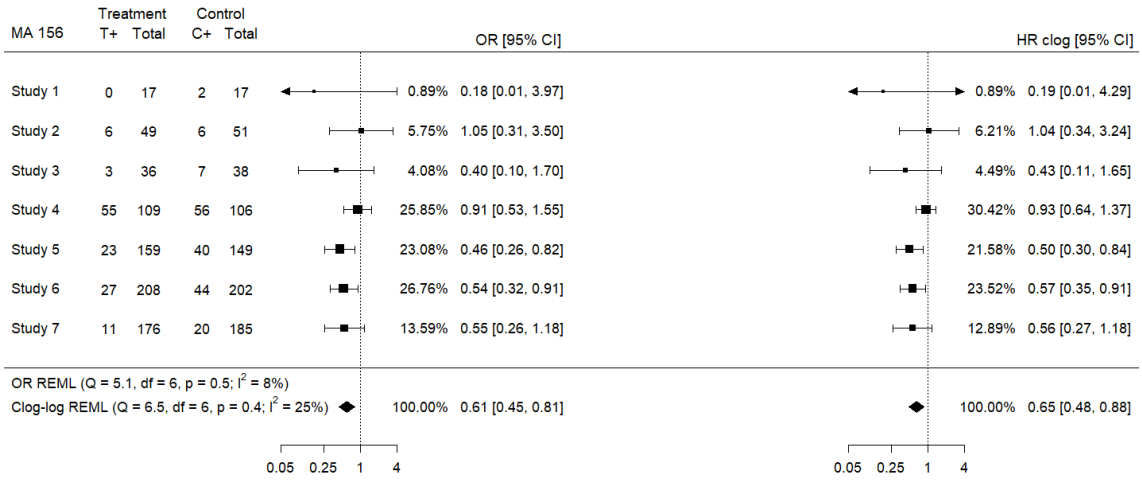
Two-Stage Random-Effects Model			
MA Identifier	OR (95% CI) vs. HR (95% CI)	τ^2 OR vs. τ^2 HR	I^2 OR vs. I^2 HR
7	1.036 (0.946, 1.135) vs. 1.019 (0.916, 1.134)	0.001 vs. 0.003	3% vs. 8%
156	0.606 (0.451, 0.813) vs. 0.652 (0.484, 0.878)	0.012 vs. 0.038	8% vs. 25%
158	0.744 (0.524, 1.058) vs. 0.797 (0.584, 1.086)	0.030 vs. 0.033	31% vs. 44%
177	0.662 (0.495, 0.886) vs. 0.701 (0.532, 0.926)	0.028 vs. 0.036	22% vs. 32%
201	0.190 (0.065, 0.555) vs. 0.250 (0.085, 0.730)	0.333 vs. 0.386	20% vs. 25%
214	0.733 (0.586, 0.916) vs. 0.791 (0.678, 0.924)	0.005 vs. 0.000	4% vs. 0%
246	0.674 (0.448, 1.013) vs. 0.877 (0.722, 1.066)	0.000 vs. 0.007	0% vs. 9%
296	0.262 (0.044, 1.569) vs. 0.382 (0.146, 1.005)	1.146 vs. 0.242	45% vs. 30%
322	0.537 (0.343, 0.842) vs. 0.694 (0.522, 0.923)	0.010 vs. 0.017	6% vs. 26%
327	0.522 (0.275, 0.993) vs. 0.681 (0.469, 0.988)	0.000 vs. 0.081	0% vs. 32%
330	0.795 (0.575, 1.098) vs. 0.824 (0.618, 1.098)	0.033 vs. 0.083	15% vs. 52%
331	0.910 (0.622, 1.331) vs. 0.963 (0.790, 1.174)	0.000 vs. 0.013	0% vs. 16%
373	1.363 (1.050, 1.770) vs. 1.350 (1.086, 1.678)	0.027 vs. 0.012	12% vs. 6%
394	0.662 (0.161, 2.726) vs. 0.733 (0.334, 1.608)	1.974 vs. 0.515	82% vs. 75%
417	0.742 (0.344, 1.599) vs. 0.712 (0.390, 1.300)	0.226 vs. 0.109	21% vs. 16%
431	0.501 (0.413, 0.609) vs. 0.569 (0.492, 0.658)	0.116 vs. 0.056	70% vs. 62%
434	1.193 (0.461, 3.089) vs. 1.135 (0.588, 2.190)	0.150 vs. 0.139	12% vs. 31%
448	0.784 (0.547, 1.123) vs. 0.853 (0.707, 1.030)	0.027 vs. 0.000	13% vs. 0%
506	0.673 (0.404, 1.121) vs. 0.807 (0.533, 1.222)	0.146 vs. 0.162	22% vs. 38%
507	0.442 (0.321, 0.608) vs. 0.603 (0.485, 0.750)	0.128 vs. 0.068	29% vs. 33%
525	0.686 (0.541, 0.870) vs. 0.746 (0.592, 0.940)	0.000 vs. 0.012	0% vs. 14%
558	1.263 (0.621, 2.565) vs. 0.964 (0.556, 1.670)	0.241 vs. 0.239	49% vs. 78%

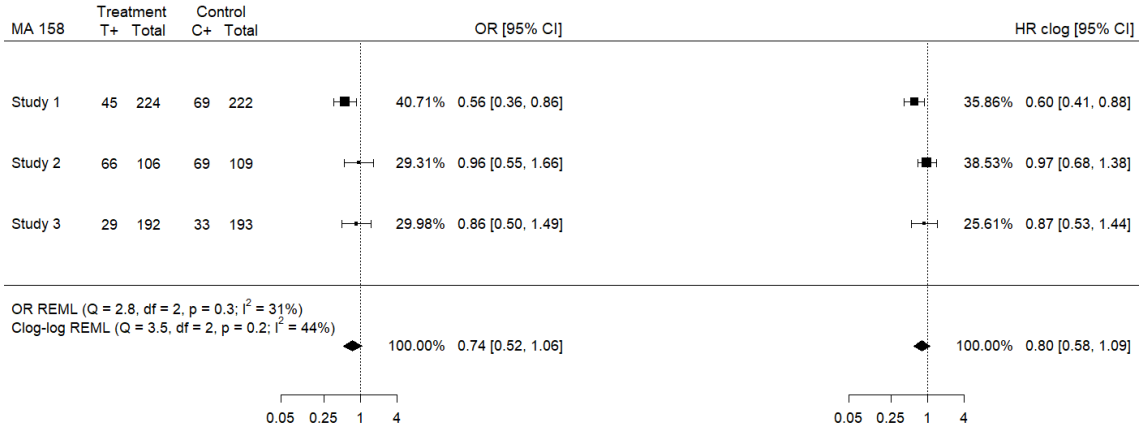
559	1.274 (0.482, 3.368) vs. 1.046 (0.665, 1.644)	0.197 vs. 0.071	26% vs. 44%
560	1.429 (0.819, 2.492) vs. 1.278 (0.759, 2.150)	0.000 vs. 0.057	0% vs. 26%
574	0.835 (0.552, 1.262) vs. 0.778 (0.522, 1.159)	0.080 vs. 0.097	34% vs. 46%
580	0.606 (0.116, 3.168) vs. 0.922 (0.419, 2.031)	0.990 vs. 0.000	47% vs. 0%
621	1.033 (0.700, 1.524) vs. 1.082 (0.857, 1.366)	0.062 vs. 0.000	32% vs. 0%
647	1.038 (0.719, 1.497) vs. 1.053 (0.745, 1.490)	0.000 vs. 0.045	0% vs. 36%
711	0.607 (0.437, 0.845) vs. 0.789 (0.665, 0.935)	0.000 vs. 0.007	0% vs. 19%

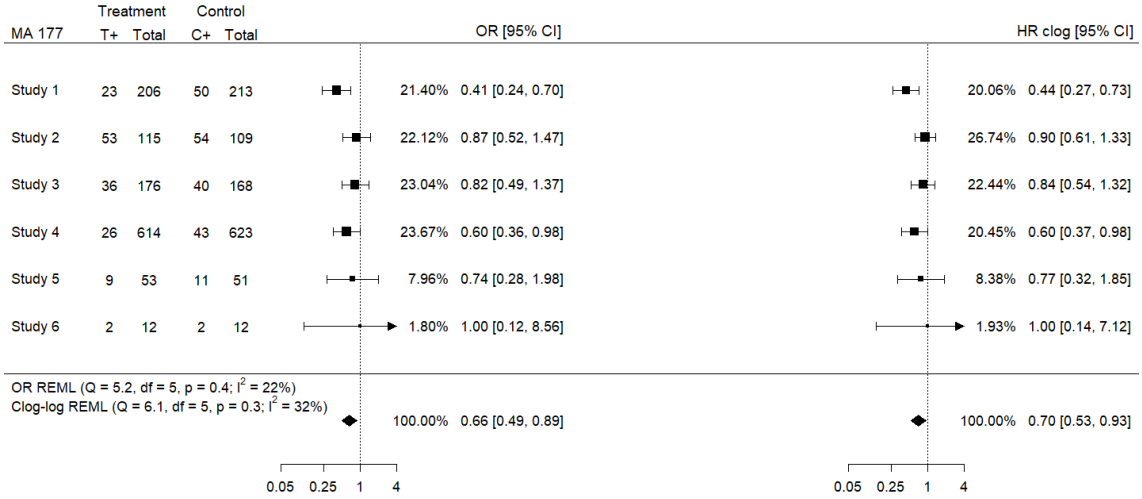
Table 3.5: Characteristics of meta-analyses outside the 95% limits of agreement based on difference of standardised estimates and difference in I^2 (Two-stage models).

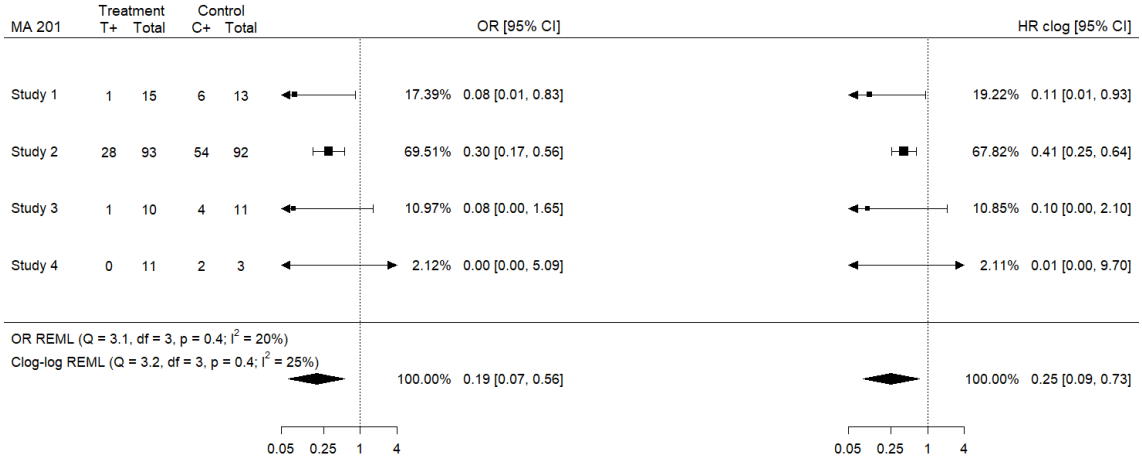
MA coloured in **blue** represent characteristics of studies outside the 95% limits of agreement based on difference of standardised estimates. MA coloured in **red** represent characteristics of studies outside the 95% limits of agreement based on difference in I^2. MA coloured in **black** represent characteristics of studies outside the 95% limits of agreement based on difference of standardised estimates and difference in I^2.

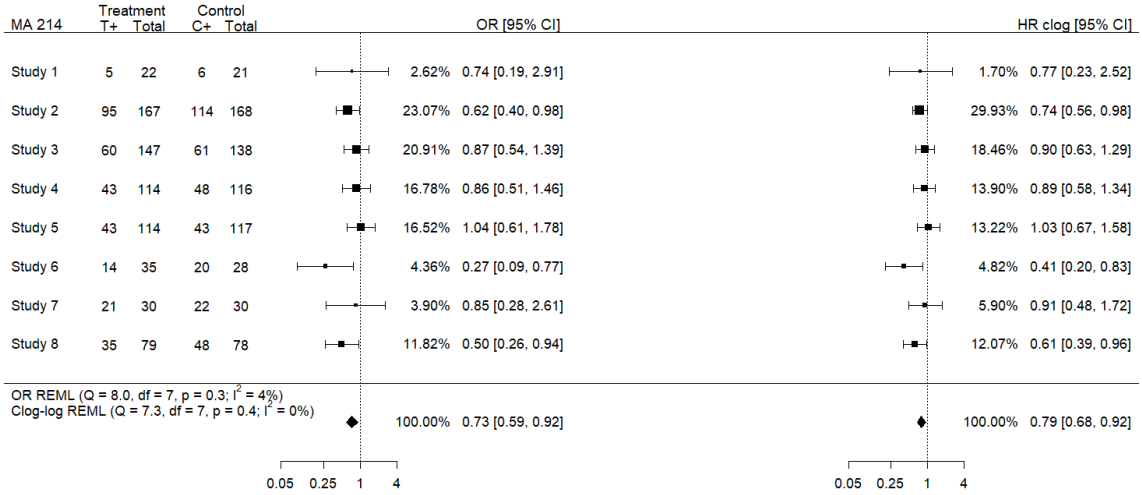
The meta-analysis forest plots below correspond to the meta-analyses presented in Table 3.5. The meta-analyses already presented in Chapter 3 were omitted from the figures below.

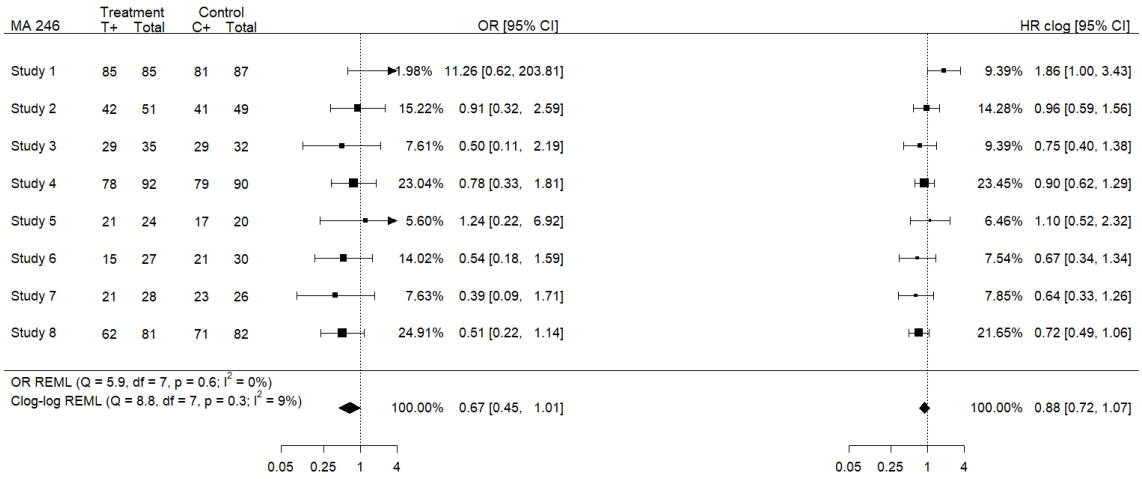


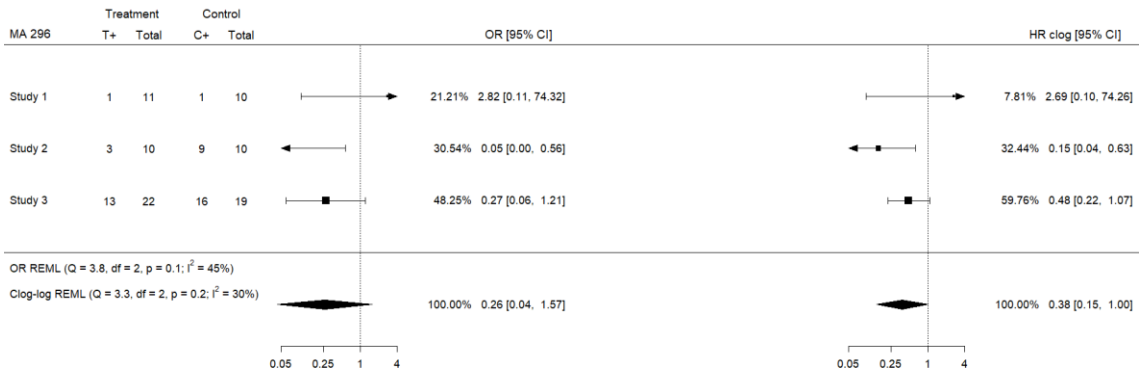

MA 158
Treatment
T+ Total
Control
C+ Total
OR [95% CI]
HR clog [95% CI]
Study 1 45 224 69 222 40.71% 0.56 [0.36, 0.86] 35.86% 0.60 [0.41, 0.88]
Study 2 66 106 69 109 29.31% 0.96 [0.55, 1.66] 38.53% 0.97 [0.68, 1.38]
Study 3 29 192 33 193 29.98% 0.86 [0.50, 1.49] 25.61% 0.87 [0.53, 1.44]
OR REML (Q = 2.8, df = 2, p = 0.3; I² = 31%) 100.00% 0.74 [0.52, 1.06]
Clog-log REML (Q = 3.5, df = 2, p = 0.2; I² = 44%) 100.00% 0.80 [0.58, 1.09]
0.05 0.25 1 4
0.05 0.25 1 4

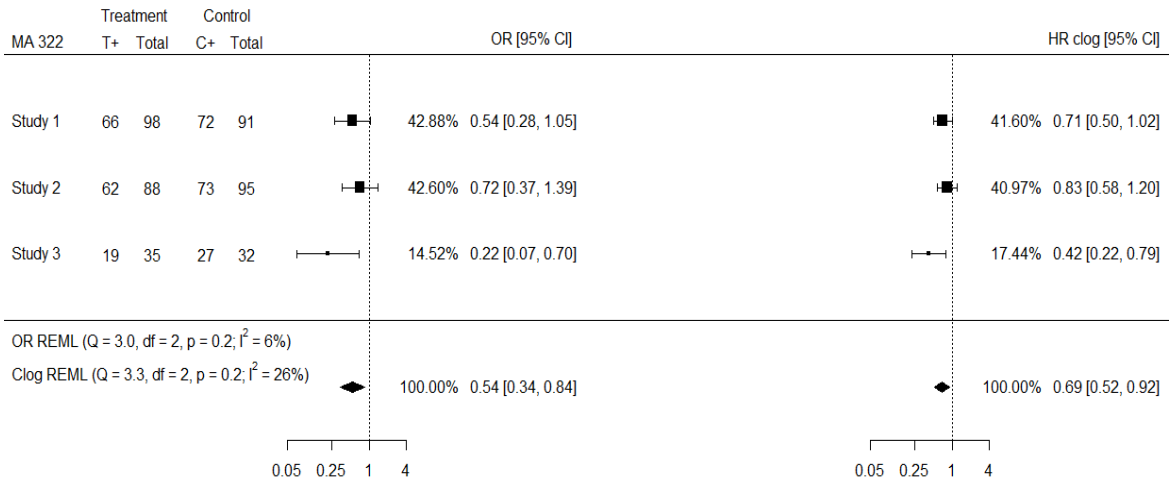

MA 177
Treatment
T+ Total
Control
C+ Total
OR [95% CI]
HR clog [95% CI]
Study 1 23 206 50 213 21.40% 0.41 [0.24, 0.70] 20.06% 0.44 [0.27, 0.73]
Study 2 53 115 54 109 22.12% 0.87 [0.52, 1.47] 26.74% 0.90 [0.61, 1.33]
Study 3 36 176 40 168 23.04% 0.82 [0.49, 1.37] 22.44% 0.84 [0.54, 1.32]
Study 4 26 614 43 623 23.67% 0.60 [0.36, 0.98] 20.45% 0.60 [0.37, 0.98]
Study 5 9 53 11 51 7.96% 0.74 [0.28, 1.98] 8.38% 0.77 [0.32, 1.85]
Study 6 2 12 2 12 1.80% 1.00 [0.12, 8.56] 1.93% 1.00 [0.14, 7.12]
OR REML (Q = 5.2, df = 5, p = 0.4; I² = 22%) 100.00% 0.66 [0.49, 0.89]
Clog-log REML (Q = 6.1, df = 5, p = 0.3; I² = 32%) 100.00% 0.70 [0.53, 0.93]
0.05 0.25 1 4
0.05 0.25 1 4

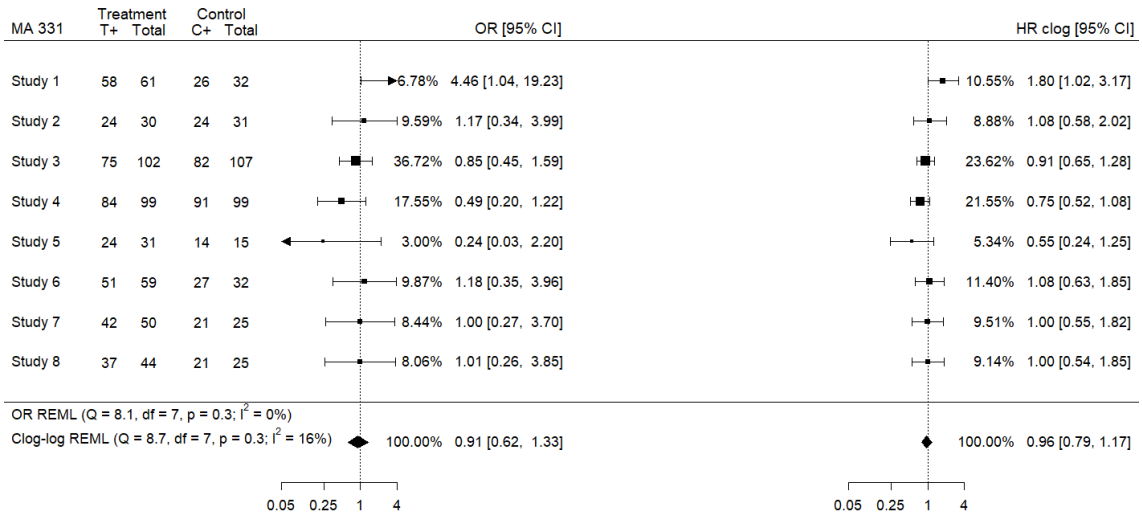

MA 201
Treatment
T+ Total
Control
C+ Total
OR [95% CI]
HR clog [95% CI]
Study 1 1 15 6 13 17.39% 0.08 [0.01, 0.83] 19.22% 0.11 [0.01, 0.93]
Study 2 28 93 54 92 69.51% 0.30 [0.17, 0.56] 67.82% 0.41 [0.25, 0.64]
Study 3 1 10 4 11 10.97% 0.08 [0.00, 1.65] 10.85% 0.10 [0.00, 2.10]
Study 4 0 11 2 3 2.12% 0.00 [0.00, 5.09] 2.11% 0.01 [0.00, 9.70]
OR REML (Q = 3.1, df = 3, p = 0.4; I² = 20%)
Clog-log REML (Q = 3.2, df = 3, p = 0.4; I² = 25%) 100.00% 0.19 [0.07, 0.56] 100.00% 0.25 [0.09, 0.73]
0.05 0.25 1 4
0.05 0.25 1 4

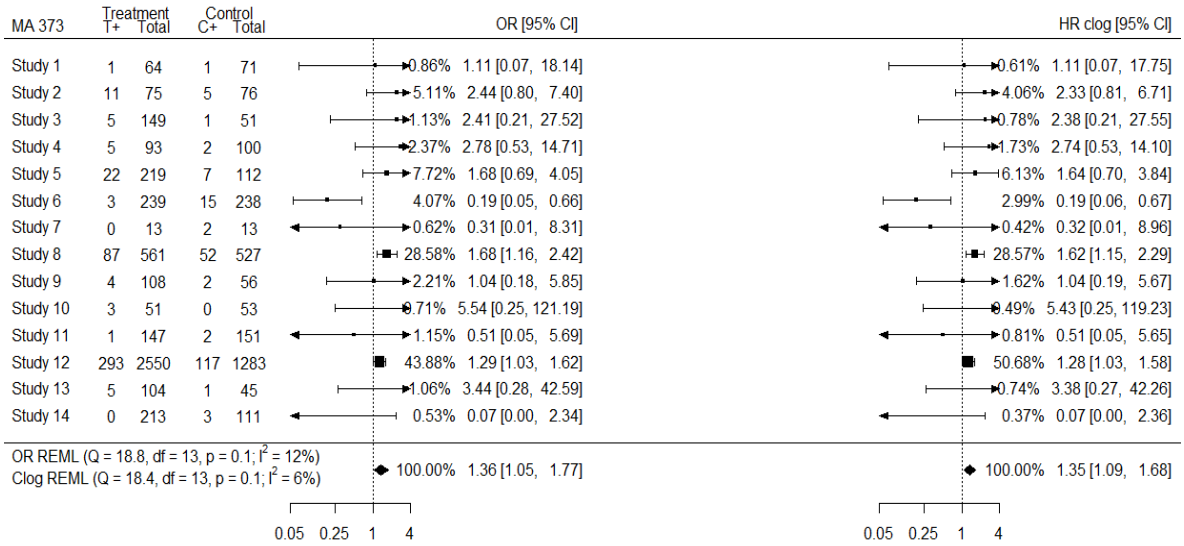

MA 214
Treatment
T+ Total
Control
C+ Total
OR [95% CI]
HR clog [95% CI]
Study 1 5 22 6 21 2.62% 0.74 [0.19, 2.91] 1.70% 0.77 [0.23, 2.52]
Study 2 95 167 114 168 23.07% 0.62 [0.40, 0.98] 29.93% 0.74 [0.56, 0.98]
Study 3 60 147 61 138 20.91% 0.87 [0.54, 1.39] 18.46% 0.90 [0.63, 1.29]
Study 4 43 114 48 116 16.78% 0.86 [0.51, 1.46] 13.90% 0.89 [0.58, 1.34]
Study 5 43 114 43 117 16.52% 1.04 [0.61, 1.78] 13.22% 1.03 [0.67, 1.58]
Study 6 14 35 20 28 4.36% 0.27 [0.09, 0.77] 4.82% 0.41 [0.20, 0.83]
Study 7 21 30 22 30 3.90% 0.85 [0.28, 2.61] 5.90% 0.91 [0.48, 1.72]
Study 8 35 79 48 78 11.82% 0.50 [0.26, 0.94] 12.07% 0.61 [0.39, 0.96]
OR REML (Q = 8.0, df = 7, p = 0.3; I² = 4%)
Clog-log REML (Q = 7.3, df = 7, p = 0.4; I² = 0%)
100.00% 0.73 [0.59, 0.92] 100.00% 0.79 [0.68, 0.92]
0.05 0.25 1 4
0.05 0.25 1 4

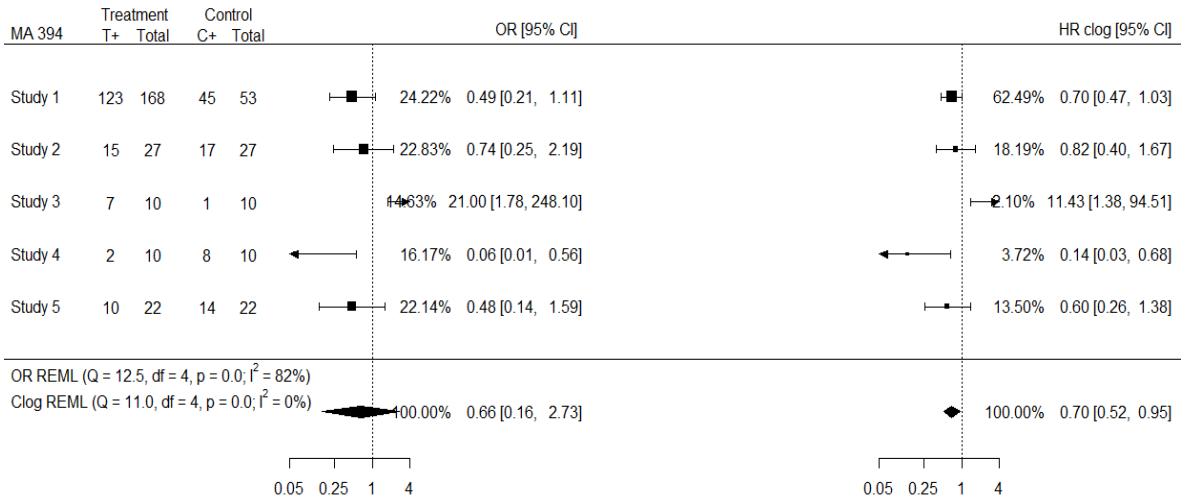

MA 246
Treatment
T+ Total
Control
C+ Total
OR [95% CI]
HR clog [95% CI]
Study 1 85 85 81 87 1.98% 11.26 [0.62, 203.81] 9.39% 1.86 [1.00, 3.43]
Study 2 42 51 41 49 15.22% 0.91 [0.32, 2.59] 14.28% 0.96 [0.59, 1.56]
Study 3 29 35 29 32 7.61% 0.50 [0.11, 2.19] 9.39% 0.75 [0.40, 1.38]
Study 4 78 92 79 90 23.04% 0.78 [0.33, 1.81] 23.45% 0.90 [0.62, 1.29]
Study 5 21 24 17 20 5.60% 1.24 [0.22, 6.92] 6.46% 1.10 [0.52, 2.32]
Study 6 15 27 21 30 14.02% 0.54 [0.18, 1.59] 7.54% 0.67 [0.34, 1.34]
Study 7 21 28 23 26 7.63% 0.39 [0.09, 1.71] 7.85% 0.64 [0.33, 1.26]
Study 8 62 81 71 82 24.91% 0.51 [0.22, 1.14] 21.65% 0.72 [0.49, 1.06]
OR REML (Q = 5.9, df = 7, p = 0.6; I² = 0%)
Clog-log REML (Q = 8.8, df = 7, p = 0.3; I² = 9%)
100.00% 0.67 [0.45, 1.01] 100.00% 0.88 [0.72, 1.07]
0.05 0.25 1 4
0.05 0.25 1 4

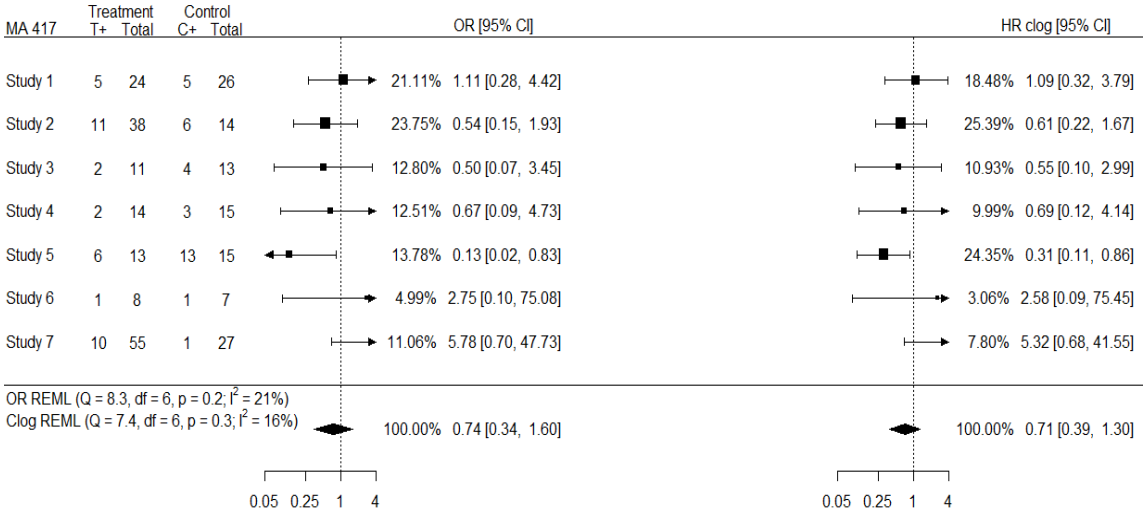

MA 296
Treatment
T+ Total
Control
C+ Total
OR [95% CI]
HR clog [95% CI]
Study 1 1 11 1 10 21.21% 2.82 [0.11, 74.32] 7.81% 2.69 [0.10, 74.26]
Study 2 3 10 9 10 30.54% 0.05 [0.00, 0.56] 32.44% 0.15 [0.04, 0.63]
Study 3 13 22 16 19 48.25% 0.27 [0.06, 1.21] 59.76% 0.48 [0.22, 1.07]
OR REML (Q = 3.8, df = 2, p = 0.1; I² = 45%)
Clog-log REML (Q = 3.3, df = 2, p = 0.2; I² = 30%)
100.00% 0.26 [0.04, 1.57] 100.00% 0.38 [0.15, 1.00]
0.05 0.25 1 4
0.05 0.25 1 4

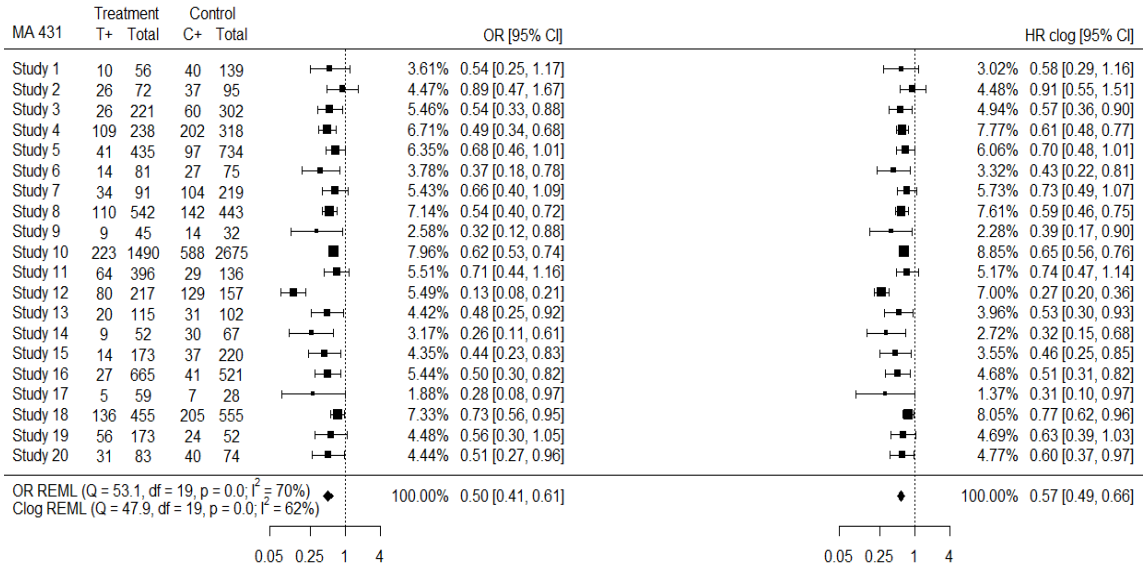

MA 322
Treatment
T+ Total
Control
C+ Total
OR [95% CI]
HR clog [95% CI]
Study 1 66 98 72 91 42.88% 0.54 [0.28, 1.05] 41.60% 0.71 [0.50, 1.02]
Study 2 62 88 73 95 42.60% 0.72 [0.37, 1.39] 40.97% 0.83 [0.58, 1.20]
Study 3 19 35 27 32 14.52% 0.22 [0.07, 0.70] 17.44% 0.42 [0.22, 0.79]
OR REML (Q = 3.0, df = 2, p = 0.2; I² = 6%) 100.00% 0.54 [0.34, 0.84]
Clog REML (Q = 3.3, df = 2, p = 0.2; I² = 26%) 100.00% 0.69 [0.52, 0.92]
0.05 0.25 1 4
0.05 0.25 1 4

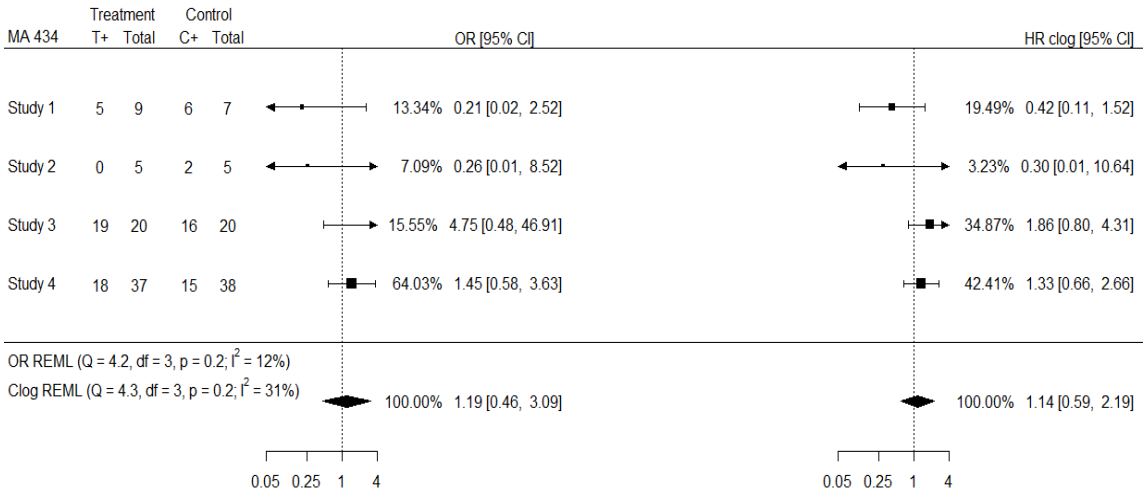

MA 331
Treatment
T+ Total
Control
C+ Total
OR [95% CI]
HR clog [95% CI]
Study 1 58 61 26 32 6.78% 4.46 [1.04, 19.23] 10.55% 1.80 [1.02, 3.17]
Study 2 24 30 24 31 9.59% 1.17 [0.34, 3.99] 8.88% 1.08 [0.58, 2.02]
Study 3 75 102 82 107 36.72% 0.85 [0.45, 1.59] 23.62% 0.91 [0.65, 1.28]
Study 4 84 99 91 99 17.55% 0.49 [0.20, 1.22] 21.55% 0.75 [0.52, 1.08]
Study 5 24 31 14 15 3.00% 0.24 [0.03, 2.20] 5.34% 0.55 [0.24, 1.25]
Study 6 51 59 27 32 9.87% 1.18 [0.35, 3.96] 11.40% 1.08 [0.63, 1.85]
Study 7 42 50 21 25 8.44% 1.00 [0.27, 3.70] 9.51% 1.00 [0.55, 1.82]
Study 8 37 44 21 25 8.06% 1.01 [0.26, 3.85] 9.14% 1.00 [0.54, 1.85]
OR REML (Q = 8.1, df = 7, p = 0.3; I² = 0%) 100.00% 0.91 [0.62, 1.33]
Clog-log REML (Q = 8.7, df = 7, p = 0.3; I² = 16%) 100.00% 0.96 [0.79, 1.17]
0.05 0.25 1 4
0.05 0.25 1 4

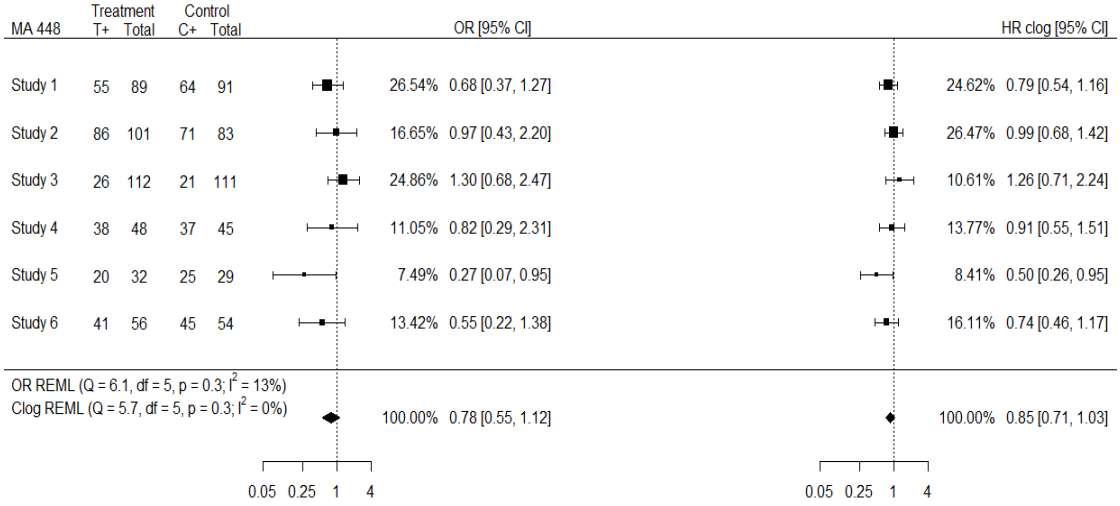

MA 373
Treatment
T+ Total
Control
C+ Total
OR [95% CI]
HR clog [95% CI]
Study 1 1 64 1 71 0.86% 1.11 [0.07, 18.14] 0.61% 1.11 [0.07, 17.75]
Study 2 11 75 5 76 5.11% 2.44 [0.80, 7.40] 4.06% 2.33 [0.81, 6.71]
Study 3 5 149 1 51 1.13% 2.41 [0.21, 27.52] 0.78% 2.38 [0.21, 27.55]
Study 4 5 93 2 100 2.37% 2.78 [0.53, 14.71] 1.73% 2.74 [0.53, 14.10]
Study 5 22 219 7 112 7.72% 1.68 [0.69, 4.05] 6.13% 1.64 [0.70, 3.84]
Study 6 3 239 15 238 4.07% 0.19 [0.05, 0.66] 2.99% 0.19 [0.06, 0.67]
Study 7 0 13 2 13 0.62% 0.31 [0.01, 8.31] 0.42% 0.32 [0.01, 8.96]
Study 8 87 561 52 527 28.58% 1.68 [1.16, 2.42] 28.57% 1.62 [1.15, 2.29]
Study 9 4 108 2 56 2.21% 1.04 [0.18, 5.85] 1.62% 1.04 [0.19, 5.67]
Study 10 3 51 0 53 0.71% 5.54 [0.25, 121.19] 0.49% 5.43 [0.25, 119.23]
Study 11 1 147 2 151 1.15% 0.51 [0.05, 5.69] 0.81% 0.51 [0.05, 5.65]
Study 12 293 2550 117 1283 43.88% 1.29 [1.03, 1.62] 50.68% 1.28 [1.03, 1.58]
Study 13 5 104 1 45 1.06% 3.44 [0.28, 42.59] 0.74% 3.38 [0.27, 42.26]
Study 14 0 213 3 111 0.53% 0.07 [0.00, 2.34] 0.37% 0.07 [0.00, 2.36]
OR REML (Q = 18.8, df = 13, p = 0.1; I² = 12%) 100.00% 1.36 [1.05, 1.77]
Clog REML (Q = 18.4, df = 13, p = 0.1; I² = 6%) 100.00% 1.35 [1.09, 1.68]
0.05 0.25 1 4
0.05 0.25 1 4

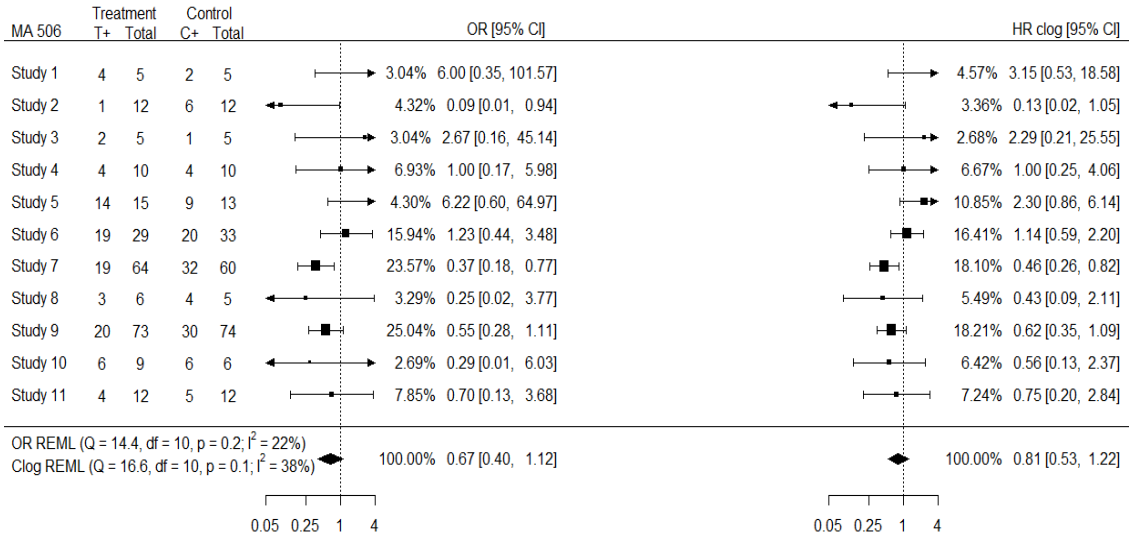

MA 394
Treatment
T+ Total
Control
C+ Total
OR [95% CI]
HR clog [95% CI]
Study 1 123 168 45 53 24.22% 0.49 [0.21, 1.11] 62.49% 0.70 [0.47, 1.03]
Study 2 15 27 17 27 22.83% 0.74 [0.25, 2.19] 18.19% 0.82 [0.40, 1.67]
Study 3 7 10 1 10 14.63% 21.00 [1.78, 248.10] 2.10% 11.43 [1.38, 94.51]
Study 4 2 10 8 10 16.17% 0.06 [0.01, 0.56] 3.72% 0.14 [0.03, 0.68]
Study 5 10 22 14 22 22.14% 0.48 [0.14, 1.59] 13.50% 0.60 [0.26, 1.38]
OR REML (Q = 12.5, df = 4, p = 0.0; I² = 82%) 100.00% 0.66 [0.16, 2.73]
Clog REML (Q = 11.0, df = 4, p = 0.0; I² = 0%) 100.00% 0.70 [0.52, 0.95]
0.05 0.25 1 4
0.05 0.25 1 4

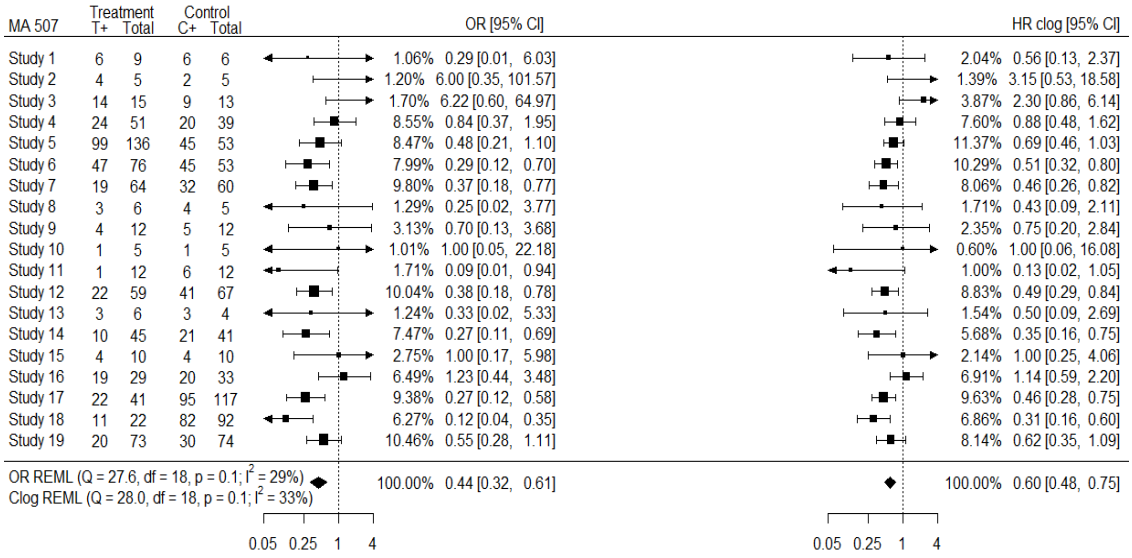

MA 417
Treatment
T+ Total
Control
C+ Total
OR [95% CI]
HR clog [95% CI]
Study 1 5 24 5 26 21.11% 1.11 [0.28, 4.42] 18.48% 1.09 [0.32, 3.79]
Study 2 11 38 6 14 23.75% 0.54 [0.15, 1.93] 25.39% 0.61 [0.22, 1.67]
Study 3 2 11 4 13 12.80% 0.50 [0.07, 3.45] 10.93% 0.55 [0.10, 2.99]
Study 4 2 14 3 15 12.51% 0.67 [0.09, 4.73] 9.99% 0.69 [0.12, 4.14]
Study 5 6 13 13 15 13.78% 0.13 [0.02, 0.83] 24.35% 0.31 [0.11, 0.86]
Study 6 1 8 1 7 4.99% 2.75 [0.10, 75.08] 3.06% 2.58 [0.09, 75.45]
Study 7 10 55 1 27 11.06% 5.78 [0.70, 47.73] 7.80% 5.32 [0.68, 41.55]
OR REML (Q = 8.3, df = 6, p = 0.2; I² = 21%) 100.00% 0.74 [0.34, 1.60]
Clog REML (Q = 7.4, df = 6, p = 0.3; I² = 16%) 100.00% 0.71 [0.39, 1.30]
0.05 0.25 1 4
0.05 0.25 1 4

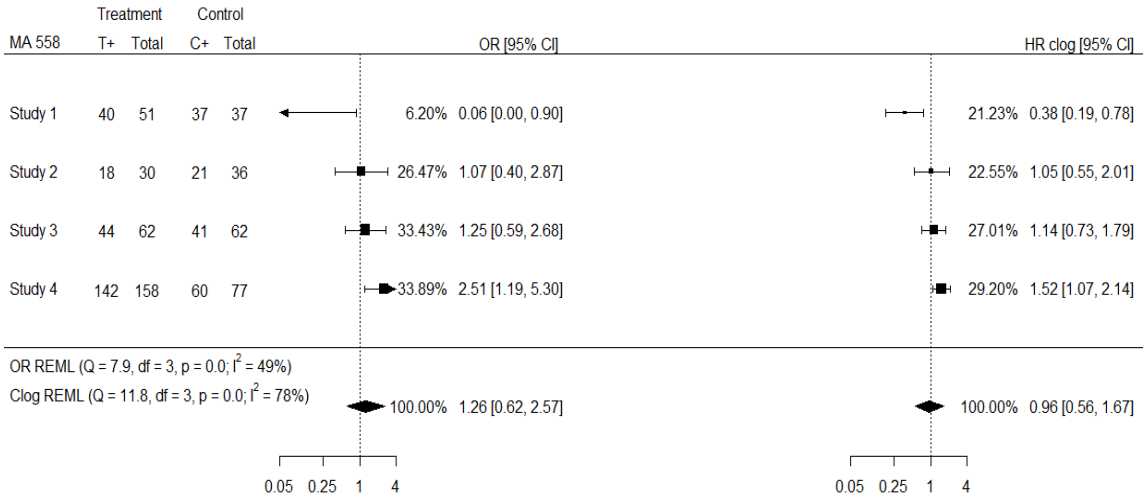

MA 431
Treatment
T+ Total
Control
C+ Total
OR [95% CI]
HR clog [95% CI]
Study 1 10 56 40 139 3.61% 0.54 [0.25, 1.17] 3.02% 0.58 [0.29, 1.16]
Study 2 26 72 37 95 4.47% 0.89 [0.47, 1.67] 4.48% 0.91 [0.55, 1.51]
Study 3 26 221 60 302 5.46% 0.54 [0.33, 0.88] 4.94% 0.57 [0.36, 0.90]
Study 4 109 238 202 318 6.71% 0.49 [0.34, 0.68] 7.77% 0.61 [0.48, 0.77]
Study 5 41 435 97 734 6.35% 0.68 [0.46, 1.01] 6.06% 0.70 [0.48, 1.01]
Study 6 14 81 27 75 3.78% 0.37 [0.18, 0.78] 3.32% 0.43 [0.22, 0.81]
Study 7 34 91 104 219 5.43% 0.66 [0.40, 1.09] 5.73% 0.73 [0.49, 1.07]
Study 8 110 542 142 443 7.14% 0.54 [0.40, 0.72] 7.61% 0.59 [0.46, 0.75]
Study 9 9 45 14 32 2.58% 0.32 [0.12, 0.88] 2.28% 0.39 [0.17, 0.90]
Study 10 223 1490 588 2675 7.96% 0.62 [0.53, 0.74] 8.85% 0.65 [0.56, 0.76]
Study 11 64 396 29 136 5.51% 0.71 [0.44, 1.16] 5.17% 0.74 [0.47, 1.14]
Study 12 80 217 129 157 5.49% 0.13 [0.08, 0.21] 7.00% 0.27 [0.20, 0.36]
Study 13 20 115 31 102 4.42% 0.48 [0.25, 0.92] 3.96% 0.53 [0.30, 0.93]
Study 14 9 52 30 67 3.17% 0.26 [0.11, 0.61] 2.72% 0.32 [0.15, 0.68]
Study 15 14 173 37 220 4.35% 0.44 [0.23, 0.83] 3.55% 0.46 [0.25, 0.85]
Study 16 27 665 41 521 5.44% 0.50 [0.30, 0.82] 4.68% 0.51 [0.31, 0.82]
Study 17 5 59 7 28 1.88% 0.28 [0.08, 0.97] 1.37% 0.31 [0.10, 0.97]
Study 18 136 455 205 555 7.33% 0.73 [0.56, 0.95] 8.05% 0.77 [0.62, 0.96]
Study 19 56 173 24 52 4.48% 0.56 [0.30, 1.05] 4.69% 0.63 [0.39, 1.03]
Study 20 31 83 40 74 4.44% 0.51 [0.27, 0.96] 4.77% 0.60 [0.37, 0.97]
OR REML (Q = 53.1, df = 19, p = 0.0; I² = 70%) 100.00% 0.50 [0.41, 0.61]
Clog REML (Q = 47.9, df = 19, p = 0.0; I² = 62%) 100.00% 0.57 [0.49, 0.66]
0.05 0.25 1 4
0.05 0.25 1 4

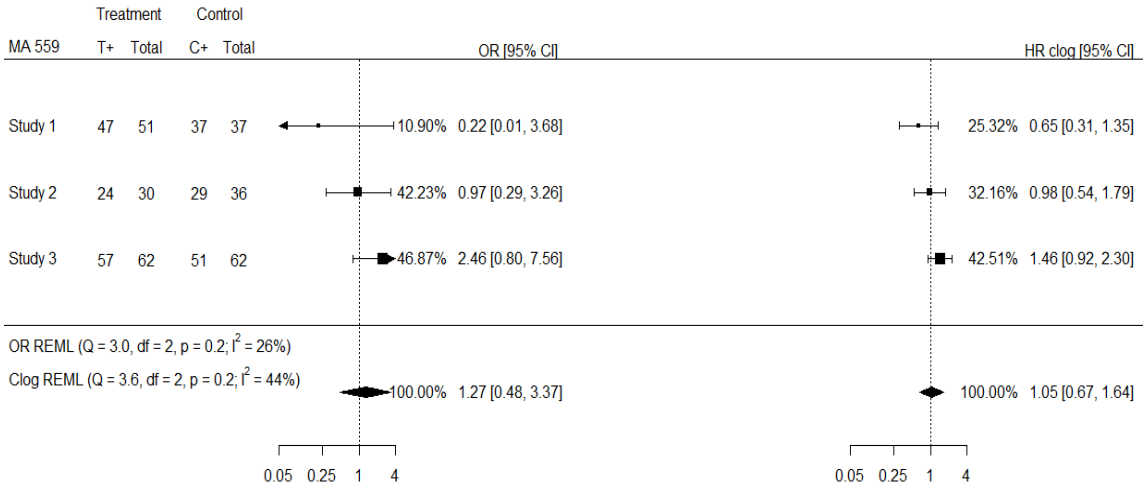

MA 434
Treatment T+ Total
Control C+ Total
OR [95% CI]
HR clog [95% CI]
Study 1 5 9 6 7 13.34% 0.21 [0.02, 2.52] 19.49% 0.42 [0.11, 1.52]
Study 2 0 5 2 5 7.09% 0.26 [0.01, 8.52] 3.23% 0.30 [0.01, 10.64]
Study 3 19 20 16 20 15.55% 4.75 [0.48, 46.91] 34.87% 1.86 [0.80, 4.31]
Study 4 18 37 15 38 64.03% 1.45 [0.58, 3.63] 42.41% 1.33 [0.66, 2.66]
OR REML (Q = 4.2, df = 3, p = 0.2; I² = 12%) 100.00% 1.19 [0.46, 3.09]
Clog REML (Q = 4.3, df = 3, p = 0.2; I² = 31%) 100.00% 1.14 [0.59, 2.19]
0.05 0.25 1 4
0.05 0.25 1 4

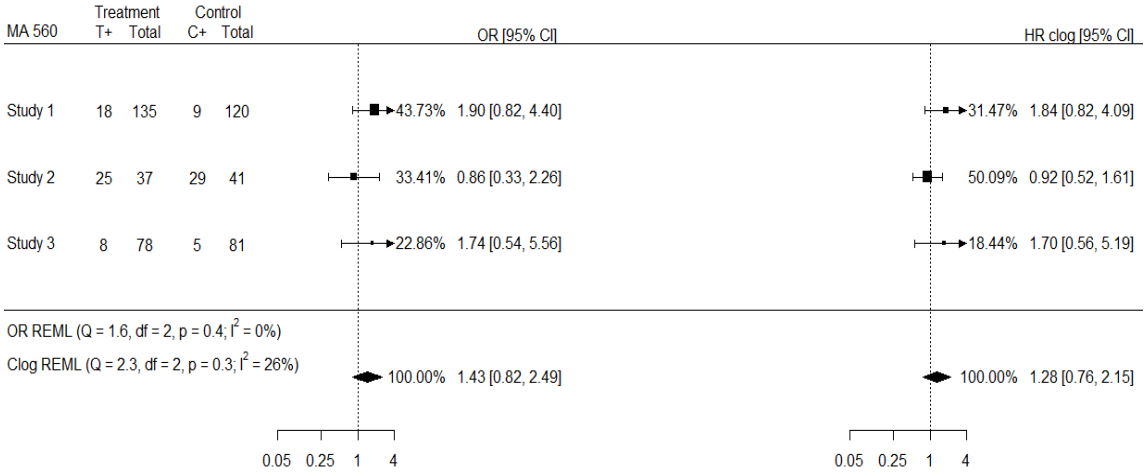

MA 448
Treatment T+ Total
Control C+ Total
OR [95% CI]
HR clog [95% CI]
Study 1 55 89 64 91 26.54% 0.68 [0.37, 1.27] 24.62% 0.79 [0.54, 1.16]
Study 2 86 101 71 83 16.65% 0.97 [0.43, 2.20] 26.47% 0.99 [0.68, 1.42]
Study 3 26 112 21 111 24.86% 1.30 [0.68, 2.47] 10.61% 1.26 [0.71, 2.24]
Study 4 38 48 37 45 11.05% 0.82 [0.29, 2.31] 13.77% 0.91 [0.55, 1.51]
Study 5 20 32 25 29 7.49% 0.27 [0.07, 0.95] 8.41% 0.50 [0.26, 0.95]
Study 6 41 56 45 54 13.42% 0.55 [0.22, 1.38] 16.11% 0.74 [0.46, 1.17]
OR REML (Q = 6.1, df = 5, p = 0.3; I² = 13%) 100.00% 0.78 [0.55, 1.12]
Clog REML (Q = 5.7, df = 5, p = 0.3; I² = 0%) 100.00% 0.85 [0.71, 1.03]
0.05 0.25 1 4
0.05 0.25 1 4

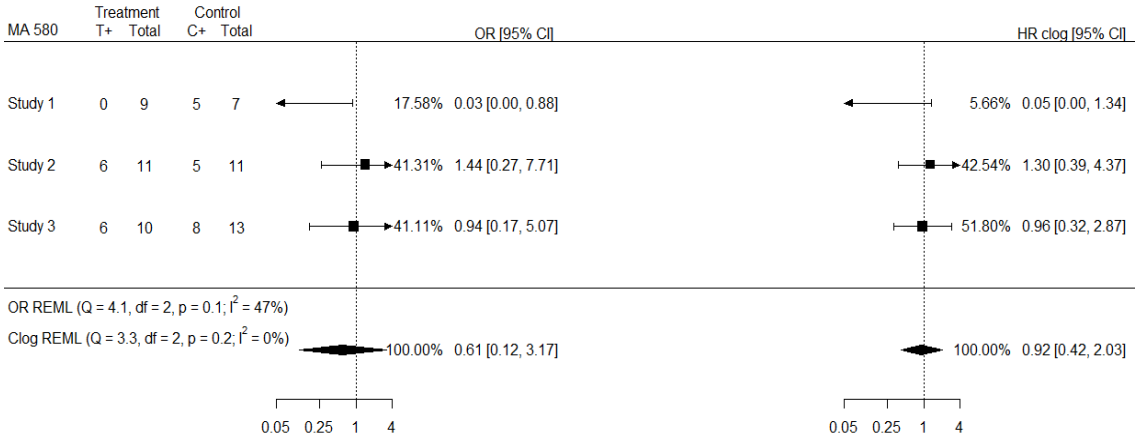

MA 506
Treatment T+ Total
Control C+ Total
OR [95% CI]
HR clog [95% CI]
Study 1 4 5 2 5 3.04% 6.00 [0.35, 101.57] 4.57% 3.15 [0.53, 18.58]
Study 2 1 12 6 12 4.32% 0.09 [0.01, 0.94] 3.36% 0.13 [0.02, 1.05]
Study 3 2 5 1 5 3.04% 2.67 [0.16, 45.14] 2.68% 2.29 [0.21, 25.55]
Study 4 4 10 4 10 6.93% 1.00 [0.17, 5.98] 6.67% 1.00 [0.25, 4.06]
Study 5 14 15 9 13 4.30% 6.22 [0.60, 64.97] 10.85% 2.30 [0.86, 6.14]
Study 6 19 29 20 33 15.94% 1.23 [0.44, 3.48] 16.41% 1.14 [0.59, 2.20]
Study 7 19 64 32 60 23.57% 0.37 [0.18, 0.77] 18.10% 0.46 [0.26, 0.82]
Study 8 3 6 4 5 3.29% 0.25 [0.02, 3.77] 5.49% 0.43 [0.09, 2.11]
Study 9 20 73 30 74 25.04% 0.55 [0.28, 1.11] 18.21% 0.62 [0.35, 1.09]
Study 10 6 9 6 6 2.69% 0.29 [0.01, 6.03] 6.42% 0.56 [0.13, 2.37]
Study 11 4 12 5 12 7.85% 0.70 [0.13, 3.68] 7.24% 0.75 [0.20, 2.84]
OR REML (Q = 14.4, df = 10, p = 0.2; I² = 22%) 100.00% 0.67 [0.40, 1.12]
Clog REML (Q = 16.6, df = 10, p = 0.1; I² = 38%) 100.00% 0.81 [0.53, 1.22]
0.05 0.25 1 4
0.05 0.25 1 4

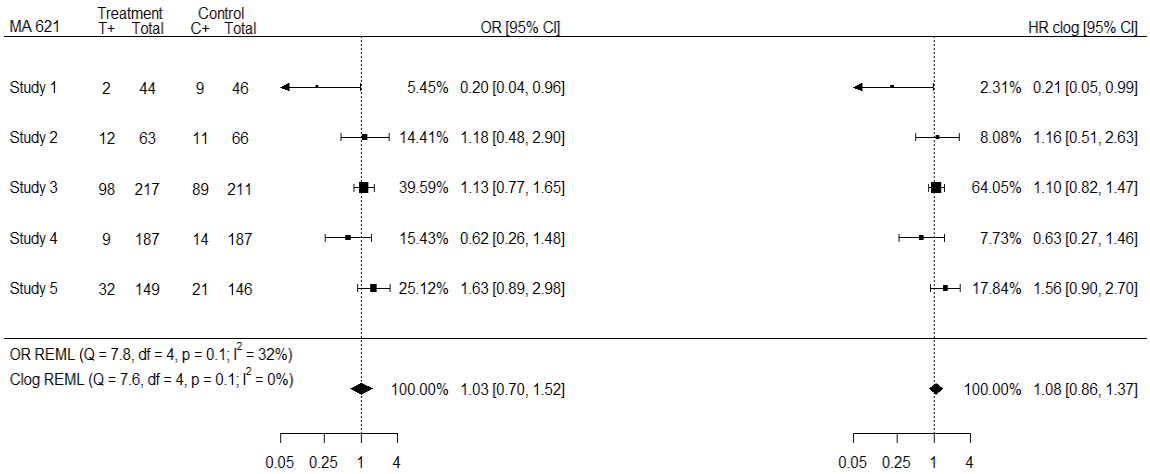

MA 507
Treatment T+ Total
Control C+ Total
OR [95% CI]
HR clog [95% CI]
Study 1 6 9 6 6 1.06% 0.29 [0.01, 6.03] 2.04% 0.56 [0.13, 2.37]
Study 2 4 5 2 5 1.20% 6.00 [0.35, 101.57] 1.39% 3.15 [0.53, 18.58]
Study 3 14 15 9 13 1.70% 6.22 [0.60, 64.97] 3.87% 2.30 [0.86, 6.14]
Study 4 24 51 20 39 8.55% 0.84 [0.37, 1.95] 7.60% 0.88 [0.48, 1.62]
Study 5 99 136 45 53 8.47% 0.48 [0.21, 1.10] 11.37% 0.69 [0.46, 1.03]
Study 6 47 76 45 53 7.99% 0.29 [0.12, 0.70] 10.29% 0.51 [0.32, 0.80]
Study 7 19 64 32 60 9.80% 0.37 [0.18, 0.77] 8.06% 0.46 [0.26, 0.82]
Study 8 3 6 4 5 1.29% 0.25 [0.02, 3.77] 1.71% 0.43 [0.09, 2.11]
Study 9 4 12 5 12 3.13% 0.70 [0.13, 3.68] 2.35% 0.75 [0.20, 2.84]
Study 10 1 5 1 5 1.01% 1.00 [0.05, 22.18] 0.60% 1.00 [0.06, 16.08]
Study 11 1 12 6 12 1.71% 0.09 [0.01, 0.94] 1.00% 0.13 [0.02, 1.05]
Study 12 22 59 41 67 10.04% 0.38 [0.18, 0.78] 8.83% 0.49 [0.29, 0.84]
Study 13 3 6 3 4 1.24% 0.33 [0.02, 5.33] 1.54% 0.50 [0.09, 2.69]
Study 14 10 45 21 41 7.47% 0.27 [0.11, 0.69] 5.68% 0.35 [0.16, 0.75]
Study 15 4 10 4 10 2.75% 1.00 [0.17, 5.98] 2.14% 1.00 [0.25, 4.06]
Study 16 19 29 20 33 6.49% 1.23 [0.44, 3.48] 6.91% 1.14 [0.59, 2.20]
Study 17 22 41 95 117 9.38% 0.27 [0.12, 0.58] 9.63% 0.46 [0.28, 0.75]
Study 18 11 22 82 92 6.27% 0.12 [0.04, 0.35] 6.86% 0.31 [0.16, 0.60]
Study 19 20 73 30 74 10.46% 0.55 [0.28, 1.11] 8.14% 0.62 [0.35, 1.09]
OR REML (Q = 27.6, df = 18, p = 0.1; I² = 29%) 100.00% 0.44 [0.32, 0.61]
Clog REML (Q = 28.0, df = 18, p = 0.1; I² = 33%) 100.00% 0.60 [0.48, 0.75]
0.05 0.25 1 4
0.05 0.25 1 4

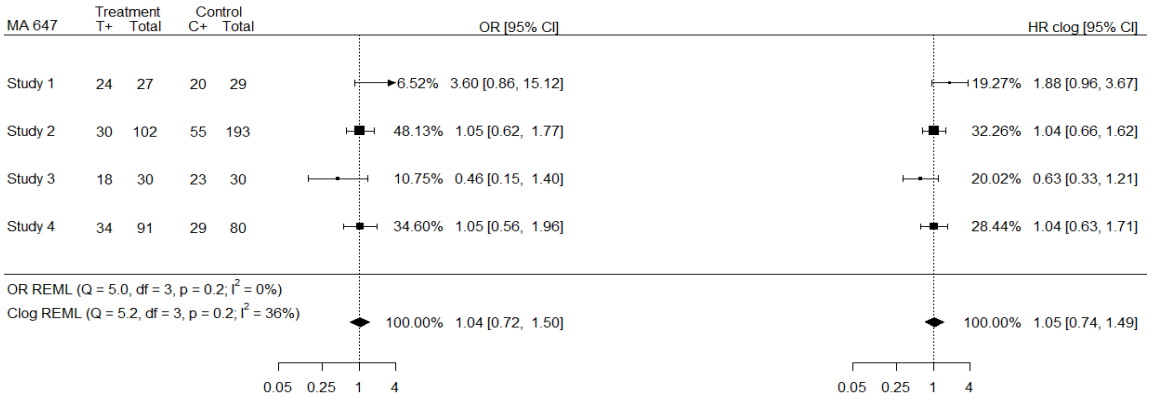

MA 558
Treatment T+ Total
Control C+ Total
OR [95% CI]
HR clog [95% CI]
Study 1 40 51 37 37 6.20% 0.06 [0.00, 0.90] 21.23% 0.38 [0.19, 0.78]
Study 2 18 30 21 36 26.47% 1.07 [0.40, 2.87] 22.55% 1.05 [0.55, 2.01]
Study 3 44 62 41 62 33.43% 1.25 [0.59, 2.68] 27.01% 1.14 [0.73, 1.79]
Study 4 142 158 60 77 33.89% 2.51 [1.19, 5.30] 29.20% 1.52 [1.07, 2.14]
OR REML (Q = 7.9, df = 3, p = 0.0; I² = 49%) 100.00% 1.26 [0.62, 2.57]
Clog REML (Q = 11.8, df = 3, p = 0.0; I² = 78%) 100.00% 0.96 [0.56, 1.67]
0.05 0.25 1 4
0.05 0.25 1 4

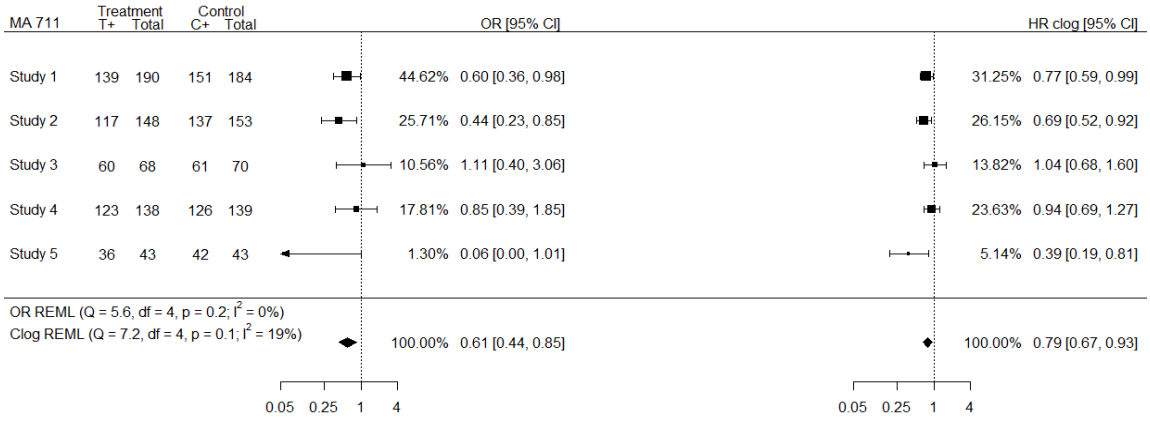

MA 559
Treatment T+ Total
Control C+ Total
OR [95% CI]
HR clog [95% CI]
Study 1 47 51 37 37 10.90% 0.22 [0.01, 3.68] 25.32% 0.65 [0.31, 1.35]
Study 2 24 30 29 36 42.23% 0.97 [0.29, 3.26] 32.16% 0.98 [0.54, 1.79]
Study 3 57 62 51 62 46.87% 2.46 [0.80, 7.56] 42.51% 1.46 [0.92, 2.30]
OR REML (Q = 3.0, df = 2, p = 0.2; I² = 26%) 100.00% 1.27 [0.48, 3.37]
Clog REML (Q = 3.6, df = 2, p = 0.2; I² = 44%) 100.00% 1.05 [0.67, 1.64]
0.05 0.25 1 4
0.05 0.25 1 4


MA 560
Treatment T+ Total
Control C+ Total
OR [95% CI]
HR clog [95% CI]
Study 1 18 135 9 120 43.73% 1.90 [0.82, 4.40] 31.47% 1.84 [0.82, 4.09]
Study 2 25 37 29 41 33.41% 0.86 [0.33, 2.26] 50.09% 0.92 [0.52, 1.61]
Study 3 8 78 5 81 22.86% 1.74 [0.54, 5.56] 18.44% 1.70 [0.56, 5.19]
OR REML (Q = 1.6, df = 2, p = 0.4; I² = 0%) 100.00% 1.43 [0.82, 2.49]
Clog REML (Q = 2.3, df = 2, p = 0.3; I² = 26%) 100.00% 1.28 [0.76, 2.15]
0.05 0.25 1 4
0.05 0.25 1 4


MA 580
Treatment T+ Total
Control C+ Total
OR [95% CI]
HR clog [95% CI]
Study 1 0 9 5 7 17.58% 0.03 [0.00, 0.88] 5.66% 0.05 [0.00, 1.34]
Study 2 6 11 5 11 41.31% 1.44 [0.27, 7.71] 42.54% 1.30 [0.39, 4.37]
Study 3 6 10 8 13 41.11% 0.94 [0.17, 5.07] 51.80% 0.96 [0.32, 2.87]
OR REML (Q = 4.1, df = 2, p = 0.1; I² = 47%) 100.00% 0.61 [0.12, 3.17]
Clog REML (Q = 3.3, df = 2, p = 0.2; I² = 0%) 100.00% 0.92 [0.42, 2.03]
0.05 0.25 1 4
0.05 0.25 1 4


MA 621
Treatment T+ Total
Control C+ Total
OR [95% CI]
HR clog [95% CI]
Study 1 2 44 9 46 5.45% 0.20 [0.04, 0.96] 2.31% 0.21 [0.05, 0.99]
Study 2 12 63 11 66 14.41% 1.18 [0.48, 2.90] 8.08% 1.16 [0.51, 2.63]
Study 3 98 217 89 211 39.59% 1.13 [0.77, 1.65] 64.05% 1.10 [0.82, 1.47]
Study 4 9 187 14 187 15.43% 0.62 [0.26, 1.48] 7.73% 0.63 [0.27, 1.46]
Study 5 32 149 21 146 25.12% 1.63 [0.89, 2.98] 17.84% 1.56 [0.90, 2.70]
OR REML (Q = 7.8, df = 4, p = 0.1; I² = 32%) 100.00% 1.03 [0.70, 1.52]
Clog REML (Q = 7.6, df = 4, p = 0.1; I² = 0%) 100.00% 1.08 [0.86, 1.37]
0.05 0.25 1 4
0.05 0.25 1 4


MA 647
Treatment
T+ Total
Control
C+ Total
OR [95% CI]
HR clog [95% CI]
Study 1 24 27 20 29 6.52% 3.60 [0.86, 15.12] 19.27% 1.88 [0.96, 3.67]
Study 2 30 102 55 193 48.13% 1.05 [0.62, 1.77] 32.26% 1.04 [0.66, 1.62]
Study 3 18 30 23 30 10.75% 0.46 [0.15, 1.40] 20.02% 0.63 [0.33, 1.21]
Study 4 34 91 29 80 34.60% 1.05 [0.56, 1.96] 28.44% 1.04 [0.63, 1.71]
OR REML (Q = 5.0, df = 3, p = 0.2; I² = 0%) 100.00% 1.04 [0.72, 1.50]
Clog REML (Q = 5.2, df = 3, p = 0.2; I² = 36%) 100.00% 1.05 [0.74, 1.49]
0.05 0.25 1 4
0.05 0.25 1 4


MA 711
Treatment
T+ Total
Control
C+ Total
OR [95% CI]
HR clog [95% CI]
Study 1 139 190 151 184 44.62% 0.60 [0.36, 0.98] 31.25% 0.77 [0.59, 0.99]
Study 2 117 148 137 153 25.71% 0.44 [0.23, 0.85] 26.15% 0.69 [0.52, 0.92]
Study 3 60 68 61 70 10.56% 1.11 [0.40, 3.06] 13.82% 1.04 [0.68, 1.60]
Study 4 123 138 126 139 17.81% 0.85 [0.39, 1.85] 23.63% 0.94 [0.69, 1.27]
Study 5 36 43 42 43 1.30% 0.06 [0.00, 1.01] 5.14% 0.39 [0.19, 0.81]
OR REML (Q = 5.6, df = 4, p = 0.2; I² = 0%) 100.00% 0.61 [0.44, 0.85]
Clog REML (Q = 7.2, df = 4, p = 0.1; I² = 19%) 100.00% 0.79 [0.67, 0.93]
0.05 0.25 1 4
0.05 0.25 1 4

C – Additional material relating to the “OEV” data meta-analyses analysed in Chapter 4

C.1 – Bland-Altman plots for IPD, Non IPD and baseline risk

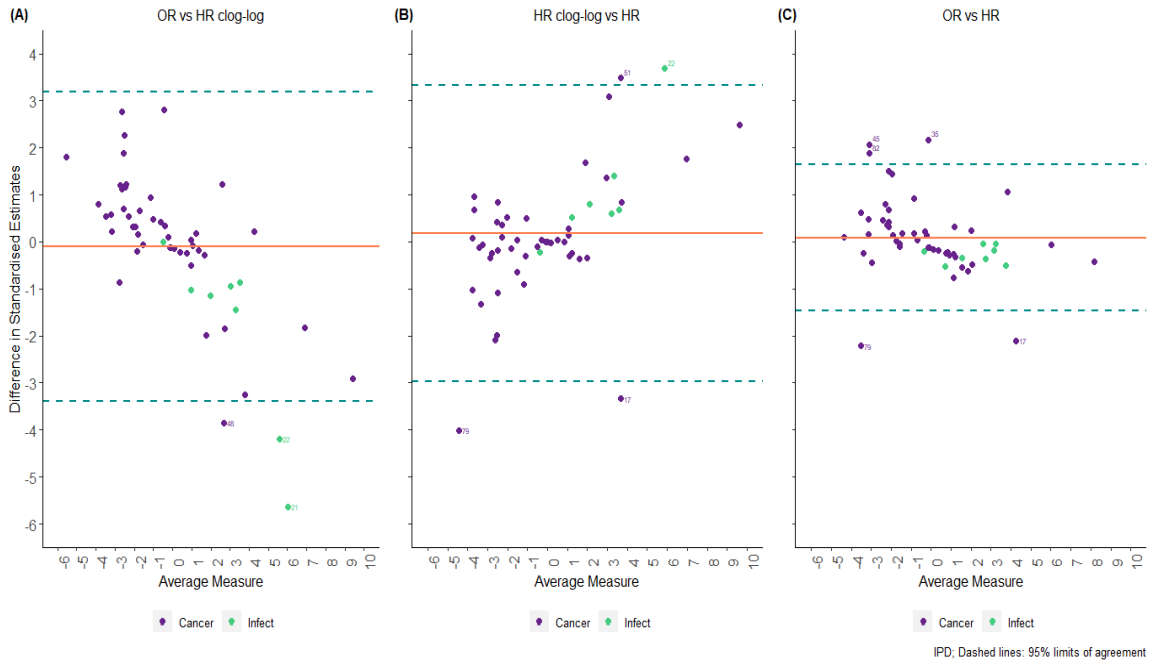


Figure 4.9: IPD - Bland-Altman Plot comparing standardised OR vs. HR estimates for two-stage models in “OEV” data.

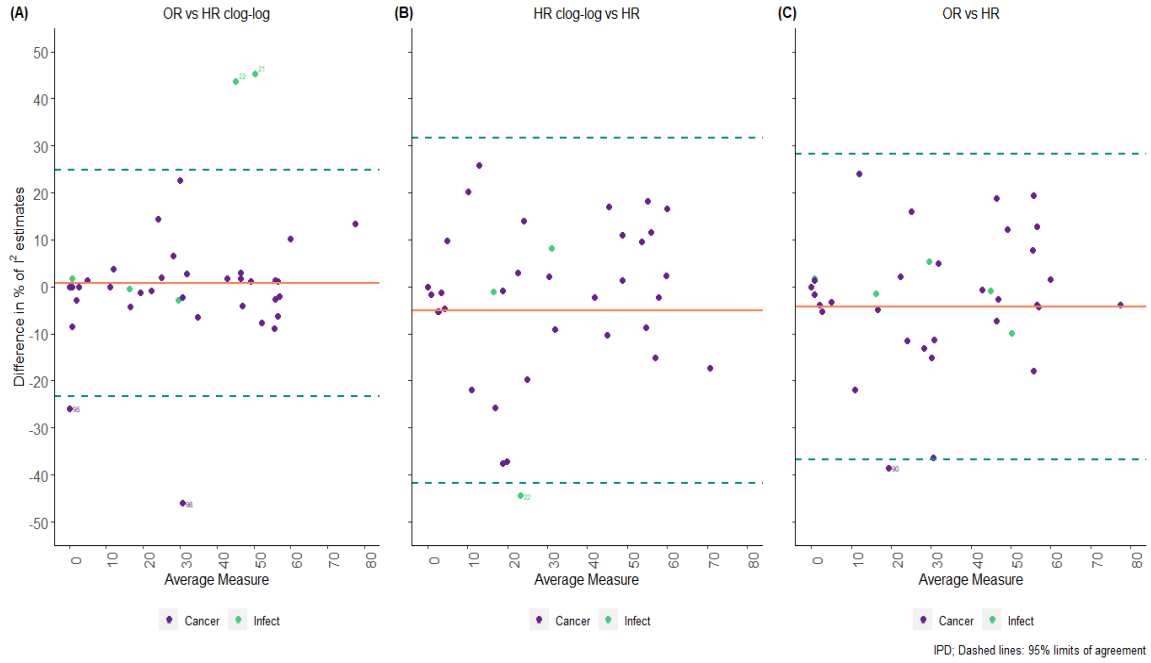


Figure 4.10: IPD - Bland-Altman Plot comparing I^2 estimates (OR vs. HR) for two-stage models in “OEV” data.

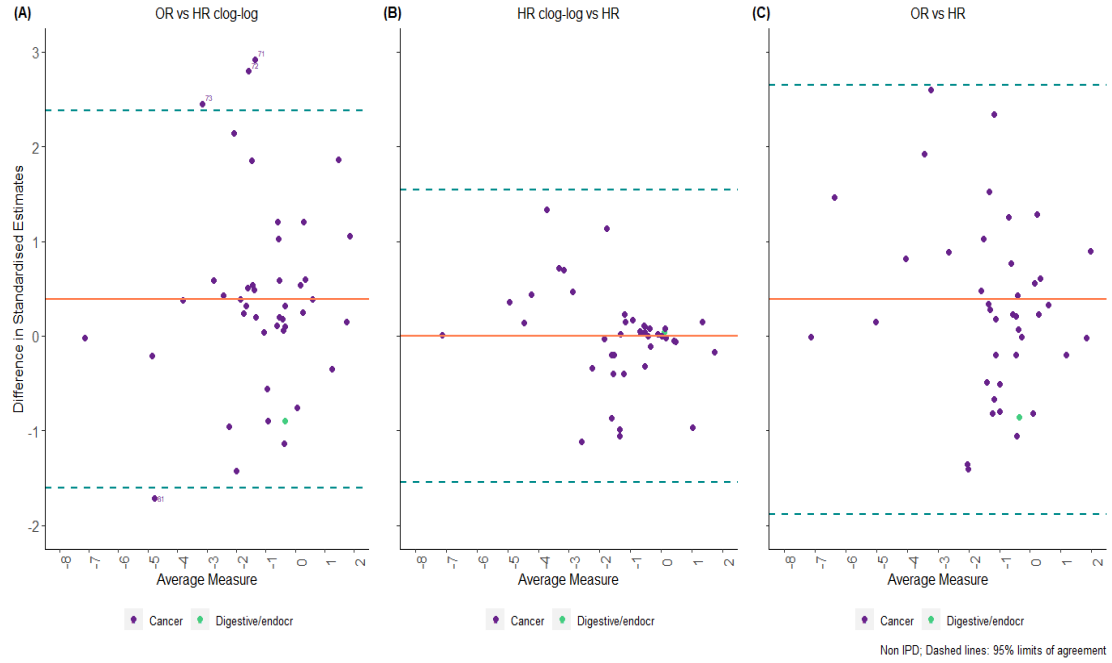


Figure 4.11: Non IPD - Bland-Altman Plot comparing standardised OR vs. HR estimates for two-stage models in “OEV” data.

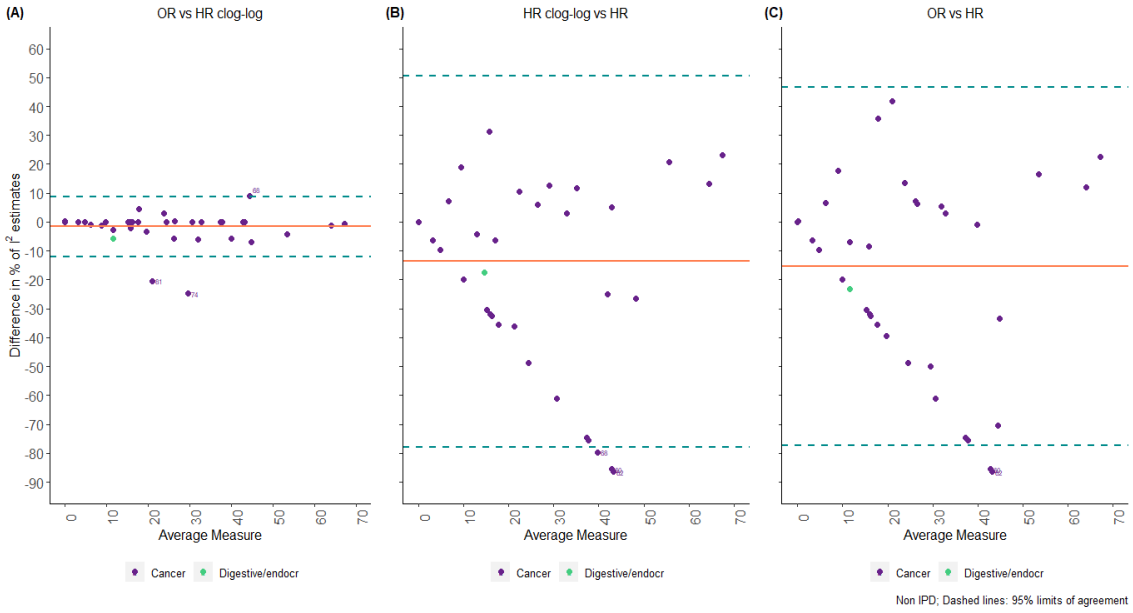


Figure 4.12: Non IPD - Bland-Altman Plot comparing I^2 estimates (OR vs. HR) for two-stage models in “OEV” data.

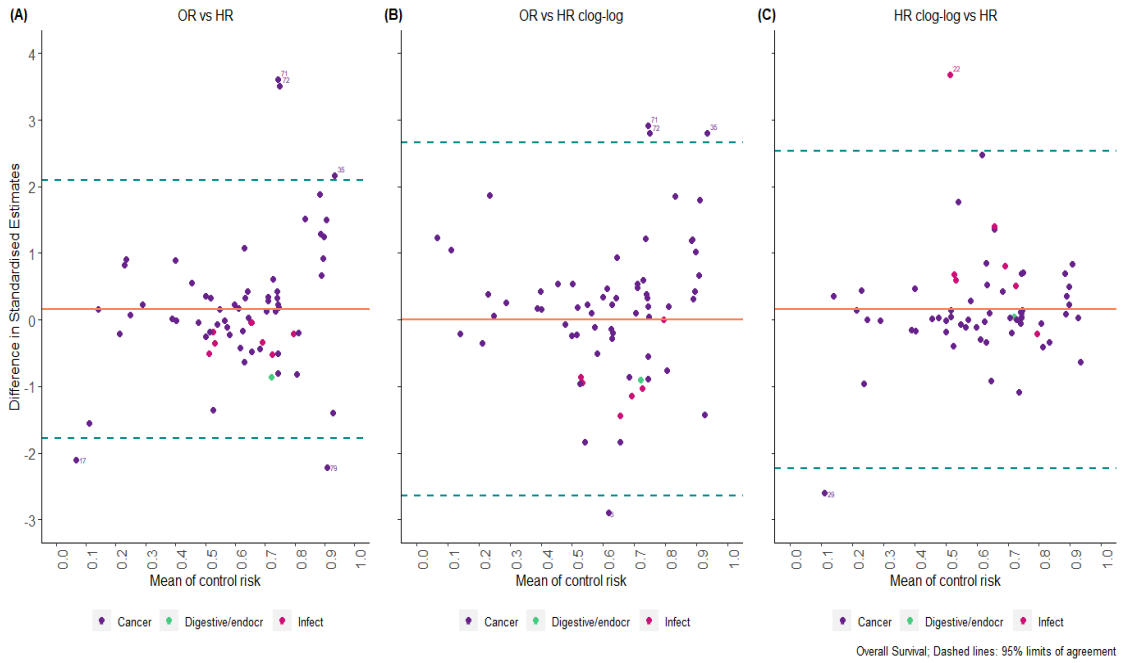


Figure 4.13: Overall Survival – Bland-Altman plot examining the association between the difference in the scales to baseline risk.

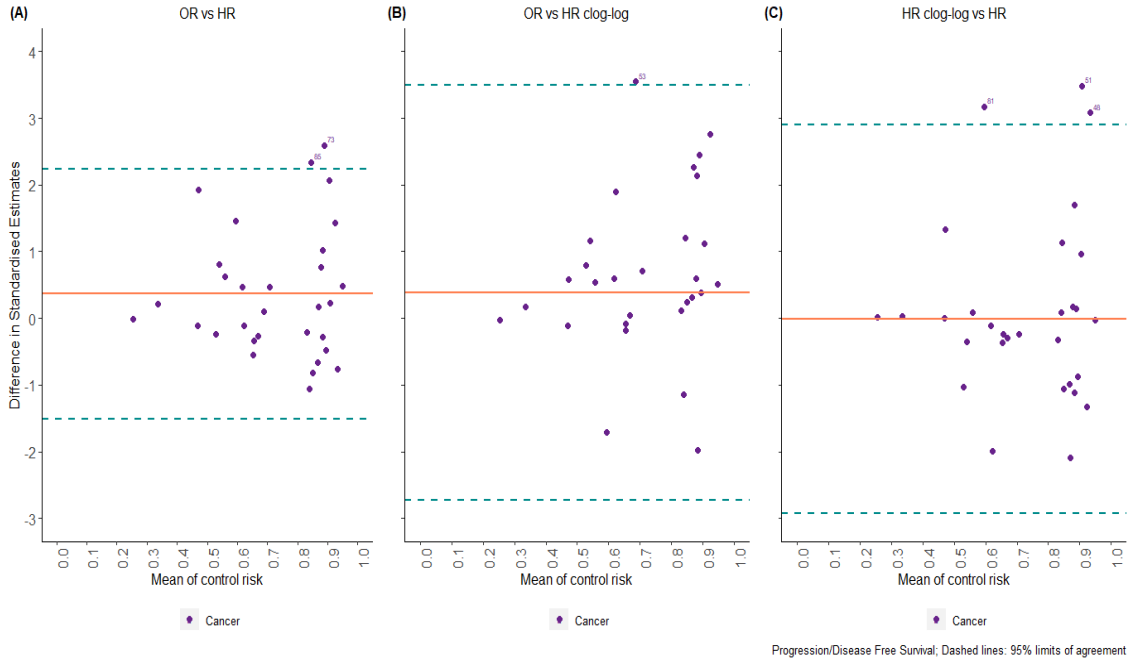


Figure 4.14: Progression/Disease Free Survival – Bland-Altman plot examining the association between the difference in the scales to baseline risk.

C.2 – Model Implementation

Two-stage MA for ORs

```
for (i in unique(CDSR_2008$ma)) {
  cat(i,"\n")
  try.fit<- try(rma.uni(ai = treat_n, bi = nontreat_n, ci = ctrl_n, di = nonctrl_n,
                        data = CDSR_2008[CDSR_2008$ma==i,], measure =
"OR",method="REML",
                        control=list(maxiter=500, verbose=TRUE, stepadj=0.5),
verbose=TRUE))
  resultsREML[i,7]<-i
  resultsREML[i,8]<-unique(CDSR_2008[CDSR_2008$ma==i,]$scode)
  if (class(try.fit)!="try-error") {
    CDSR.2008[[i]]<- try.fit
    resultsREML[i,1]<-as.numeric(exp(CDSR.2008[[i]]$b))
    resultsREML[i,2]<-as.numeric(CDSR.2008[[i]]$se)
    resultsREML[i,3]<-as.numeric(exp(CDSR.2008[[i]]$ci.lb))
    resultsREML[i,4]<-as.numeric(exp(CDSR.2008[[i]]$ci.ub))
    resultsREML[i,5]<-as.numeric(CDSR.2008[[i]]$tau2)
    resultsREML[i,6]<-as.numeric(CDSR.2008[[i]]$I2)
  } else {
    CDSR.2008[[i]] <- NULL }}
```

Two-stage MA for HRs

```
CDSR_2008$logHR<-CDSR_2008$o_e/CDSR_2008$variance   # Log HR
calculation
CDSR_2008$varHR<-1/CDSR_2008$variance                    # Variance
HR calculation
for (i in unique(CDSR_2008$ma)) {
  cat(i,"\n")
  try.fit1<- try(rma.uni(yi = logHR, vi = varHR, data =
CDSR_2008[CDSR_2008$ma==i,],
                         method="REML",control=list(maxiter=10e9, verbose=TRUE,
stepadj=0.2), verbose=TRUE))
  resultsREMLHR[i,7]<-i
  resultsREMLHR[i,8]<-unique(CDSR_2008[CDSR_2008$ma==i,]$scode)
```

```
 if (class(try.fit1)!="try-error") {
   CDSR.2008HR[[i]]<- try.fit1
   resultsREMLHR[i,1]<-as.numeric(exp(CDSR.2008HR[[i]]$b))
   resultsREMLHR[i,2]<-as.numeric(CDSR.2008HR[[i]]$se)
   resultsREMLHR[i,3]<-as.numeric(exp(CDSR.2008HR[[i]]$ci.lb))
   resultsREMLHR[i,4]<-as.numeric(exp(CDSR.2008HR[[i]]$ci.ub))
   resultsREMLHR[i,5]<-as.numeric(CDSR.2008HR[[i]]$tau2)
   resultsREMLHR[i,6]<-as.numeric(CDSR.2008HR[[i]]$I2)
 } else {
   CDSR.2008HR[[i]] <- NULL }}
```

Two-stage MA for HRs using the Clog-log link

A) Calculation of HR clog-log and corresponding variance

```
CDSR_2008$proptreat<-
CDSR_2008$treat_n/(CDSR_2008$treat_n+CDSR_2008$nontreat_n)

CDSR_2008$propctrl<-
CDSR_2008$ctrl_n/(CDSR_2008$ctrl_n+CDSR_2008$nonctrl_n)

CDSR_2008$logHRclog<-(log(-log(1-CDSR_2008$proptreat)))-log(-log(1-
CDSR_2008$propctrl))

CDSR_2008$derivTreat<-1/((log(1-
CDSR_2008$proptreat))*(CDSR_2008$proptreat-1))

CDSR_2008$derivCtrl<-1/((log(1-
CDSR_2008$propctrl))*(CDSR_2008$propctrl-1))

CDSR_2008$varTreat<-
(CDSR_2008$derivTreat^2)*((CDSR_2008$proptreat*(1-
CDSR_2008$proptreat))/CDSR_2008$treat_total)

CDSR_2008$varCtrl<-(CDSR_2008$derivCtrl^2)*((CDSR_2008$propctrl*(1-
CDSR_2008$propctrl))/CDSR_2008$ctrl_total)

CDSR_2008$varHRclog<-CDSR_2008$varTreat+CDSR_2008$varCtrl
```

B) Model Implementation

```
CDSR.2008HRclog = list()

resREMLclogHR=data.frame(matrix(NA, max(CDSR_2008$ma), 8))

colnames(resREMLclogHR)<-c("estimates_HRclog","SE_HRclog",
"LowerCI_HRclog", "UpperCI_HRclog","Tau_HRclog", "Isq_HRclog", "MA",
"Med_Area")


for (i in unique(CDSR_2008$ma)) {
```

```
  cat(i,"\n")
  try.fit2<- try(rma.uni(yi = logHRclog, vi = varHRclog, data =
CDSR_2008[CDSR_2008$ma==i,],
                         method="REML",control=list(maxiter=10e9, verbose=TRUE,
stepadj=0.2), verbose=TRUE))
  resREMLclogHR[i,7]<-i
  resREMLclogHR[i,8]<-unique(CDSR_2008[CDSR_2008$ma==i,]$scode)
  if (class(try.fit1)!="try-error") {
    CDSR.2008HRclog[[i]]<- try.fit2
    resREMLclogHR[i,1]<-as.numeric(exp(CDSR.2008HRclog[[i]]$b))
    resREMLclogHR[i,2]<-as.numeric(CDSR.2008HRclog[[i]]$se)
    resREMLclogHR[i,3]<-as.numeric(exp(CDSR.2008HRclog[[i]]$ci.lb))
    resREMLclogHR[i,4]<-as.numeric(exp(CDSR.2008HRclog[[i]]$ci.ub))
    resREMLclogHR[i,5]<-as.numeric(CDSR.2008HRclog[[i]]$tau2)
    resREMLclogHR[i,6]<-as.numeric(CDSR.2008HRclog[[i]]$I2)
  } else {
    CDSR.2008HRclog[[i]] <- NULL }}
```

C.3 – Table containing the exact results from the two-stage meta-analysis models & additional forest plots considered as outliers from the Bland-Altman plots

Two-Stage Random-Effects Model – Overall Survival				
MA Identifier	OR (95% CI) vs. HR (95% CI)	τ^2 OR vs. τ^2 HR	I^2 OR vs. I^2 HR	IPD (Yes/No)
03	1.206 (1.152, 1.263) vs 1.172 (1.129, 1.216)	0.000 vs. 0.000	0% vs. 3%	Yes
17	2.144 (1.347, 3.410) vs. 1.535 (1.311, 1.797)	0.000 vs. 0.002	0% vs. 5%	Yes
21	1.482 (1.166, 1.885) vs. 1.433 (1.154, 1.779)	0.003 vs. 0.003	45% vs. 0%	Yes
22	1.464 (1.182, 1.812) vs. 1.382 (1.180, 1.619)	0.003 vs. 0.001	45% vs. 1%	Yes

29	1.841 (1.112, 3.050) vs. 1.349 (1.162, 1.567)	0.204 vs. 0.011	31% vs. 28%	No
35	1.470 (0.663, 3.263) vs. 0.785 (0.533, 1.157)	0.221 vs. 0.224	19% vs. 85%	Yes
42	1.075 (0.748, 1.546) vs. 1.022 (0.781, 1.337)	0.000 vs. 0.050	0% vs. 49%	No
71	1.009 (0.848, 1.201) vs. 0.868 (0.802, 0.939)	0.068 vs. 0.012	35% vs. 29%	No
72	0.986 (0.843, 1.153) vs. 0.875 (0.815, 0.939)	0.055 vs. 0.009	30% vs. 23%	No
74	1.088 (0.837, 1.413) vs. 1.003 (0.818, 1.228)	0.007 vs. 0.045	5% vs. 55%	No
79	0.753 (0.668, 0.849) vs. 0.854 (0.752, 0.970)	0.019 vs. 0.030	25% vs. 36%	Yes
82	0.771 (0.609, 0.977) vs. 0.849 (0.784, 0.919)	0.000 vs. 0.000	0% vs. 0%	Yes
87	1.110 (0.694, 1.773) vs. 0.982 (0.744, 1.297)	0.061 vs. 0.000	36% vs. 0%	No
95	0.857 (0.735, 1.000) vs. 0.895 (0.813, 0.984)	0.000 vs. 0.000	0% vs. 26%	Yes
96	0.748 (0.627, 0.894) vs. 0.821 (0.821, 0.944)	0.010 vs. 0.023	12% vs. 58%	Yes

Table 4.4: Characteristics of meta-analyses outside the 95% limits of agreement based on difference of standardised estimates and difference in I^2.

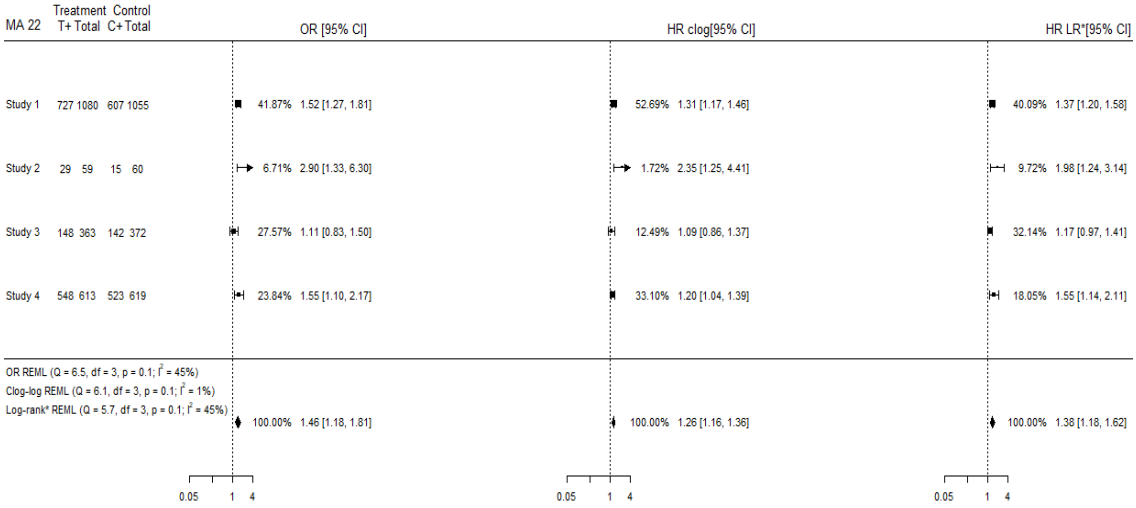
MA coloured in **blue** represent characteristics of studies outside the 95% limits of agreement based on difference of standardised estimates. MA coloured in **red** represent characteristics of studies outside the 95% limits of agreement based on difference in I^2. MA coloured in **black** represent characteristics of studies outside the 95% limits of agreement based on difference of standardised estimates and difference in I^2.

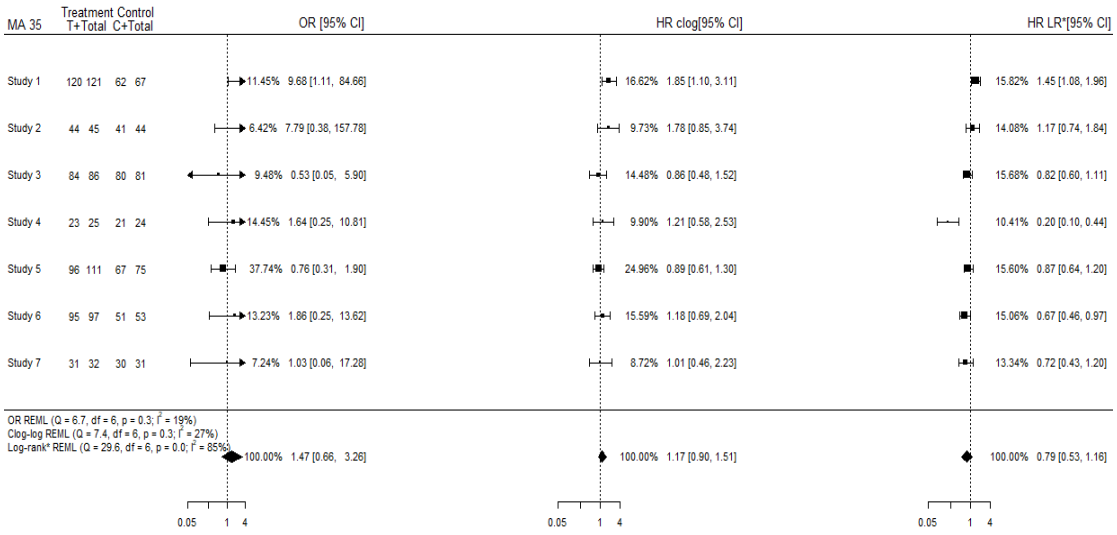
Two-Stage Random-Effects Model – Progression/disease free survival				
MA Identifier	OR (95% CI) vs. HR (95% CI)	τ^2 OR vs. τ^2 HR	I^2 OR vs. I^2 HR	IPD (Yes/No)
45	0.655 (0.440, 0.975) vs. 0.747 (0.651, 0.857)	0.005 vs. 0.000	1% vs. 0%	Yes
48	1.465 (0.548, 3.914) vs. 2.120 (0.809, 5.554)	0.000 vs. 0.000	0% vs. 0%	Yes
51	1.465 (1.034, 2.076) vs. 1.374 (0.992, 1.903)	0.000 vs. 0.000	0% vs. 0%	Yes
53	0.438 (0.301, 0.638) vs. 0.503 (0.371, 0.682)	0.373 vs. 0.267	76% vs. 62%	Yes
56	0.810 (0.627, 1.045) vs. 0.842 (0.674, 1.051)	0.036 vs. 0.044	23% vs. 0%	Yes
60	0.856 (0.699, 1.047) vs. 0.919 (0.756, 1.117)	0.000 vs. 0.100	0% vs. 85%	No
62	0.865 (0.726, 1.031) vs. 0.923 (0.760, 1.120)	0.000 vs. 0.099	0% vs. 86%	No
68	0.847 (0.478, 1.502) vs. 0.921 (0.592, 1.434)	0.025 vs. 0.120	9% vs. 0%	No
73	0.778 (0.602, 1.004) vs. 0.785 (0.707, 0.872)	0.092 vs. 0.033	28% vs. 61%	No
81	0.463 (0.354, 0.605) vs. 0.624 (0.548, 0.711)	0.052 vs. 0.000	52% vs. 62%	No
83	0.805 (0.573, 1.129) vs. 0.767 (0.632, 0.931)	0.000 vs. 0.043	0% vs. 70%	Yes
85	0.996 (0.396, 2.510) vs. 0.801 (0.665, 0.964)	0.684 vs. 0.022	62% vs. 45%	No
90	0.723 (0.603, 0.868) vs. 0.758 (0.641, 0.895)	0.000 vs. 0.028	0% vs. 39%	Yes

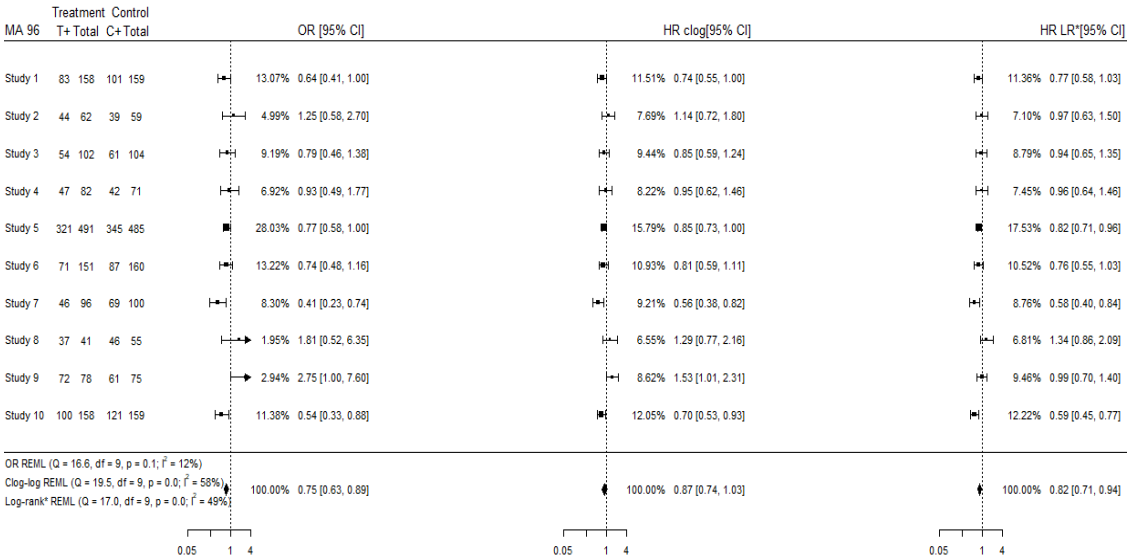
Table 4.5: Characteristics of meta-analyses outside the 95% limits of agreement based on difference of standardised estimates and difference in I^2.

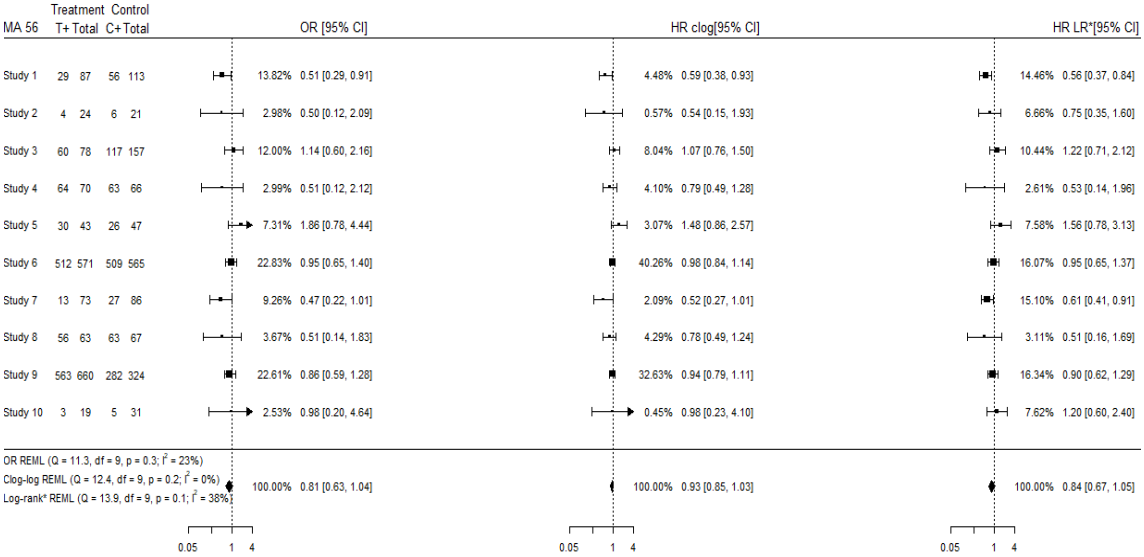
MA coloured in **blue** represent characteristics of studies outside the 95% limits of agreement based on difference of standardised estimates. MA coloured in **red** represent characteristics of studies outside the 95% limits of agreement based on difference in I^2.

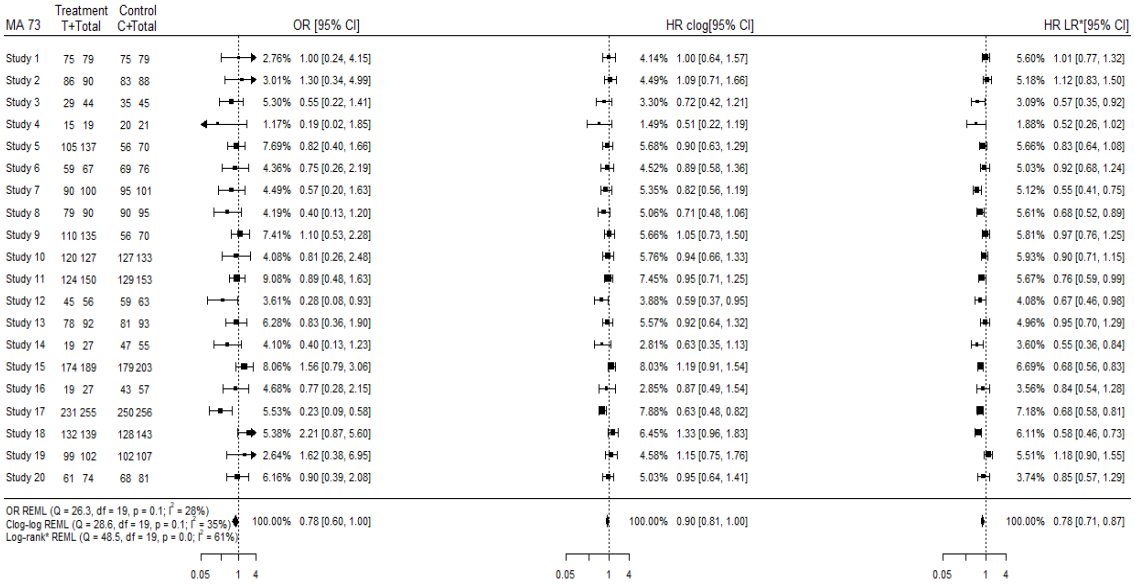
The meta-analysis forest plots below correspond to the meta-analyses presented in Table 4.4-5. The meta-analyses already presented in Chapter 4 were omitted from the figures below.

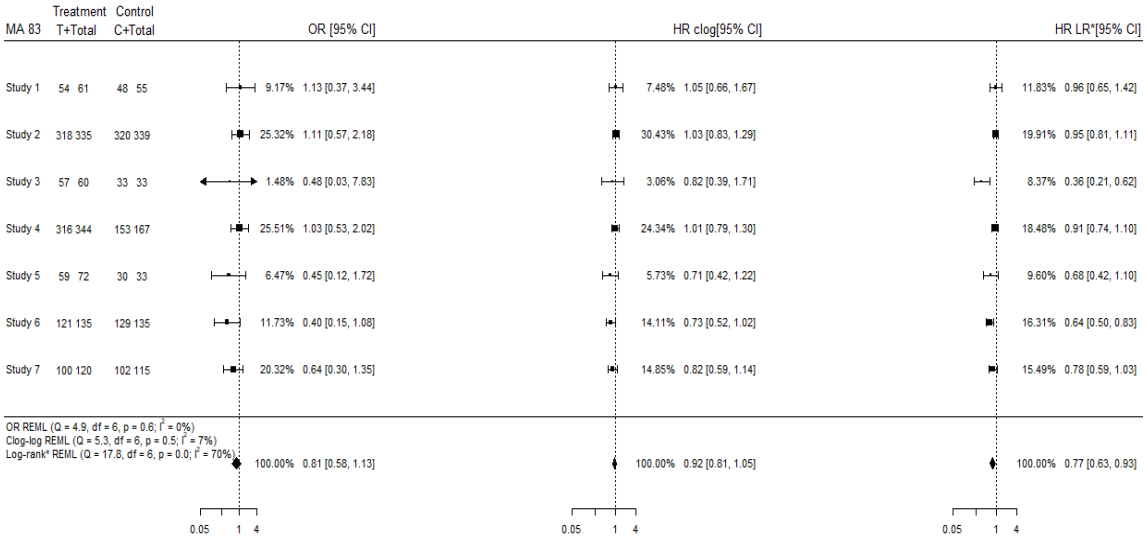


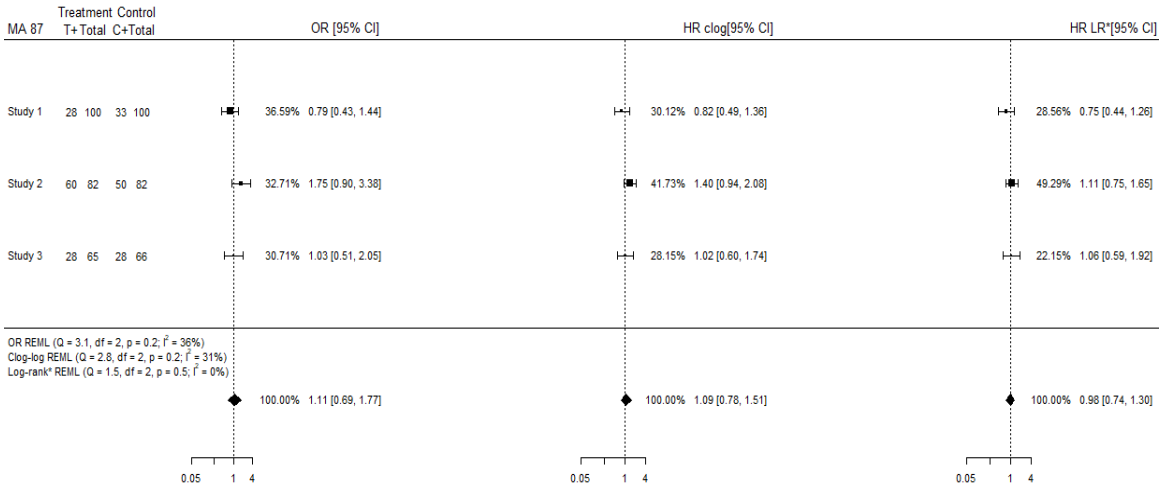


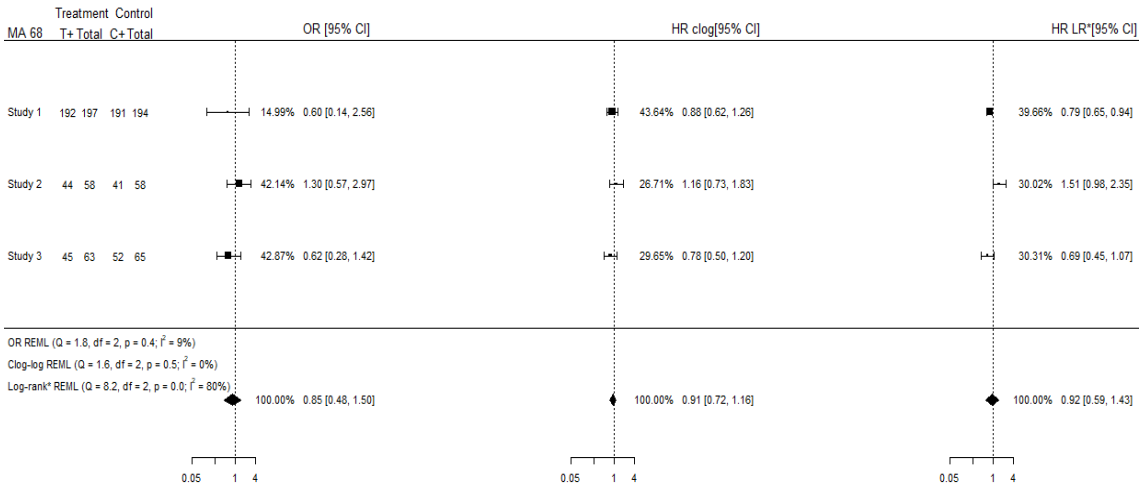



MA 56
Treatment T+ Total
Control C+ Total
OR [95% CI]
HR clog[95% CI]
HR LR*[95% CI]
Study 1 29 87 56 113 13.82% 0.51 [0.29, 0.91] 4.48% 0.59 [0.38, 0.93] 14.46% 0.56 [0.37, 0.84]
Study 2 4 24 6 21 2.98% 0.50 [0.12, 2.09] 0.57% 0.54 [0.15, 1.93] 6.66% 0.75 [0.35, 1.60]
Study 3 60 78 117 157 12.00% 1.14 [0.60, 2.16] 8.04% 1.07 [0.76, 1.50] 10.44% 1.22 [0.71, 2.12]
Study 4 64 70 63 66 2.99% 0.51 [0.12, 2.12] 4.10% 0.79 [0.49, 1.28] 2.61% 0.53 [0.14, 1.96]
Study 5 30 43 26 47 7.31% 1.86 [0.78, 4.44] 3.07% 1.48 [0.86, 2.57] 7.58% 1.56 [0.78, 3.13]
Study 6 512 571 509 565 22.83% 0.95 [0.65, 1.40] 40.26% 0.98 [0.84, 1.14] 16.07% 0.95 [0.65, 1.37]
Study 7 13 73 27 86 9.26% 0.47 [0.22, 1.01] 2.09% 0.52 [0.27, 1.01] 15.10% 0.61 [0.41, 0.91]
Study 8 56 63 63 67 3.67% 0.51 [0.14, 1.83] 4.29% 0.78 [0.49, 1.24] 3.11% 0.51 [0.16, 1.69]
Study 9 563 660 282 324 22.61% 0.86 [0.59, 1.28] 32.63% 0.94 [0.79, 1.11] 16.34% 0.90 [0.62, 1.29]
Study 10 3 19 5 31 2.53% 0.98 [0.20, 4.64] 0.45% 0.98 [0.23, 4.10] 7.62% 1.20 [0.60, 2.40]
OR REML (Q = 11.3, df = 9, p = 0.3; I² = 23%)
Clog-log REML (Q = 12.4, df = 9, p = 0.2; I² = 0%)
Log-rank* REML (Q = 13.9, df = 9, p = 0.1; I² = 38%)
100.00% 0.81 [0.63, 1.04]
100.00% 0.93 [0.85, 1.03]
100.00% 0.84 [0.67, 1.05]
0.05 1 4
0.05 1 4
0.05 1 4


MA 73
Treatment T+Total
Control C+Total
OR [95% CI]
HR clog[95% CI]
HR LR*[95% CI]
Study 1 75 79 75 79 2.76% 1.00 [0.24, 4.15] 4.14% 1.00 [0.64, 1.57] 5.60% 1.01 [0.77, 1.32]
Study 2 86 90 83 88 3.01% 1.30 [0.34, 4.99] 4.49% 1.09 [0.71, 1.66] 5.18% 1.12 [0.83, 1.50]
Study 3 29 44 35 45 5.30% 0.55 [0.22, 1.41] 3.30% 0.72 [0.42, 1.21] 3.09% 0.57 [0.35, 0.92]
Study 4 15 19 20 21 1.17% 0.19 [0.02, 1.85] 1.49% 0.51 [0.22, 1.19] 1.88% 0.52 [0.26, 1.02]
Study 5 105 137 56 70 7.69% 0.82 [0.40, 1.66] 5.68% 0.90 [0.63, 1.29] 5.66% 0.83 [0.64, 1.08]
Study 6 59 67 69 76 4.36% 0.75 [0.26, 2.19] 4.52% 0.89 [0.58, 1.36] 5.03% 0.92 [0.68, 1.24]
Study 7 90 100 95 101 4.49% 0.57 [0.20, 1.63] 5.35% 0.82 [0.56, 1.19] 5.12% 0.55 [0.41, 0.75]
Study 8 79 90 90 95 4.19% 0.40 [0.13, 1.20] 5.06% 0.71 [0.48, 1.06] 5.61% 0.68 [0.52, 0.89]
Study 9 110 135 56 70 7.41% 1.10 [0.53, 2.28] 5.66% 1.05 [0.73, 1.50] 5.81% 0.97 [0.76, 1.25]
Study 10 120 127 127 133 4.08% 0.81 [0.26, 2.48] 5.76% 0.94 [0.66, 1.33] 5.93% 0.90 [0.71, 1.15]
Study 11 124 150 129 153 9.08% 0.89 [0.48, 1.63] 7.45% 0.95 [0.71, 1.25] 5.67% 0.76 [0.59, 0.99]
Study 12 45 56 59 63 3.61% 0.28 [0.08, 0.93] 3.88% 0.59 [0.37, 0.95] 4.08% 0.67 [0.46, 0.98]
Study 13 78 92 81 93 6.28% 0.83 [0.36, 1.90] 5.57% 0.92 [0.64, 1.32] 4.96% 0.95 [0.70, 1.29]
Study 14 19 27 47 55 4.10% 0.40 [0.13, 1.23] 2.81% 0.63 [0.35, 1.13] 3.60% 0.55 [0.36, 0.84]
Study 15 174 189 179 203 8.06% 1.56 [0.79, 3.06] 8.03% 1.19 [0.91, 1.54] 6.69% 0.68 [0.56, 0.83]
Study 16 19 27 43 57 4.68% 0.77 [0.28, 2.15] 2.85% 0.87 [0.49, 1.54] 3.56% 0.84 [0.54, 1.28]
Study 17 231 255 250 256 5.53% 0.23 [0.09, 0.58] 7.88% 0.63 [0.48, 0.82] 7.18% 0.68 [0.58, 0.81]
Study 18 132 139 128 143 5.38% 2.21 [0.87, 5.60] 6.45% 1.33 [0.96, 1.83] 6.11% 0.58 [0.46, 0.73]
Study 19 99 102 102 107 2.64% 1.62 [0.38, 6.95] 4.58% 1.15 [0.75, 1.76] 5.51% 1.18 [0.90, 1.55]
Study 20 61 74 68 81 6.16% 0.90 [0.39, 2.08] 5.03% 0.95 [0.64, 1.41] 3.74% 0.85 [0.57, 1.29]
OR REML (Q = 26.3, df = 19, p = 0.1; I² = 28%)
Clog-log REML (Q = 28.6, df = 19, p = 0.1; I² = 35%)
Log-rank* REML (Q = 48.5, df = 19, p = 0.0; I² = 61%)
100.00% 0.78 [0.60, 1.00]
100.00% 0.90 [0.81, 1.00]
100.00% 0.78 [0.71, 0.87]
0.05 1 4
0.05 1 4
0.05 1 4


MA 83
Treatment T+Total
Control C+Total
OR [95% CI]
HR clog[95% CI]
HR LR*[95% CI]
Study 1 54 61 48 55 9.17% 1.13 [0.37, 3.44] 7.48% 1.05 [0.66, 1.67] 11.83% 0.96 [0.65, 1.42]
Study 2 318 335 320 339 25.32% 1.11 [0.57, 2.18] 30.43% 1.03 [0.83, 1.29] 19.91% 0.95 [0.81, 1.11]
Study 3 57 60 33 33 1.48% 0.48 [0.03, 7.83] 3.06% 0.82 [0.39, 1.71] 8.37% 0.36 [0.21, 0.62]
Study 4 316 344 153 167 25.51% 1.03 [0.53, 2.02] 24.34% 1.01 [0.79, 1.30] 18.48% 0.91 [0.74, 1.10]
Study 5 59 72 30 33 6.47% 0.45 [0.12, 1.72] 5.73% 0.71 [0.42, 1.22] 9.60% 0.68 [0.42, 1.10]
Study 6 121 135 129 135 11.73% 0.40 [0.15, 1.08] 14.11% 0.73 [0.52, 1.02] 16.31% 0.64 [0.50, 0.83]
Study 7 100 120 102 115 20.32% 0.64 [0.30, 1.35] 14.85% 0.82 [0.59, 1.14] 15.49% 0.78 [0.59, 1.03]
OR REML (Q = 4.9, df = 6, p = 0.6; I² = 0%)
Clog-log REML (Q = 5.3, df = 6, p = 0.5; I² = 7%)
Log-rank* REML (Q = 17.8, df = 6, p = 0.0; I² = 70%)
100.00% 0.81 [0.58, 1.13]
100.00% 0.92 [0.81, 1.05]
100.00% 0.77 [0.63, 0.93]
0.05 1 4
0.05 1 4
0.05 1 4


MA 87
Treatment T+Total
Control C+Total
OR [95% CI]
HR clog[95% CI]
HR LR*[95% CI]
Study 1 28 100 33 100 36.59% 0.79 [0.43, 1.44] 30.12% 0.82 [0.49, 1.36] 28.56% 0.75 [0.44, 1.26]
Study 2 60 82 50 82 32.71% 1.75 [0.90, 3.38] 41.73% 1.40 [0.94, 2.08] 49.29% 1.11 [0.75, 1.65]
Study 3 28 65 28 66 30.71% 1.03 [0.51, 2.05] 28.15% 1.02 [0.60, 1.74] 22.15% 1.06 [0.59, 1.92]
OR REML (Q = 3.1, df = 2, p = 0.2; I² = 36%)
Clog-log REML (Q = 2.8, df = 2, p = 0.2; I² = 31%)
Log-rank* REML (Q = 1.5, df = 2, p = 0.5; I² = 0%)
100.00% 1.11 [0.69, 1.77]
100.00% 1.09 [0.78, 1.51]
100.00% 0.98 [0.74, 1.30]
0.05 1 4
0.05 1 4
0.05 1 4


MA 68
Treatment T+ Total
Control C+ Total
OR [95% CI]
HR clog[95% CI]
HR LR*[95% CI]
Study 1 192 197 191 194 14.99% 0.60 [0.14, 2.56] 43.64% 0.88 [0.62, 1.26] 39.66% 0.79 [0.65, 0.94]
Study 2 44 58 41 58 42.14% 1.30 [0.57, 2.97] 26.71% 1.16 [0.73, 1.83] 30.02% 1.51 [0.98, 2.35]
Study 3 45 63 52 65 42.87% 0.62 [0.28, 1.42] 29.65% 0.78 [0.50, 1.20] 30.31% 0.69 [0.45, 1.07]
OR REML (Q = 1.8, df = 2, p = 0.4; I² = 9%)
Clog-log REML (Q = 1.6, df = 2, p = 0.5; I² = 0%)
Log-rank* REML (Q = 8.2, df = 2, p = 0.0; I² = 80%)
100.00% 0.85 [0.48, 1.50]
100.00% 0.91 [0.72, 1.16]
100.00% 0.92 [0.59, 1.43]
0.05 1 4
0.05 1 4
0.05 1 4

D –Additional material relating to the Individual Participant Data Meta-analysis analysed in Chapter 5

D.1- Kaplan-Meier Plots for time-to-event outcomes in IPD

Event Free Survival

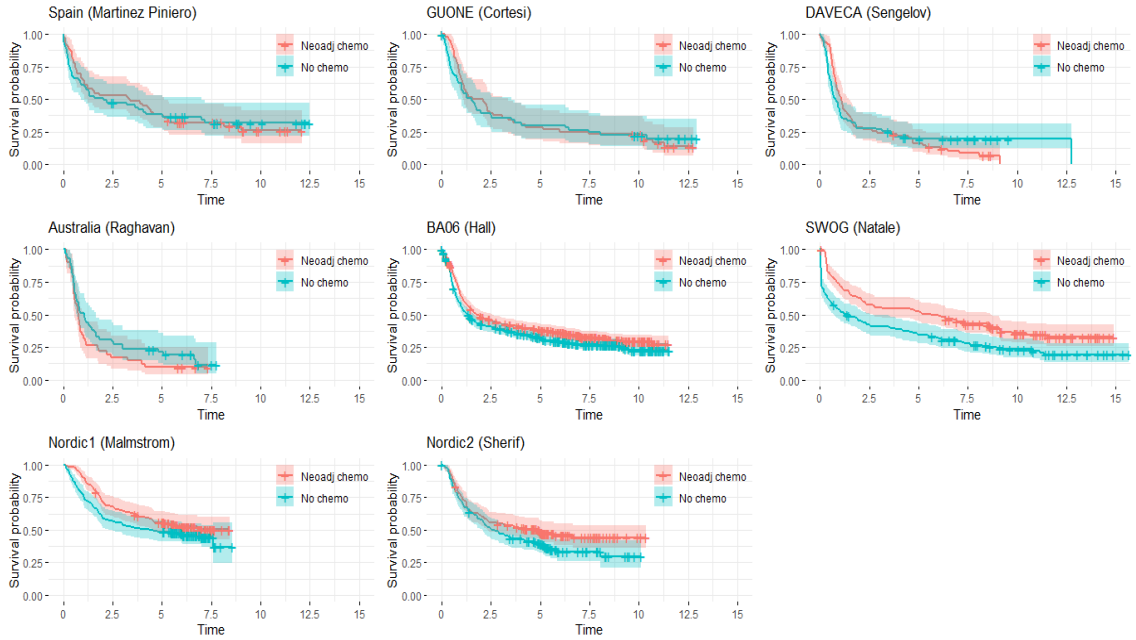


Figure 5.8: Kaplan-Meier plot for the outcome of event free survival

Local Recurrence Free Survival

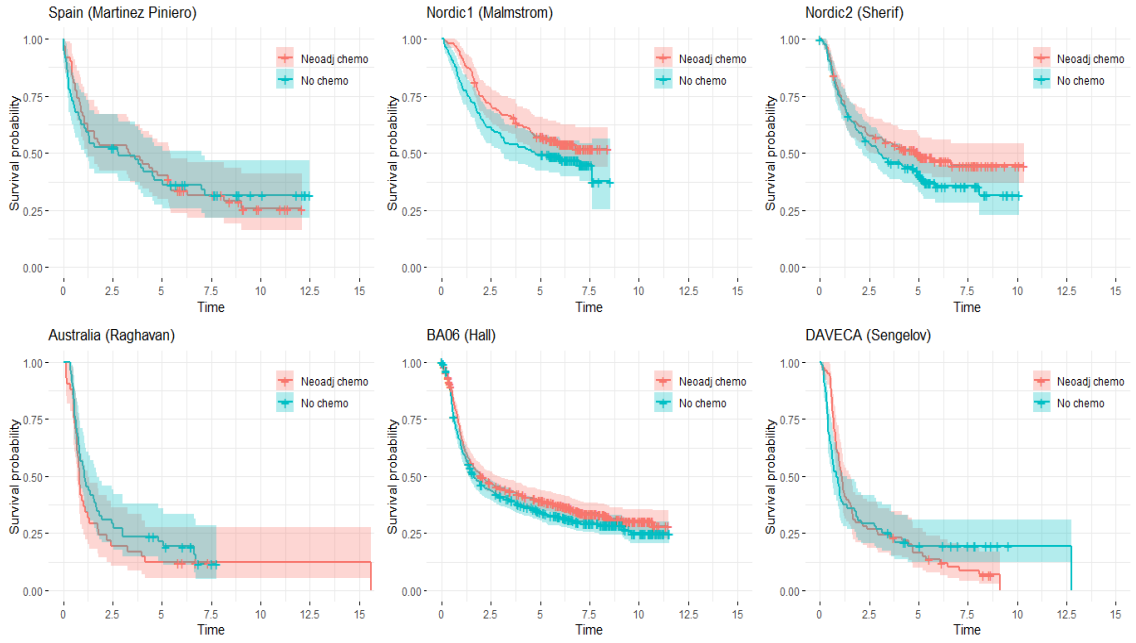


Figure 5.9: Kaplan-Meier plot for the outcome of local recurrence free survival

Metastasis Free Survival

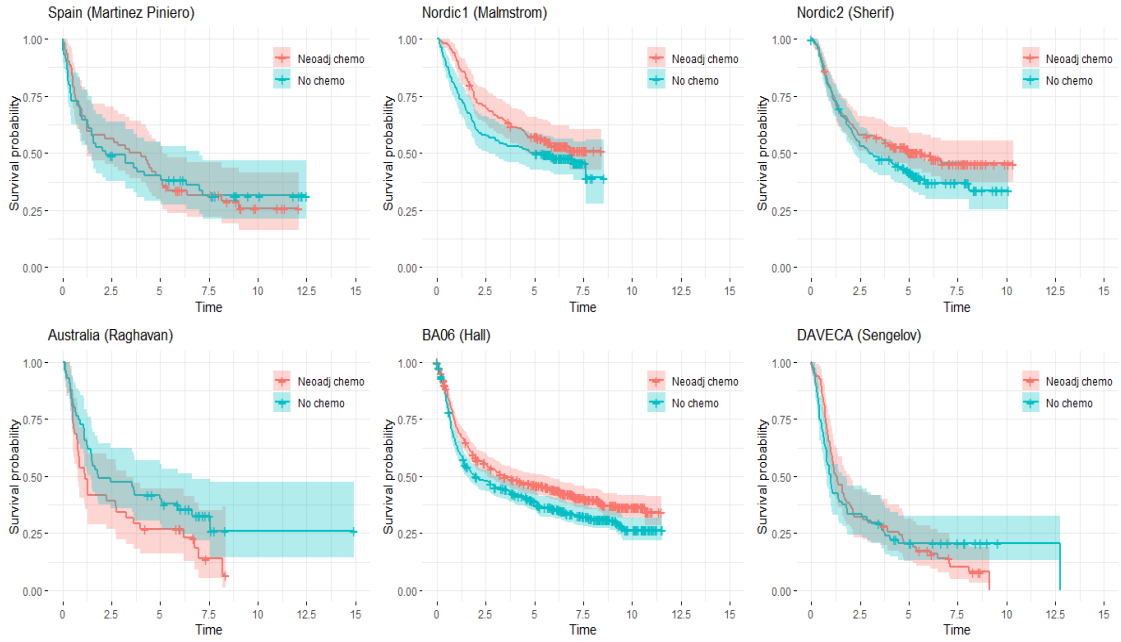


Figure 5.10: Kaplan-Meier plot for the outcome of metastasis free survival

D.2- Model Implementation

Initially, I obtained the HR and associated standard error per trial and O-E and V statistics from the log-rank test. The code below is an example using the outcome of overall survival. Similar code was used for the outcomes of event free survival, metastasis free survival and local recurrence free survival.

```
for (i in unique(IPD$TrialID)) {

cat(i,"\n")

res.O_E <- try(survdiff(Surv(Survtime_year, Surv) ~ Arm, data =
IPD[IPD$TrialID==i,]))

res.cox <- try(coxph(Surv(Survtime_year, Surv) ~ Arm, data =
IPD[IPD$TrialID==i,]))

IPDSurv[i,1]<-i

IPDSurv[i,2]<-res.O_E[["obs"]][[1]]

IPDSurv[i,3]<-res.O_E[["n"]][[1]]

IPDSurv[i,4]<-res.O_E[["n"]][[1]]-res.O_E[["obs"]][[1]]
```

```
IPDSurv[i,5]<-res.O_E[["obs"]][[2]]

IPDSurv[i,6]<-res.O_E[["n"]][[2]]

IPDSurv[i,7]<-res.O_E[["n"]][[2]]-res.O_E[["obs"]][[2]]

IPDSurv[i,8]<-res.O_E[["obs"]][[1]]-res.O_E[["exp"]][[1]]  #O-E from treatment only#

IPDSurv[i,9]<-res.O_E[["var"]][[1,1]]

IPDSurv[i,10]<-coef(summary(res.cox))[,1]

IPDSurv[i,11]<-coef(summary(res.cox))[,3]}

IPDSurv<-na.omit(IPDSurv)

IPDSurv$TrialID<-ordered(IPDSurv$TrialID, levels = c(1,2,3,4,7,8,9,10,11), c("Spain","Australia","Nordic 1","UK","GUONE","BA06","Nordic 2","DAVECA","SWOG"))

IPDSurv$Outcome<-1
```

Useful functions facilitating model implementation at later stages

```
cloglogfun<-function(dat)

{  dat$proptreat<-dat$treat_n/(dat$treat_n+dat$nontreat_n)

  dat$propctrl<-dat$ctrl_n/(dat$ctrl_n+dat$nonctrl_n)

  dat$logHRcloglog<-(log(-log(1-dat$proptreat)))-log(-log(1-dat$propctrl))

  dat$derivTreat<-1/((log(1-dat$proptreat))*(dat$proptreat-1))

  dat$derivCtrl<-1/((log(1-dat$propctrl))*(dat$propctrl-1))

  dat$varTreat<-(dat$derivTreat^2)*((dat$proptreat*(1-dat$proptreat))/dat$treat_total)

  dat$varCtrl<-(dat$derivCtrl^2)*((dat$propctrl*(1-dat$propctrl))/dat$ctrl_total)

  dat$varHRcloglog<-dat$varTreat+dat$varCtrl

  dat$logOR<-log((dat$treat_n*dat$nonctrl_n)/(dat$ctrl_n*dat$nontreat_n))
```

```
  dat$varOR<-
(1/dat$treat_n)+(1/dat$nontreat_n)+(1/dat$ctrl_n)+(1/dat$nonctrl_n)

  dat$O_ElogHR<-dat$O_E/dat$var_O_E

  dat$O_EvarHR<-1/dat$var_O_E

  dat$within_prec<-1/dat$varHRcloglog

  dat$within_prec2<-(1/dat$varHRcloglog)^2

  dat_table <- data.frame(dat)

  dat_table}

IPDSurv<-cloglogfun(IPDSurv)


#Create data in long format in order to create dataset forms necessary for one-
stage models

golong <- function(dat)

{n <- c(dat$treat_n+dat$nontreat_n, dat$ctrl_n+dat$nonctrl_n)

  event <- c(dat$treat_n, dat$ctrl_n)

  study <- c(1:nrow(dat), 1:nrow(dat))

  obs <- 1:length(n)

  treat <- c(rep(1, length(n)/2), rep(0, length(n)/2))

  control <- 1-treat

  treat12 <- treat - 0.5

  outcome.num<-c(dat$Outcome)

dat_long <- data.frame(n, event, study, obs, treat, control, treat12,
outcome.num)

dat_long}

datlong.IPDEvent <- golong(IPDEvent)

datlong.IPDLRFS <- golong(IPDLRFS)
```

```
datlong.IPDMFS <- golong(IPDMFS)

datlong.IPDSurv <- golong(IPDSurv)
```

#Function extracting information from one-stage cox models.

```
se <- function(object)

sqrt(diag(vcov(object)))

confint.coxme <- function(object, level = .95, digits = 9) {

z <- qnorm(1 - (1 - level)/2)

b <- coef(object)

s <- se(object)

ci.lb <- b - z * s

ci.ub <- b + z * s

out <- data.frame(b, ci.lb, ci.ub, s, exp(b), exp(ci.lb), exp(ci.ub))

out <- round(out, digits = digits)

colnames(out) <- c("coef", "ci.lb(coef)", "ci.ub(coef)", "se(coef)", "exp(coef)",
"ci.lb(exp(coef))", "ci.ub(exp(coef))")

out$`Wald p` <- round(pnorm(b/s, lower.tail = F) * 2, digits + 1)

out$CI <- paste(out$`ci.lb(exp(coef))`, " to ", out$`ci.ub(exp(coef))`, sep = "",
collapse = NULL) out}
```

Model Implementation

```
try.fit1<- try(rma.uni(ai = treat_n, bi = nontreat_n, ci = ctrl_n, di = nonctrl_n, data
= IPDSurv, measure = "OR",method="REML",slab=paste(TrialID),
control=list(maxiter=500, verbose=TRUE, stepadj=0.5), verbose=TRUE))

try.fit2<- try(rma.uni(yi = logHRcloglog, vi = varHRcloglog, data = IPDSurv,
slab=paste(TrialID), method="REML",control=list(maxiter=10e9,
verbose=TRUE, stepadj=0.2), verbose=TRUE))
```

```
try.fit3<- try(rma.uni(yi = O_ElogHR, vi = O_EvarHR, data = IPDSurv,
slab=paste(TrialID), method="REML",control=list(maxiter=10e9,
verbose=TRUE, stepadj=0.2), verbose=TRUE))

try.fit4<- try(rma.uni(yi = -logHR, sei = logHR_se, data = IPDSurv,
slab=paste(TrialID), method="REML",control=list(maxiter=10e9,
verbose=TRUE, stepadj=0.2), verbose=TRUE))

try.fit5<- try(rma.glmm(ai = treat_n, bi = nontreat_n, ci = ctrl_n, di = nonctrl_n,
data = IPDSurv, measure = "OR",model="UM.FS", drop00=F,nAGQ=7))

try.fit6<-try(glmer(cbind(event,n-event) ~ factor(treat) + factor(study) + (treat12-
1|study), data=datlong.IPDSurv, family=binomial(link="cloglog"),nAGQ=7))

try.fit7<-try(coxme(Surv(Survtime_year, Surv) ~ Arm + (1+Arm |TrialID), data =
IPD))
```

E – Additional material relating to the Simulation Study presented in Chapter 6

E.1-Simulation Code

```
#Reproducibility: Set the seed at the beginning of the DGM script
set.seed(2109990)
# This function simulates the data
simdata <- function(j, dgm, n=5, prob = 0.5,mean.trialsize=1000, mean.trialsize.sd=100, mean.hr=0, hr.tau=0,
                    mean.fu=5, mean.fu.sd=1, Size=1,lambdae = 0.1,lambdac=0.05, gamma = 2) {
  trialsize=list()
  trialsize[[1]] <- data.frame(rep(floor(rnorm(n=n, mean=mean.trialsize, sd=mean.trialsize.sd)),1))
  colnames(trialsize[[1]])<-"TrialSize"
  for (i in 1:length(trialsize[[1]][,])) {
    trialsize[[1]]$Beta[i]<-rnorm(n=1, mean=mean.hr, sd=sqrt(hr.tau)) #0.0002 when tau changes
    trialsize[[1]]$ExpBeta[i]<-exp(trialsize[[1]]$Beta[i])
    trialsize[[1]]$followUp[i]<-round(rnorm(n=1, mean=mean.fu, sd=mean.fu.sd), 1) }

  # Generate a data set with ID and a binary variable treatment group indicator:
  df=list()
  for (i in seq_along(trialsize[[1]]$TrialSize)) {
```

```
  cat(i,"\n")
  df[[i]] <- data.frame( id = 1:trialsize[[1]]$TrialSize[i], trt = rbinom(n = trialsize[[1]]$TrialSize[i], size = Size, prob = prob) ) }


# Simulate survival time & censoring time with maximum follow-up time of x years
s=list()
#c=list() #censoring
for (i in seq_along(df)) {
  cat(i,"\n")
  s[[i]] <- simsurv(dist = "weibull", lambdas = lambdae, gammas = gamma, betas = c(trt = trialsize[[1]]$Beta[i]), x = df[[i]], maxt = trialsize[[1]]$followUp[i])
  #c[[i]] <- simsurv(dist = "weibull", lambdas = lambdac, gammas = gamma, betas = c(trt = trialsize[[1]]$Beta[i]), x = df[[i]], maxt = trialsize[[1]]$followUp[i])    #Run this if you want to add censoring
  #output: id=identifier, eventtime=simulated event(censoring time), status=event indicator,1=failure,0=censored}
# without censoring in your data
for (i in seq_along(df)){
  for (l in 1:length(df[[i]]$id)){
    s[[i]]$eventtime[l]<-s[[i]]$eventtime[l]
    #s[[i]]$randcenstime[l]<-c[[i]]$eventtime[l]
    s[[i]]$fixedcenstime[l]<-max(s[[i]]$eventtime)
```

```
    s[[i]]$Observed_time[l] <- s[[i]]$eventtime[l]
    s[[i]]$case[l] <- ifelse(s[[i]]$Observed_time[l] == s[[i]]$eventtime[l], 1, 2)
    s[[i]]$status_new[l] <- ifelse(s[[i]]$case[l] == 1, s[[i]]$status[l], 0) }}

#include censoring in your data
for (i in seq_along(df)){
  for (l in 1:length(df[[i]]$id)){
    s[[i]]$eventtime[l]<-s[[i]]$eventtime[l]
    s[[i]]$randcenstime[l]<-c[[i]]$eventtime[l]
    s[[i]]$fixedcenstime[l]<-trialsize[[1]]$followUp[i]
    s[[i]]$Observed_time[l] <- pmin(s[[i]]$eventtime[l], s[[i]]$randcenstime[l], s[[i]]$fixedcenstime[l])
    s[[i]]$case[l] <- ifelse(s[[i]]$Observed_time[l] == s[[i]]$eventtime[l], 1, ifelse(s[[i]]$Observed_time[l] == s[[i]]$randcenstime[l], 2, 3))
    s[[i]]$status_new[l] <- ifelse(s[[i]]$case[l] == 1, s[[i]]$status[l], 0) }}

simuldata<-list()
for (i in seq_along(df)) {
  simuldata[[i]] <- merge(df[[i]], s[[i]])
  simuldata[[i]]$TrialID<-i
```

```
    simuldata[[i]]$SimID<-j
    simuldata[[i]]$dgm<-dgm}

  # Merge all data in a data matrix
  simuldata<-data.frame(do.call(rbind, simuldata))

  # Merge covariates data and survival times:
  #We save the current seed as an attribute of each data set
  attr(simuldata, "seed") <- .Random.seed
  return(simuldata)}

# This function fits the models using the simulated data
simfit <- function(k, dgm,simuldata) {
  # For each Trial apply log-rank and Cox proportional hazards model
  IPDSim=data.frame(matrix(NA, max(simuldata$TrialID), 11))
  colnames(IPDSim)<-c("TrialID","treat_n", "treat_total","nontreat_n",
                      "ctrl_n", "ctrl_total","nonctrl_n",
                      "O_E", "var_O_E", "logHR", "logHR_se")
```

```
 for (i in unique(simuldata$TrialID)) {
   cat(i,"\n")
   res.O_E <- tryCatch(survdiff(Surv(Observed_time, status_new) ~ trt,
                                data = simuldata[simuldata$TrialID==i,]), error = function(e) NULL)
   res.cox <- tryCatch(coxph(Surv(Observed_time, status_new) ~ trt,
                             data = simuldata[simuldata$TrialID==i,]), error = function(e) NULL)
   IPDSim[i,1]<-i
   IPDSim[i,2]<-res.O_E[["obs"]][[2]]
   IPDSim[i,3]<-res.O_E[["n"]][[2]]
   IPDSim[i,4]<-res.O_E[["n"]][[2]]-res.O_E[["obs"]][[2]]
   IPDSim[i,5]<-res.O_E[["obs"]][[1]]
   IPDSim[i,6]<-res.O_E[["n"]][[1]]
   IPDSim[i,7]<-res.O_E[["n"]][[1]]-res.O_E[["obs"]][[1]]
   IPDSim[i,8]<-res.O_E[["obs"]][[2]]-res.O_E[["exp"]][[2]]  #O-E from treatment only#
   IPDSim[i,9]<-res.O_E[["var"]][[2,2]]
   IPDSim[i,10]<-coef(summary(res.cox))[,1]
   IPDSim[i,11]<-coef(summary(res.cox))[,3]
}
```

```
# #Identify rare events and apply continuity correction
IPDSim$rare<-ifelse(IPDSim$treat_n<1, 1,
                    ifelse(IPDSim$ctrl_n<1, 1,
                           ifelse(IPDSim$nontreat_n<1, 1,
                                  ifelse(IPDSim$nonctrl_n<1, 1,0))))
table(IPDSim$rare)

 #CC adding the reciprocal of the opposite treatment arm size to those with rare events
IPDSim$result<-ifelse(IPDSim$rare==1,IPDSim$treat_total/IPDSim$ctrl_total,0)
IPDSim$armtreat<-1/IPDSim$treat_total
IPDSim$armctrl<-1/IPDSim$ctrl_total
IPDSim$TCC<-ifelse(IPDSim$result!=0, 1/(IPDSim$result+1),0)
IPDSim$CCC<-ifelse(IPDSim$result!=0, IPDSim$result/(IPDSim$result+1),0)
IPDSim$treat_n<-ifelse(IPDSim$result!=0, IPDSim$TCC+IPDSim$treat_n, IPDSim$treat_n)
IPDSim$ctrl_n<-ifelse(IPDSim$result!=0, IPDSim$CCC+IPDSim$ctrl_n, IPDSim$ctrl_n)
IPDSim$nontreat_n<-ifelse(IPDSim$result!=0, IPDSim$TCC+IPDSim$nontreat_n, IPDSim$nontreat_n)
IPDSim$nonctrl_n<-ifelse(IPDSim$result!=0,IPDSim$CCC+IPDSim$nonctrl_n, IPDSim$nonctrl_n)
```

```
# Calculation facilitating implementation of a cloglog model
  cloglogfun<-function(dat)
  {dat$proptreat<-dat$treat_n/(dat$treat_n+dat$nontreat_n)
    dat$propctrl<-dat$ctrl_n/(dat$ctrl_n+dat$nonctrl_n)
    dat$logHRcloglog<-(log(-log(1-dat$proptreat)))-log(-log(1-dat$propctrl))
    dat$derivTreat<-1/((log(1-dat$proptreat))*(dat$proptreat-1))
    dat$derivCtrl<-1/((log(1-dat$propctrl))*(dat$propctrl-1))
    dat$varTreat<-(dat$derivTreat^2)*((dat$proptreat*(1-dat$proptreat))/dat$treat_total)
    dat$varCtrl<-(dat$derivCtrl^2)*((dat$propctrl*(1-dat$propctrl))/dat$ctrl_total)
    dat$varHRcloglog<-dat$varTreat+dat$varCtrl
    dat$logOR<-log((dat$treat_n*dat$nonctrl_n)/(dat$ctrl_n*dat$nontreat_n))
    dat$varOR<-(1/dat$treat_n)+(1/dat$nontreat_n)+(1/dat$ctrl_n)+(1/dat$nonctrl_n)
    dat$O_ElogHR<-dat$O_E/dat$var_O_E
    dat$O_EvarHR<-1/dat$var_O_E
    dat$within_prec<-1/dat$varHRcloglog
    dat$within_prec2<-(1/dat$varHRcloglog)^2
    dat_table <- data.frame(dat)
    dat_table}
```

```
IPDSim<-cloglogfun(IPDSim)
# Run two-stage MA models and compare.
IPDsimRes=data.frame(matrix(NA, 4, 6))
colnames(IPDsimRes)<-c("logestimates","SE", "LowerCI", "UpperCI","Tau", "Isq")
try.fit1<- tryCatch(rma.uni(ai = treat_n, bi = nontreat_n, ci = ctrl_n, di = nonctrl_n,
                    data = IPDSim, measure = "OR",method="REML", slab=paste(TrialID),
                    control=list(maxiter=500, verbose=TRUE, stepadj=0.5), verbose=TRUE), error = function(e) NULL)
try.fit2<- tryCatch(rma.uni(yi = logHRcloglog, vi = varHRcloglog, data = IPDSim, slab=paste(TrialID),
                    method="REML",control=list(maxiter=10e9, verbose=TRUE, stepadj=0.2), verbose=TRUE), error = function(e) NULL)
try.fit3<- tryCatch(rma.uni(yi = O_ElogHR, vi = O_EvarHR, data = IPDSim, slab=paste(TrialID),
                    method="REML",control=list(maxiter=10e9, verbose=TRUE, stepadj=0.2), verbose=TRUE), error = function(e) NULL)
try.fit4<- tryCatch(rma.uni(yi = logHR, sei = logHR_se, data = IPDSim, slab=paste(TrialID),
                    method="REML",control=list(maxiter=10e9, verbose=TRUE, stepadj=0.2), verbose=TRUE), error = function(e) NULL)
try.fit<-list(try.fit1,try.fit2,try.fit3,try.fit4)
#Extract results from models
for (i in 1:4) {
  cat(i,"\n")
  if(!is.null(try.fit[i])) {
```

```
    IPDsimRes[i,1]<-as.numeric(try.fit[[i]][["b"]])
    IPDsimRes[i,2]<-as.numeric(try.fit[[i]][["se"]])
    IPDsimRes[i,3]<-as.numeric(try.fit[[i]][["ci.lb"]])
    IPDsimRes[i,4]<-as.numeric(try.fit[[i]][["ci.ub"]])
    IPDsimRes[i,5]<-as.numeric(try.fit[[i]][["tau2"]])
    IPDsimRes[i,6]<-as.numeric(try.fit[[i]][["I2"]])
  } else {
    IPDsimRes[i,1]<-NA
    IPDsimRes[i,2]<-NA
    IPDsimRes[i,3]<-NA
    IPDsimRes[i,4]<-NA
    IPDsimRes[i,5]<-NA
    IPDsimRes[i,6]<-NA
  }
}
remove(try.fit1, try.fit2, try.fit3, try.fit4)
#Indicate the model applied
modelfun<-function(dat)
```

```
  { dat$Model<-c("OR REML (2-stage)","HR cloglog (2-stage)","HR O-E & V (2-stage)",
                 "HR CoxPH (2-stage)")
    dat_table <- data.frame(dat)
    dat_table }
  IPDsimRes<-modelfun(IPDsimRes)
  out <- data.frame( k = k, dgm = dgm,
                     IPDsimRes = IPDsimRes)
  return(out) }
#Run Simulation 1000 times
B <- 1000
dgm<-1:1
set.seed(2109990)
datares<-foreach (k = 1:B, .combine=rbind, .packages= "foreach") %do%
{
simdata(j=k, dgm = 1)}
results <- foreach (k = 1:B, .combine=rbind, .packages= "foreach") %do% {
  simfit(k = k, dgm=1, simuldata=datares[datares$SimID==k,])
}
```

E.2- Information obtained from the literature facilitating the decision of simulation scenarios and exact tables containing the results from the simulation scenarios

Source	No. of studies per MA	Study Sample Size	Follow-up	Censoring Rate	τ^2	I^2	Log HR	Survival times/other
MRC CTU 18 IPD MA*	5-19	491-8447	<1 month-(approx. 10 y)	-	-	40%-70%	(-0.43,-0.19)	-
Bowden et al.[159]	4,8,10, 16,25	50, 100, 150	-	10%	0.1	42%-69%	(0, 0.4, 0.8)	Exponential
Simmonds et al.[21]	-	100-1000	5 or 10 years	0%-20%	-	-	(0-1)	Weibull
Hirooka et al.[109]	-	100, 300 per group	5 years	0%, 30%	-	-	1, 0.9, 0.8, 0.7, 0.6	Exponential/ Surv rate: 20%, 50%, 80%
Katsahian et al.[108]	3,5,10, 20,30	240, 600, 2400	-	-	0, 0.15, 0.6	-	0, -0.223	30 studies only for 2400 participants
Tudur-Smith et al.[107]	5	100 per group	-	-	0, 0.01, 0.03, 0.07, 0.1	14%, 25%, 43%, 62%, 70%	0, 0.1, 0.5, 0.9	Exponential

IPD MA (Chapter 5)	9	96-976	1-5 years	15%-51%	0-0.017	0%-37%	-0.147, -0.052	Overall survival outcome
"OEV" data (Chapter 3)	3, 6, 10, 14, 32	72, 116, 160, 279, 739, 985	-	-	0, 0, 0.01, 0.04, 0.13	0%, 0%, 21%, 46%, 75%	0.68, 0.82, 0.93, 1.03, 1.35	-

*Based on Bowden et al.[159] paper & research papers from the literature review chapter.

Table 6.3: Information obtained from the literature facilitating the decision of parameters for the simulation scenarios.

Run	Participants per trial (Mean, SD)	Log HR (M)	τ^2	Follow-up (M, sd)	$\gamma_e = \gamma_c$	λ_e	λ_c	p(E>FU)	p(C>FU)	P(min(E, C) > FU)	P(min(E, C) < FU)	P(E <C < FU)	P(C < E < FU)
5 trial per MA													
Scenario 0*	(1000, 100)	0	0	(5, 1)	2	0.1	0	0.08	1.00	0.08	0.92	0.92	0.00
Base Case**	(400, 40)	-0.3	0.05	(3, 0.3)	2	0.1	0.05	0.41	0.64	0.26	0.74	0.49	0.25
Short F-up	(400, 40)	-0.3	0.05	(1, 0.2)	2	0.1	0	0.90	1.00	0.90	0.10	0.10	0.00
Long F-up	(400, 40)	-0.3	0.05	(5, 1)	2	0.1	0.07	0.08	0.17	0.01	0.99	0.58	0.41
Large heterogeneity	(400, 40)	-0.3	0.1	(3, 0.3)	2	0.1	0.05	0.41	0.64	0.26	0.74	0.49	0.25
Small heterogeneity	(400, 40)	-0.3	0.001	(3, 0.3)	2	0.1	0.05	0.41	0.64	0.26	0.74	0.49	0.25

Large effect size	(400, 40)	-0.8	0.05	(3, 0.3)	2	0.1	0.05	0.41	0.64	0.26	0.74	0.49	0.25
No effect size	(400, 40)	0	0.05	(3, 0.3)	2	0.1	0.05	0.41	0.64	0.26	0.74	0.49	0.25
Small sample size	(100, 15)	-0.3	0.05	(3, 0.3)	2	0.1	0.05	0.41	0.64	0.26	0.74	0.49	0.25
Large sample size	(1000, 100)	-0.3	0.05	(3, 0.3)	2	0.1	0.05	0.41	0.64	0.26	0.74	0.49	0.25
Small P(Event)	(400, 40)	-0.3	0.05	(3, 0.3)	2	0.05	0.04	0.64	0.70	0.44	0.56	0.31	0.25
Large % R_cens+Small P(Event)	(400, 40)	-0.3	0.05	(5, 1)	2	0.05	0.04	0.29	0.37	0.11	0.89	0.50	0.40
80% Power	(400, 40)	-0.3	0.027	(3, 0.3)	2	0.1	0.05	0.41	0.64	0.26	0.74	0.49	0.25
Long Follow-up+0% R_cens	(400, 40)	-0.3	0.05	(5, 1)	2	0.05	0	0.29	1.00	0.29	0.71	0.71	0.00
20 trial per MA													
Scenario 0*	(1000, 100)	0	0	(5, 1)	2	0.1	0	0.08	1.00	0.08	0.92	0.92	0.00
Base Case**	(400, 40)	-0.3	0.05	(3, 0.3)	2	0.1	0.05	0.41	0.64	0.26	0.74	0.49	0.25
Short F-up	(400, 40)	-0.3	0.05	(1, 0.2)	2	0.1	0	0.90	1.00	0.90	0.10	0.10	0.00
Long F-up	(400, 40)	-0.3	0.05	(5, 1)	2	0.1	0.07	0.08	0.17	0.01	0.99	0.58	0.41
Large heterogeneity	(400, 40)	-0.3	0.1	(3, 0.3)	2	0.1	0.05	0.41	0.64	0.26	0.74	0.49	0.25

Small heterogeneity	(400, 40)	-0.3	0.001	(3, 0.3)	2	0.1	0.05	0.41	0.64	0.26	0.74	0.49	0.25
Large effect size	(400, 40)	-0.8	0.05	(3, 0.3)	2	0.1	0.05	0.41	0.64	0.26	0.74	0.49	0.25
No effect size	(400, 40)	0	0.05	(3, 0.3)	2	0.1	0.05	0.41	0.64	0.26	0.74	0.49	0.25
Small sample size	(100, 15)	-0.3	0.05	(3, 0.3)	2	0.1	0.05	0.41	0.64	0.26	0.74	0.49	0.25
Large sample size	(1000, 100)	-0.3	0.05	(3, 0.3)	2	0.1	0.05	0.41	0.64	0.26	0.74	0.49	0.25
Small P(Event)	(400, 40)	-0.3	0.05	(3, 0.3)	2	0.05	0.04	0.64	0.70	0.44	0.56	0.31	0.25
Large % R_cens+Small P(Event)	(400, 40)	-0.3	0.05	(5, 1)	2	0.05	0.04	0.29	0.37	0.11	0.89	0.50	0.40
80% Power	(400, 40)	-0.3	0.2	(3, 0.3)	2	0.1	0.05	0.41	0.64	0.26	0.74	0.49	0.25
Long Follow-up+0% R_cens	(400, 40)	-0.3	0.05	(5, 1)	2	0.05	0	0.29	1.00	0.29	0.71	0.71	0.00

Scenario 0*: Large sample size, No effect size, No heterogeneity, Long F-Up; Base Case**: Medium sample size, medium effect size, medium heterogeneity, medium follow up; Other scenarios change from base case; R_cens=Random censoring; P(E)=P(Event)

Table 6.4: Presentation of simulation scenarios and event, censoring probabilities for the different simulation scenarios applied.

Simulation Scenarios					Methods			
Run	No. of participants per trial (Mean, SD)	Mean Log HR	τ	Follow-up time (Mean, SD)	Two-stage Cox PH Model (HR)	Two-stage ("O-E" & V) (HR)	Two-stage clog-log (HR)	Two-stage Logit link (OR)
5 trials per MA								
Scenario 0*	(1000, 100)	0	0	(5, 1)	-0.0006	-0.0006	-0.0003	-0.0005
Base Case**	(400, 40)	-0.3	0.05	(3, 0.3)	0.0072	0.005	0.0822	0.01
Short F-up	(400, 40)	-0.3	0.05	(1, 0.2)	0.0096	0.0081	0.0092	-0.0062
Long F-up	(400, 40)	-0.3	0.05	(5, 1)	0.0043	-0.0026	0.2321	0.2009
Large heterogeneity	(400, 40)	-0.3	0.1	(3, 0.3)	0.0085	0.0062	0.0764	0.0058
Small heterogeneity	(400, 40)	-0.3	0.001	(3, 0.3)	0.0025	0.0009	0.0855	0.0124
Large effect size	(400, 40)	-0.8	0.05	(3, 0.3)	0.0082	0.0012	0.1739	0.0017
No effect size	(400, 40)	0	0.05	(3, 0.3)	0.006	0.0063	-0.006	-0.007
Small sample size	(100, 15)	-0.3	0.05	(3, 0.3)	0.0057	0.0013	0.0807	0.0048
Large sample size	(1000, 100)	-0.3	0.05	(3, 0.3)	0.0048	0.0031	0.0795	0.0088
Small P(Event)	(400, 40)	-0.3	0.05	(3, 0.3)	0.0094	0.0092	0.0599	0.0181
Large % R_cens+Small P(Event)	(400, 40)	-0.3	0.05	(5, 1)	0.005	0.0004	0.1566	0.1077

Long Follow-up+0% R_cens	(400, 40)	-0.3	0.05	(5, 1)	0.0048	0.0028	0.0027	-0.187
80% Power	(400, 40)	-0.3	0.027	(3, 0.3)	0.0061	0.0041	0.0844	0.0117
20 trials per MA								
Scenario 0*	(1000, 100)	0	0	(5, 1)	-0.0006	-0.0006	-0.0006	-0.0019
Base Case**	(400, 40)	-0.3	0.05	(3, 0.3)	0.006	0.0036	0.0818	0.0082
Short F-up	(400, 40)	-0.3	0.05	(1, 0.2)	0.0242	0.0176	0.0235	0.0079
Long F-up	(400, 40)	-0.3	0.05	(5, 1)	0.0027	-0.004	0.2327	0.2002
Large heterogeneity	(400, 40)	-0.3	0.1	(3, 0.3)	0.0075	0.0049	0.0763	0.0044
Small heterogeneity	(400, 40)	-0.3	0.001	(3, 0.3)	0.0006	-0.0012	0.0838	0.0095
Large effect size	(400, 40)	-0.8	0.05	(3, 0.3)	0.0093	0.0013	0.1758	0.0025
No effect size	(400, 40)	0	0.05	(3, 0.3)	0.0048	0.0049	-0.0057	-0.0078
Small sample size	(100, 15)	-0.3	0.05	(3, 0.3)	0.0123	0.0067	0.0845	0.0065
Large sample size	(1000, 100)	-0.3	0.05	(3, 0.3)	0.0037	0.0018	0.0793	0.0079
Small P(Event)	(400, 40)	-0.3	0.05	(3, 0.3)	0.0117	0.0107	0.0634	0.0209
Large % R_cens+Small P(Event)	(400, 40)	-0.3	0.05	(5, 1)	0.0042	-0.0005	0.1586	0.1086
Long Follow-up+0% R_cens	(400, 40)	-0.3	0.05	(5, 1)	0.0042	0.002	0.0019	-0.1901
80% Power	(400, 40)	-0.3	0.2	(3, 0.3)	0.0092	0.0075	0.064	-0.0048

Scenario 0*: Large sample size, No effect size, No heterogeneity, Long F-Up; Base Case**: Medium sample size, medium effect size, medium heterogeneity, medium follow up; Other scenarios change from base case; R_cens=Random censoring; P(E)=P(Event)

Table 6.5: Bias observed per simulation scenario across different meta-analysis models.

Simulation Scenarios					Methods			
Run	No. of participants per trial (Mean, SD)	Mean Log HR	τ	Follow-up time (Mean, SD)	Two-stage Cox PH Model (HR)	Two-stage ("O-E" & V) (HR)	Two-stage clog-log (HR)	Two-stage Logit link (OR)
5 trials per MA								
Scenario 0*	(1000, 100)	0	0	(5, 1)	10%	10%	9%	10%
Base Case**	(400, 40)	-0.3	0.05	(3, 0.3)	-4%	-4%	-2%	-2%
Short F-up	(400, 40)	-0.3	0.05	(1, 0.2)	6%	4%	6%	6%
Long F-up	(400, 40)	-0.3	0.05	(5, 1)	-3%	-3%	9%	9%
Large heterogeneity	(400, 40)	-0.3	0.1	(3, 0.3)	-3%	-3%	-3%	-2%
Small heterogeneity	(400, 40)	-0.3	0.001	(3, 0.3)	7%	6%	9%	10%
Large effect size	(400, 40)	-0.8	0.05	(3, 0.3)	-3%	-3%	-1%	-1%
No effect size	(400, 40)	0	0.05	(3, 0.3)	-3%	-3%	0%	0%

Small sample size	(100, 15)	-0.3	0.05	(3, 0.3)	0%	-1%	4%	5%
Large sample size	(1000, 100)	-0.3	0.05	(3, 0.3)	0%	0%	1%	1%
Small P(Event)	(400, 40)	-0.3	0.05	(3, 0.3)	-1%	-1%	0%	0%
Large % R_cens+Small P(Event)	(400, 40)	-0.3	0.05	(5, 1)	-2%	-2%	4%	4%
Long Follow-up+0% R_cens	(400, 40)	-0.3	0.05	(5, 1)	-1%	-1%	-1%	1%
80% Power	(400, 40)	-0.3	0.027	(3, 0.3)	-3%	-3%	0%	0%
20 trials per MA								
Scenario 0*	(1000, 100)	0	0	(5, 1)	8%	8%	7%	7%
Base Case**	(400, 40)	-0.3	0.05	(3, 0.3)	-1%	-1%	-2%	-2%
Short F-up	(400, 40)	-0.3	0.05	(1, 0.2)	2%	2%	2%	2%
Long F-up	(400, 40)	-0.3	0.05	(5, 1)	0%	0%	1%	1%
Large heterogeneity	(400, 40)	-0.3	0.1	(3, 0.3)	0%	0%	-2%	-1%
Small heterogeneity	(400, 40)	-0.3	0.001	(3, 0.3)	5%	5%	5%	5%
Large effect size	(400, 40)	-0.8	0.05	(3, 0.3)	-1%	-1%	-3%	-3%
No effect size	(400, 40)	0	0.05	(3, 0.3)	0%	0%	-1%	-1%
Small sample size	(100, 15)	-0.3	0.05	(3, 0.3)	1%	0%	1%	1%
Large sample size	(1000, 100)	-0.3	0.05	(3, 0.3)	-1%	-1%	-1%	-1%
Small P(Event)	(400, 40)	-0.3	0.05	(3, 0.3)	0%	0%	-1%	-1%

Large % R_cens+Small P(Event)	(400, 40)	-0.3	0.05	(5, 1)	-1%	-1%	-5%	0%
Long Follow-up+0% R_cens	(400, 40)	-0.3	0.05	(5, 1)	2%	2%	3%	4%
80% Power	(400, 40)	-0.3	0.2	(3, 0.3)	0%	0%	-1%	-1%

Scenario 0*: Large sample size, No effect size, No heterogeneity, Long F-Up; Base Case**: Medium sample size, medium effect size, medium heterogeneity, medium follow up; Other scenarios change from base case; R_cens=Random censoring; P(E)=P(Event)

Table 6.6: Percent difference between the modelled and empirical standard errors per simulation scenario across different meta-analysis models

Simulation Scenarios					**Methods**			
Run	No. of participants per trial (Mean, SD)	Mean Log HR	τ	Follow-up time (Mean, SD)	Two-stage Cox PH Model (HR)	Two-stage (“O-E” & V) (HR)	Two-stage clog-log (HR)	Two-stage Logit link (OR)
5 trials per MA								
Scenario 0*	(1000, 100)	0	0	(5, 1)	0%	0%	-44%	-91%
Base Case**	(400, 40)	-0.3	0.05	(3, 0.3)	0%	-2%	38%	-20%

Short F-up	(400, 40)	-0.3	0.05	(1, 0.2)	0%	2%	0%	-9%
Long F-up	(400, 40)	-0.3	0.05	(5, 1)	0%	-7%	182%	35%
Large heterogeneity	(400, 40)	-0.3	0.1	(3, 0.3)	0%	-2%	51%	-11%
Small heterogeneity	(400, 40)	-0.3	0.001	(3, 0.3)	0%	-2%	2%	-42%
Large effect size	(400, 40)	-0.8	0.05	(3, 0.3)	0%	-1%	17%	-21%
No effect size	(400, 40)	0	0.05	(3, 0.3)	0%	-2%	64%	-13%
Small sample size	(100, 15)	-0.3	0.05	(3, 0.3)	0%	-3%	18%	-33%
Large sample size	(1000, 100)	-0.3	0.05	(3, 0.3)	0%	-2%	53%	-9%
Small P(Event)	(400, 40)	-0.3	0.05	(3, 0.3)	0%	1%	20%	-11%
Large % R_cens+Small P(Event)	(400, 40)	-0.3	0.05	(5, 1)	0%	-4%	90%	9%
Long Follow-up+0% R_cens	(400, 40)	-0.3	0.05	(5, 1)	0%	-2%	-4%	-64%
80% Power	(400, 40)	-0.3	0.027	(3, 0.3)	0%	-2%	28%	-27%
20 trials per MA								
Scenario 0*	(1000, 100)	0	0	(5, 1)	0%	0%	-44%	-90%
Base Case**	(400, 40)	-0.3	0.05	(3, 0.3)	0%	-2%	33%	-22%
Short F-up	(400, 40)	-0.3	0.05	(1, 0.2)	0%	0%	0%	-10%
Long F-up	(400, 40)	-0.3	0.05	(5, 1)	0%	-7%	151%	18%
Large heterogeneity	(400, 40)	-0.3	0.1	(3, 0.3)	0%	-2%	48%	-13%

Small heterogeneity	(400, 40)	-0.3	0.001	(3, 0.3)	0%	-2%	-5%	-46%
Large effect size	(400, 40)	-0.8	0.05	(3, 0.3)	0%	0%	12%	-24%
No effect size	(400, 40)	0	0.05	(3, 0.3)	0%	-2%	57%	-17%
Small sample size	(100, 15)	-0.3	0.05	(3, 0.3)	0%	-4%	13%	-37%
Large sample size	(1000, 100)	-0.3	0.05	(3, 0.3)	0%	-2%	50%	-11%
Small P(Event)	(400, 40)	-0.3	0.05	(3, 0.3)	0%	0%	18%	13%
Large % R_cens+Small P(Event)	(400, 40)	-0.3	0.05	(5, 1)	0%	-5%	71%	-2%
Long Follow-up+0% R_cens	(400, 40)	-0.3	0.05	(5, 1)	0%	-2%	-2%	-62%
80% Power	(400, 40)	-0.3	0.2	(3, 0.3)	0%	-2%	61%	-3%

Scenario 0*: Large sample size, No effect size, No heterogeneity, Long F-Up; Base Case**: Medium sample size, medium effect size, medium heterogeneity, medium follow up; Other scenarios change from base case; R_cens=Random censoring; P(E)=P(Event)

Table 6.7: Relative (%) increase (or decrease) in precision per simulation scenario across different meta-analysis models.

Simulation Scenarios					Methods			
Run	No. of participants per trial (Mean, SD)	Mean Log HR	τ	Follow-up time (Mean, SD)	Two-stage Cox PH Model (HR)	Two-stage ("O-E" & V) (HR)	Two-stage clog-log (HR)	Two-stage Logit link (OR)
5 trials per MA								
Scenario 0*	(1000, 100)	0	0	(5, 1)	0.001	0.001	0.0017	0.0112
Base Case**	(400, 40)	-0.3	0.05	(3, 0.3)	0.0161	0.0163	0.0184	0.0201
Short F-up	(400, 40)	-0.3	0.05	(1, 0.2)	0.0351	0.0344	0.0352	0.0386
Long F-up	(400, 40)	-0.3	0.05	(5, 1)	0.0148	0.0159	0.0591	0.0514
Large heterogeneity	(400, 40)	-0.3	0.1	(3, 0.3)	0.0264	0.0269	0.0233	0.0296
Small heterogeneity	(400, 40)	-0.3	0.001	(3, 0.3)	0.0052	0.0052	0.0124	0.0091
Large effect size	(400, 40)	-0.8	0.05	(3, 0.3)	0.017	0.017	0.0447	0.0215
No effect size	(400, 40)	0	0.05	(3, 0.3)	0.0154	0.0157	0.0094	0.0178
Small sample size	(100, 15)	-0.3	0.05	(3, 0.3)	0.031	0.0319	0.0328	0.0465
Large sample size	(1000, 100)	-0.3	0.05	(3, 0.3)	0.0119	0.0121	0.0141	0.0133
Small P(Event)	(400, 40)	-0.3	0.05	(3, 0.3)	0.0184	0.0182	0.0188	0.021
Large % R_cens+Small P(Event)	(400, 40)	-0.3	0.05	(5, 1)	0.0155	0.0162	0.0326	0.0257

Long Follow-up+0% R_cens	(400, 40)	-0.3	0.05	(5, 1)	0.0138	0.0140	0.0143	0.0734
80% Power	(400, 40)	-0.3	0.027	(3, 0.3)	0.0111	0.0113	0.0158	0.0152
20 trials per MA								
Scenario 0*	(1000, 100)	0	0	(5, 1)	0.0002	0.0002	0.0004	0.0022
Base Case**	(400, 40)	-0.3	0.05	(3, 0.3)	0.0037	0.0037	0.0094	0.0047
Short F-up	(400, 40)	-0.3	0.05	(1, 0.2)	0.0088	0.0086	0.0089	0.0092
Long F-up	(400, 40)	-0.3	0.05	(5, 1)	0.0034	0.0037	0.0555	0.043
Large heterogeneity	(400, 40)	-0.3	0.1	(3, 0.3)	0.0061	0.0062	0.01	0.007
Small heterogeneity	(400, 40)	-0.3	0.001	(3, 0.3)	0.0012	0.0012	0.0082	0.0022
Large effect size	(400, 40)	-0.8	0.05	(3, 0.3)	0.004	0.0039	0.0344	0.0052
No effect size	(400, 40)	0	0.05	(3, 0.3)	0.0035	0.0036	0.0023	0.0043
Small sample size	(100, 15)	-0.3	0.05	(3, 0.3)	0.0072	0.0074	0.0134	0.0112
Large sample size	(1000, 100)	-0.3	0.05	(3, 0.3)	0.003	0.003	0.0083	0.0034
Small P(Event)	(400, 40)	-0.3	0.05	(3, 0.3)	0.0043	0.0043	0.0076	0.0053
Large % R_cens+Small P(Event)	(400, 40)	-0.3	0.05	(5, 1)	0.0037	0.0039	0.0273	0.0156
Long Follow-up+0% R_cens	(400, 40)	-0.3	0.05	(5, 1)	0.0032	0.0033	0.0033	0.0448
80% Power	(400, 40)	-0.3	0.2	(3, 0.3)	0.0111	0.0113	0.0109	0.0114

Scenario 0*: Large sample size, No effect size, No heterogeneity, Long F-Up; Base Case**: Medium sample size, medium effect size, medium heterogeneity, medium follow up; Other scenarios change from base case; R_cens=Random censoring; P(E)=P(Event)

Table 6.8: Mean squared error obtained per simulation scenario across different meta-analysis models.

	Simulation Scenarios				**Methods**			
Run	No. of participants per trial (Mean, SD)	Mean Log HR	τ	Follow-up time (Mean, SD)	Two-stage Cox PH Model (HR)	Two-stage ("O-E" & V) (HR)	Two-stage clog-log (HR)	Two-stage Logit link (OR)
5 trials per MA								
Scenario 0*	(1000, 100)	0	0	(5, 1)	96%	96%	95%	96%
Base Case**	(400, 40)	-0.3	0.05	(3, 0.3)	88%	88%	82%	89%
Short F-up	(400, 40)	-0.3	0.05	(1, 0.2)	95%	94%	94%	95%
Long F-up	(400, 40)	-0.3	0.05	(5, 1)	89%	88%	16%	54%
Large heterogeneity	(400, 40)	-0.3	0.1	(3, 0.3)	87%	87%	82%	87%
Small heterogeneity	(400, 40)	-0.3	0.001	(3, 0.3)	95%	94%	81%	94%
Large effect size	(400, 40)	-0.8	0.05	(3, 0.3)	89%	60%	88%	90%
No effect size	(400, 40)	0	0.05	(3, 0.3)	88%	88%	91%	91%

Small sample size	(100, 15)	-0.3	0.05	(3, 0.3)	92%	91%	91%	93%
Large sample size	(1000, 100)	-0.3	0.05	(3, 0.3)	88%	88%	77%	88%
Small P(Event)	(400, 40)	-0.3	0.05	(3, 0.3)	90%	90%	88%	92%
Large % R_cens+Small P(Event)	(400, 40)	-0.3	0.05	(5, 1)	89%	89%	53%	79%
Long Follow-up+0% R_cens	(400, 40)	-0.3	0.05	(5, 1)	87%	87%	88%	77%
80% Power	(400, 40)	-0.3	0.027	(3, 0.3)	89%	89%	80%	91%
20 trials per MA								
Scenario 0*	(1000, 100)	0	0	(5, 1)	97%	97%	96%	96%
Base Case**	(400, 40)	-0.3	0.05	(3, 0.3)	92%	93%	61%	94%
Short F-up	(400, 40)	-0.3	0.05	(1, 0.2)	93%	93%	93%	94%
Long F-up	(400, 40)	-0.3	0.05	(5, 1)	93%	93%	0%	6%
Large heterogeneity	(400, 40)	-0.3	0.1	(3, 0.3)	93%	93%	73%	93%
Small heterogeneity	(400, 40)	-0.3	0.001	(3, 0.3)	96%	96%	35%	95%
Large effect size	(400, 40)	-0.8	0.05	(3, 0.3)	92%	93%	15%	94%
No effect size	(400, 40)	0	0.05	(3, 0.3)	93%	93%	93%	93%
Small sample size	(100, 15)	-0.3	0.05	(3, 0.3)	95%	95%	78%	95%
Large sample size	(1000, 100)	-0.3	0.05	(3, 0.3)	93%	93%	54%	93%
Small P(Event)	(400, 40)	-0.3	0.05	(3, 0.3)	94%	94%	79%	94%

Large % R_cens+Small P(Event)	(400, 40)	-0.3	0.05	(5, 1)	93%	94%	9%	53%
Long Follow-up+0% R_cens	(400, 40)	-0.3	0.05	(5, 1)	93%	94%	94%	48%
80% Power	(400, 40)	-0.3	0.2	(3, 0.3)	94%	94%	84%	93%

Scenario 0*: Large sample size, No effect size, No heterogeneity, Long F-Up; Base Case**: Medium sample size, medium effect size, medium heterogeneity, medium follow up; Other scenarios change from base case; R_cens=Random censoring; P(E)=P(Event)

Table 6.9: Coverage obtained per simulation scenario across different meta-analysis models.

Simulation Scenarios					**Methods**			
Run	No. of participants per trial (Mean, SD)	Mean Log HR	τ	Follow-up time (Mean, SD)	Two-stage Cox PH Model (HR)	Two-stage (“O-E” & V) (HR)	Two-stage clog-log (HR)	Two-stage Logit link (OR)
5 trials per MA								
Scenario 0*	(1000, 100)	0	0	(5, 1)	5%	5%	5%	4%

Base Case**	(400, 40)	-0.3	0.05	(3, 0.3)	68%	68%	58%	58%
Short F-up	(400, 40)	-0.3	0.05	(1, 0.2)	34%	36%	33%	34%
Long F-up	(400, 40)	-0.3	0.05	(5, 1)	70%	70%	13%	14%
Large heterogeneity	(400, 40)	-0.3	0.1	(3, 0.3)	51%	51%	47%	48%
Small heterogeneity	(400, 40)	-0.3	0.001	(3, 0.3)	97%	97%	80%	81%
Large effect size	(400, 40)	-0.8	0.05	(3, 0.3)	100%	100%	100%	100%
No effect size	(400, 40)	0	0.05	(3, 0.3)	12%	12%	9%	10%
Small sample size	(100, 15)	-0.3	0.05	(3, 0.3)	43%	43%	28%	29%
Large sample size	(1000, 100)	-0.3	0.05	(3, 0.3)	76%	76%	71%	72%
Small P(Event)	(400, 40)	-0.3	0.05	(3, 0.3)	62%	62%	52%	53%
Large % R_cens+Small P(Event)	(400, 40)	-0.3	0.05	(5, 1)	69%	69%	36%	37%
Long Follow-up+0% R_cens	(400, 40)	-0.3	0.05	(5, 1)	72%	73%	72%	72%
80% Power	(400, 40)	-0.3	0.027	(3, 0.3)	80%	80%	66%	67%
20 trials per MA								
Scenario 0*	(1000, 100)	0	0	(5, 1)	3%	3%	4%	4%
Base Case**	(400, 40)	-0.3	0.05	(3, 0.3)	100%	100%	99%	99%
Short F-up	(400, 40)	-0.3	0.05	(1, 0.2)	86%	87%	87%	87%
Long F-up	(400, 40)	-0.3	0.05	(5, 1)	100%	100%	46%	48%

Large heterogeneity	(400, 40)	-0.3	0.1	(3, 0.3)	96%	96%	95%	95%
Small heterogeneity	(400, 40)	-0.3	0.001	(3, 0.3)	100%	100%	100%	100%
Large effect size	(400, 40)	-0.8	0.05	(3, 0.3)	100%	100%	100%	100%
No effect size	(400, 40)	0	0.05	(3, 0.3)	7%	7%	7%	8%
Small sample size	(100, 15)	-0.3	0.05	(3, 0.3)	93%	93%	76%	77%
Large sample size	(1000, 100)	-0.3	0.05	(3, 0.3)	100%	100%	100%	100%
Small P(Event)	(400, 40)	-0.3	0.05	(3, 0.3)	99%	99%	97%	97%
Large % R_cens+Small P(Event)	(400, 40)	-0.3	0.05	(5, 1)	100%	100%	88%	88%
Long Follow-up+0% R_cens	(400, 40)	-0.3	0.05	(5, 1)	100%	100%	100%	100%
80% Power	(400, 40)	-0.3	0.2	(3, 0.3)	80%	80%	84%	83%

Scenario 0*: Large sample size, No effect size, No heterogeneity, Long F-Up; Base Case**: Medium sample size, medium effect size, medium heterogeneity, medium follow up; Other scenarios change from base case; R_cens=Random censoring; P(E)=P(Event)

Table 6.10: Power obtained per simulation scenario across different meta-analysis models.

Salika et al.
BMC Medical Research Methodology (2022) 22:73
https://doi.org/10.1186/s12874-022-01541-9

BMC Medical Research Methodology

RESEARCH Open Access

Implications of analysing time-to-event outcomes as binary in meta-analysis: empirical evidence from the Cochrane Database of Systematic Reviews

Theodosia Salika*, Rebecca M. Turner, David Fisher, Jayne F. Tierney and Ian R. White

Abstract

Background: Systematic reviews and meta-analysis of time-to-event outcomes are frequently published within the Cochrane Database of Systematic Reviews (CDSR). However, these outcomes are handled differently across meta-analyses. They can be analysed on the hazard ratio (HR) scale or can be dichotomized and analysed as binary outcomes using effect measures such as odds ratios (OR) or risk ratios (RR). We investigated the impact of reanalysing meta-analyses from the CDSR that used these different effect measures.

Methods: We extracted two types of meta-analysis data from the CDSR: either recorded in a binary form only ("binary"), or in binary form together with observed minus expected and variance statistics ("OEV"). We explored how results for time-to-event outcomes originally analysed as "binary" change when analysed using the complementary log–log (clog-log) link on a HR scale. For the data originally analysed as HRs ("OEV"), we compared these results to analysing them as binary on a HR scale using the clog-log link or using a logit link on an OR scale.

Results: The pooled HR estimates were closer to 1 than the OR estimates in the majority of meta-analyses. Important differences in between-study heterogeneity between the HR and OR analyses were also observed. These changes led to discrepant conclusions between the OR and HR scales in some meta-analyses. Situations under which the clog-log link performed better than logit link and vice versa were apparent, indicating that the correct choice of the method does matter. Differences between scales arise mainly when event probability is high and may occur via differences in between-study heterogeneity or via increased within-study standard error in the OR relative to the HR analyses.

Conclusions: We identified that dichotomising time-to-event outcomes may be adequate for low event probabilities but not for high event probabilities. In meta-analyses where only binary data are available, the complementary log–log link may be a useful alternative when analysing time-to-event outcomes as binary, however the exact conditions need further exploration. These findings provide guidance on the appropriate methodology that should be used when conducting such meta-analyses.

Keywords: Time-to-event, Meta-analysis, Methodology, Survival data, Clinical trials, Cochrane database of systematic reviews

*Correspondence: theodosia.salika.18@ucl.ac.uk
MRC Clinical Trials Unit, Institute of Clinical Trials and Methodology, University College London, London, UK

Background

Systematic reviews and meta-analyses of time-to-event outcomes (e.g. time to death, recurrence of symptoms, relief of pain etc.) are frequently carried out in areas such

as cancer, respiratory and cardiovascular diseases, since event timings are crucial to assessing the impact of an intervention [1]. The decision on how time-to-event outcomes are handled in a particular meta-analysis largely depends on how eligible studies are reported, and is usually out of the control of the meta-analyst except if individual participant data (IPD) are available. The information extracted by systematic reviewers may include the total number of participants and events per arm, and/or the hazard ratio alongside its confidence interval, and/or the log-rank observed minus expected statistic ("O-E") and its variance ("V") (which are useful alternative statistics if a hazard ratio is not directly reported [1]). Time-to-event data can be analysed using the effect measure of hazard ratio (HR), or can be dichotomised and analysed as binary using effect measures such as the odds ratio (OR) or risk ratio (RR) [2]. Although HR is considered the most appropriate scale for analysis of time-to-event data, in practice OR and RR are frequently used instead due to the following reasons: unavailability of individual participant data (IPD); limitations on how these outcomes are reported in individual trial reports; lack of familiarity in handling time-to-event outcomes for meta-analysis; difficulties in understanding the methods of analysing such data without a statistician; limited available training for the majority of systematic reviewers and meta-analysts who perform such analyses [3].

In the past, research was conducted comparing the differences between the OR using logistic regression models and the HR using proportional hazard (PH) models within individual studies. Green and Symons [4] showed that logistic and Cox PH models produce similar results when the event is rare and for shorter follow-up times under a constant hazard rate. Ingram and Kleinman [5] added that important differences among the methods occur in the presence of varying censoring rates and length of follow-up. However, it has not been established yet how such results transfer to the context of an aggregate data meta-analysis for which summary data is extracted from trial reports. Further, in this context it is of interest to examine potential alternatives such as the use of the complementary log–log link, which may reduce the difference in the results between the two effect measures used. The overall meta-analytic estimate can be affected due to changes to the weighting allocated to each study, and therefore changes to the results can be unpredictable. We aimed to carry out an empirical "meta-epidemiological" study using survival meta-analysis data from the Cochrane Database of Systematic Reviews (CDSR) (Issue 1, 2008) to explore the implications of analysing time-to-event outcomes as binary in meta-analysis. We assessed the importance of extracting suitable data such as the "O-E" and "V" statistics rather than binary summaries to perform such analyses; in the occasion where binary data were available we examined whether the use of alternative methodology such as the complementary log–log link (clog-log), proven to facilitate interpretation of the results on a HR scale [6, 7] can minimise the error we may observe in the results. We assess only the differences between the OR and the HR, as the RR, according to the literature [8–11], is placed in between these measures and therefore, we expect to capture any bias within these extremes. We perform these analyses under both two- and one-stage models.

The rest of the paper is set out as follows. In the methods section, we describe the dataset we used and the statistical models that we applied. In the results, we present descriptive statistics of the database and then we describe the results obtained from reanalysing the data originally analysed as binary on an HR scale and from reanalysing the data originally analysed using "O-E" and "V" data on an OR scale. These results are followed by a discussion exploring the strengths and limitations of our findings, together with conclusions and implications.

Methods

Data

The Nordic Cochrane Centre provided the content of the first issue from 2008 of the CDSR. The database includes meta-analyses within reviews which have been classified previously by outcome type, medical specialty and types of interventions included in the pairwise comparisons [12]. The database did not record whether data type was time-to-event; however, based on the outcome classification we were able to identify (using words such as "survival", "death", "fatality") three sets of time-to-event meta-analyses:

- "binary": Those with outcome classification "all-cause mortality" where the information recorded was based only on the number of events and participants per arm;
- "OEV": Those with outcome classifications "overall survival" and "progression/disease free survival" where the information recorded was based on "binary" data in addition to log-rank "O-E" and "V" statistics"; these were originally analysed as HRs in the RevMan software;
- Those with estimated log HR and its standard error. These were removed from further analyses since there was no available information on the number of events and participants per arm and therefore no binary data meta-analysis could be conducted.

Therefore, we identified two subsets of time-to-event meta-analyses: those with binary summaries, and those

with binary summaries in addition to OEV data; we analysed each outcome per dataset separately to assess whether differences exist due to different characteristics of the outcomes. We also examined whether the information obtained from "OEV" data was based on aggregate data or IPD by examining the individual Cochrane reviews.

Eligibility Criteria

RMT (for "binary" data) and TS (for "OEV" data) initially extracted these data and conducted cleaning including examination of the outcome classification; TS repeated the "binary" data extraction to confirm the information obtained were accurate and RMT confirmed the choice of included meta-analyses obtained from "OEV" data extraction. Both datasets could contribute more than one meta-analysis per Cochrane review. RMT and TS identified 46 misclassifications due to disagreement with the original outcome classification as listed in the datasets, conflicting information in the database or unavailability of the correct version of the Cochrane review. We excluded 1,284 studies including double zero events, since they do not contribute to the meta-analysis results [12, 13]. We removed another 359 meta-analyses including fewer than 3 studies because some of the models applied below (i.e. generalised linear mixed models) will be affected by estimation issues and inevitable failures using small numbers of studies [14]; hence we wanted to make fair comparisons between the models applied. Derivation of the analysis sample is provided in Fig. 1.

Descriptive statistics

We describe the number of studies per meta-analysis, number of events and study size by the median and interquartile range. We also identify the number of medical specialities, and median number of events (and interquartile range) per medical specialty.

Model description for "binary" data

We used the following meta-analysis models to analyse the data on the OR or HR scale. The first was a model proposed for "binary" data (assuming a binomial likelihood

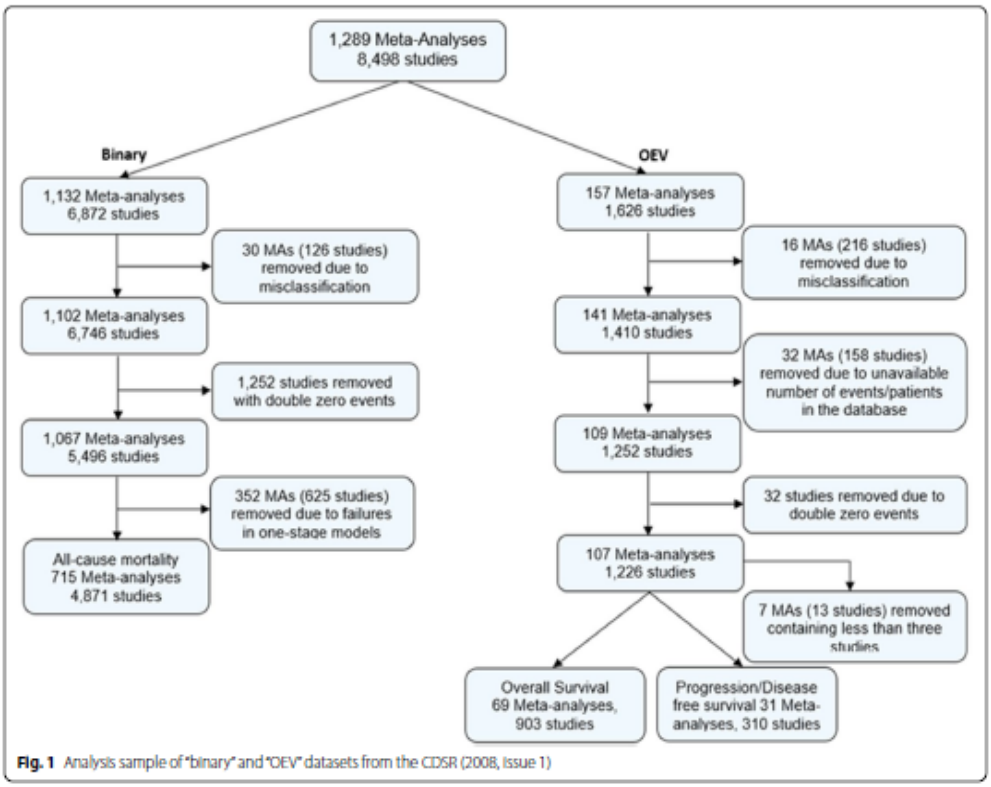


Fig. 1 Analysis sample of "binary" and "OEV" datasets from the CDSR (2008, Issue 1)

with a logit link) which is based only on the number of patients and number of events which occurred. Interpretation for the treatment effect is conducted in terms of the logarithm of an OR.

In the second approach, we modelled the binary data using a normal approximation to binomial likelihood with a complementary log–log link (clog-log), where treatment effect interpretation was based on the logarithm of a HR. This method is also based only on the number of patients and events which occurred, and ignores censoring and the time element; however it is closely related to continuous-time models, has a built-in proportional hazards assumption, and therefore has important application in survival analysis [6].

Fitting two-stage random-effects models for "binary" data

Prior to fitting the two-stage random-effects models, study arms with zero events were identified for the "binary" data. For 771 studies, a "treatment arm" continuity correction was applied as proposed by Sweeting et al. [15] and was constrained to sum to one as this ensures that the same amount of information is added to each study.

Let $i = 1, 2, \ldots, n$ denote the study. The estimated log odds and log hazard ratios were given by:

$$y_i = \begin{cases} \log\left(\frac{A_i}{B_i}\right) - \log\left(\frac{C_i}{D_i}\right) \text{for ORs} & (1) \\ \log\left[-\log\left(1-P_{Ti}\right)\right] - \log\left[-\log\left(1-P_{Ci}\right)\right] \text{for HRs} & (2) \end{cases}$$

where A_i, C_i represented number of events, B_i, D_i represented number of non-events in the treatment and control groups respectively, $P_{Ti} = \frac{A_i}{A_i+B_i}$ was the proportion of events on the treatment arm of the i^{th} study, and $P_{Ci} = \frac{C_i}{C_i+D_i}$ was the proportion of events on the control arm of the i^{th} study.

The corresponding variances were given by:

$$s_i^2 = \begin{cases} \frac{1}{A_i} + \frac{1}{B_i} + \frac{1}{C_i} + \frac{1}{D_i} \text{ for ORs} & (3) \\ \left(\frac{1}{\log\left(1-P_{Ti}\right)*\left(P_{Ti}-1\right)}\right)^2 * \left(\frac{P_{Ti}*(1-P_{Ti})}{A_i+B_i}\right) + \left(\frac{1}{\log(1-P_{Ci})*(P_{Ci}-1)}\right)^2 * \left(\frac{P_{Ci}*(1-P_{Ci})}{C_i+D_i}\right) \text{ for HRs} & (4) \end{cases}$$

Equations 2 and 4 provided a HR estimate via the use of the complementary log–log link considered as a useful link function for the discrete-time hazards models as recommended by Hedeker et al. [7] and Singer et al. [6]. We estimated the study-specific log odds ratios or log hazard ratios, y_i and their within-study variances s_i^2 as shown above and fitted a standard two-stage random-effects model to these. Additionally, we obtained the I^2 statistic from the fitted models as follows:

$$I^2 = \frac{\hat{\tau}^2}{\hat{\tau}^2 + \hat{\sigma}^2}$$

where τ^2 denotes the variance of the underlying true effects across studies and σ^2 the typical within-study variance.

To avoid downward bias in the variance components estimates, we used the REML estimator for model implementation [16]. The models were implemented via the "rma.uni" command from "metafor" package in R. We also fitted one-stage random-effects models for "binary" data. The methods related to one-stage meta-analysis models and code is available in Additional file 1.

Model description for "OEV" data

For "OEV" data, the "O-E" and "V" statistics were available in the Cochrane database alongside the number of patients and events. These data came either from published reports or from IPD; TS examined the individual reviews from the Cochrane database and assessed the data origin. Since there were more available information for these data the following three models were applied, using only two-stage meta-analysis models.

Similarly to "binary" data, we initially analysed the "OEV" data as "binary" and modelled them as described in detail in the preceding section. We also used the log-rank Observed—Expected events (O-E) and the log-rank Variance (V) statistics calculated previously from the number of events and the individual times to event on each research arm of the trial; we used the log-rank approach [17] in order to obtain another type of HR estimate. We used random-effects models to analyse the data throughout, including between-study heterogeneity to account for variation across studies.

Fitting two-stage random-effects models for "OEV" data

Similarly to the "binary" data, the estimated log odds and log hazard ratios were given by Eqs. 1 and 2 for the binary summaries while the "O-E" and "V" statistics were used as follows:

$$y_i = \frac{logrank\ Observed - Expected\ events\ (O-E)}{logrank\ Variance\ (V)} \text{ for HRs} \quad (5)$$

The corresponding variances were given by Eqs. 3 and 4 for binary summaries while for "O-E" and "V" statistics as follows:

$$s_i^2 = \frac{1}{logrank\ Variance\ (V)}\ for\ HRs \quad (6)$$

where V denotes the variance of the logrank statistic. We used the REML estimator for model implementation [16] and the models were implemented via the "rma.uni" command from "metafor" package in R.

Model comparison for "binary" data

The following model comparisons were performed. For the "binary" data set, we examined whether the results from analysing survival data as binary on an OR scale are similar to results from analysing on the HR scale using the clog-log link, both under two-stage and one-stage models. For presentation purposes, we present only comparisons of the results under two-stage models in the main paper (and for one-stage models in the Additional file 1) in order to assess the discrepancies between the model using the logit link and the model using the complementary log–log link.

First, we examined the proportion of significant and non-significant meta-analytic pooled effect estimates under the different scales used (OR vs HR scale); we identified the number of meta-analyses which were significant under one scale and non-significant under the other at a two-sided 5% level of significance.

Bland–Altman plots with associated 95% limits of agreement were constructed, with the aim of facilitating interpretation of results and producing fair comparisons between the two scales [18]. In order to create these plots, results were standardised by dividing the logarithm of the estimate by its standard error. Plots were produced for the standardised treatment effect estimates and for the I^2 statistics. I^2 represents the percentage of variability that is due to between-study heterogeneity rather than chance; I^2 values range from 0 to 100%. This measure was chosen for model comparison as it enables us to compare results directly between the two scales used. The variance of underlying true effects across studies (τ^2) was not used as it does not allow direct comparison between different outcome measures.

We identified "outliers" as meta-analyses outside the 95% limits of agreement, and we examined their characteristics. The meta-analysis characteristics we examined were the following:

- between-scale differences in the magnitude of the pooled treatment effect estimate and its 95% confidence intervals
- the levels of within-study standard error and between-study heterogeneity and study weights in the meta-analysis
- study-specific event probabilities and baseline risk

We summarised these differences by meta-analysis and reported those characteristics which were mostly associated with substantial differences between OR pooled effect estimates and corresponding HR pooled effect estimates.

Model comparison for "OEV" data

For the "OEV" data set, comparisons on overall and progression disease free survival outcomes were conducted separately; this was because differences between these outcomes might be observed in the presence of different disease severities, and therefore this would be associated with different length of follow-up and risk of the outcome.

For both outcomes, we performed comparisons by examining the differences between analysing the data as binary on an OR scale, analysing the data as binary using the clog-log link on a HR scale, or analysing the data using the "O-E" and "V" statistics on a HR scale. We assessed whether the differences observed from analysing the data as binary on an OR scale could be reduced by the use of the clog-log link. We present only comparisons of the results under two-stage models since there were no available IPD to perform comparisons under one-stage models.

Similarly to "binary" data, we examined the proportion of significant and non-significant meta-analytic pooled effect estimates under the different scales used and identified the number of meta-analyses which were significant under one scale and non-significant under the other. We created Bland–Altman plots for the standardised treatment effect estimates and for the I^2 statistics to explore the agreement among the methods producing fair comparisons between the two scales [18]. Meta-analyses outside the 95% limits of agreement were examined for their characteristics.

Results

Results for "binary" data

For the outcome of "all-cause mortality", 1,132 meta-analyses within the Cochrane database were originally analysed as binary. The median number of meta-analyses per review was 1 with IQR (1,2). The median number of studies and the median number of events are provided in Table 1, indicating that these numbers were a lot smaller than those obtained for the "OEV" data.

The distribution of medical specialities of the meta-analyses is presented in Table 2. For the "binary" data,

Table 1 Descriptive statistics for "binary" and "OEV" data from the CDSR

	"binary"		
Outcome	**All-cause Mortality**		
Total Number of MA	715		
Number of studies per MA: Median (IQR)	5 (3, 8)		
Number of events per MA: Median (IQR)	13 (4, 40)		
Median Study Size (IQR)	124 (60, 312)		
	"OEV"		
Outcome	**Overall Survival**		**Progression/ Disease Free Survival**
Total Number of MA	69		31
Number of studies per MA: Median (IQR)	10 (6, 14)		10 (7, 14)
Number of events per MA: Median (IQR)	108 (58, 254)		104 (70, 192)
Median Study Size (IQR)	182 (93, 369)		185 (90, 317)

Table 2 Distribution of medical specialties for the "binary" and "OEV" data meta-analyses in the CDSR

	"binary"			
Medical Specialty	**ACM[b] Number (%) of MAs**	**Events per MA: Median (IQR)**		
Cancer	95 (13%)	49 (17, 120)		
Cardiovascular	168 (23%)	14 (4, 43)		
Central nervous system/musculoskeletal	44 (6%)	12 (5, 33)		
Digestive/endocrine, nutritional and metabolic	71 (10%)	7 (3, 18)		
Gynaecology, pregnancy and birth	87 (12%)	7 (2, 20)		
Infectious diseases	46 (6%)	18 (8, 47)		
Mental health and behavioural conditions	21 (3%)	2 (1, 5)		
Pathological conditions, symptoms and signs	5 (1%)	9 (2, 15)		
Respiratory diseases	87 (12%)	11 (5, 36)		
Urogenital	30 (4%)	4 (2, 12)		
Other[a]	61 (9%)	9 (3, 27)		
	"OEV"			
Medical Specialty	**OS[c]: Number (%) of MAs**	**Events per MA: Median (IQR)**	**PDFS[d]: Number (%) of MAs**	**Events per MA: Median (IQR)**
Cancer	60 (87%)	104 (45, 221)	31 (100%)	116 (56, 243)
Digestive/endocrine, nutritional and metabolic	1 (1%)	52 (35, 64)	-	-
Infectious diseases	8 (12%)	482 (160, 1109)	-	-

[a] Other: Blood and immune system, General heath, Injuries, Mouth and dental, and Cystic fibrosis

[b] *ACM* All-cause mortality;

[c] *OS* Overall Survival;

[d] PDFS: Progression/Disease free survival

"Cardiovascular" (23%) is the most frequently occurring category, followed by "Cancer" (13%), "Gynaecology, pregnancy and birth" (12%) and "respiratory diseases" (12%). The median number of events in cancer substantially exceeded the median number of events in other medical areas.

Once the models were applied, we compared results between OR and HR analyses. Table 3 provides the percentages of significant and non-significant meta-analyses at a two-sided 5% level of significance indicating that there are few discrepancies present for both "binary" and "OEV" datasets under two-stage models.

According to the Bland–Altman plot (Fig. 2), the average difference between the two methods for the standardised pooled effect estimates was -0.004 units (-0.222 units, 0.214 units) and -0.1% (-10.6%, 10.3%) for the

Table 3 Number (%) of (non-)significant meta-analyses under different scales for two-stage models ("binary" and "OEV" data)

	Outcome		OR		HR (O-E & V)	
"binary"						
			Significant	Non-significant	Significant	Non-Significant
HR (clog-log)	All-cause mortality	Significant	106 (15%)	2 (0.1%)		
		Non-significant	4 (0.6%)	603 (84%)		
"OEV"						
			Significant	Non-significant	Significant	Non-Significant
HR (clog-log)	Overall Survival	Significant	20 (29%)	1 (0.2%)	18 (26%)	10 (14%)
		Non-significant	1 (0.2%)	47 (68%)	3 (4%)	38 (55%)
	Progression / Disease free Survival	Significant	9 (29%)	0 (0%)	8 (26%)	6 (19%)
		Non-significant	1 (3%)	21 (68%)	1 (3%)	16 (52%)
HR (O-E &V)	Overall Survival	Significant	18 (26%)	10 (14%)		
		Non-significant	3 (4%)	38 (55%)		
	Progression / Disease free Survival	Significant	9 (29%)	5 (16%)		
		Non-significant	1 (3%)	16 (52%)		

estimation of I^2 for two-stage models; this indicates a relatively small percentage difference between the two methods in the estimation of the measure of impact of heterogeneity I^2. The width of the 95% limits of agreement is small, indicating acceptable agreement between the two methods except in specific circumstances mentioned below. The corresponding results for one-stage models are presented in Additional file 1.

Based on Bland–Altman plots, 6% ($n=47$) of the meta-analyses were considered as outliers. In 21% of the "binary" outlying meta-analyses (e.g. MA 327; outlier obtained from I^2 estimates) a high event probability (defined here as probability greater than 0.7 for the majority of the individual studies) was observed. For example, meta-analysis 327 consists of 7 studies for which the event probability was greater than 0.7 for 5 out of 7 studies; consequently, high event probability affected substantially the differences in the individual study estimates between the OR and HR analyses, leading to different allocated relative weights for the studies, and discrepancies in the pooled effect estimates as shown in Fig. 3.

The pooled HR estimates were closer to 1 than the OR estimates in the majority of meta-analyses (Additional file 1; outlier obtained from standardised and I^2 estimates) with the exception of MA 574 for "binary" data where, even though most of the individual study HR estimates are closer to 1 than the individual OR estimates, the pooled HR estimate is further from 1 than the pooled OR estimate. Increased within-study variability on the OR scale relative to the HR scale may affect the weighting more than the actual estimates in the studies, for example within "binary" data meta-analysis 7 (Additional file 1; outlier obtained from standardised estimates), producing some differences in the pooled effect estimates between the two scales. Important differences in between-study heterogeneity between the HR and OR analyses were also observed. For example, meta-analysis 330 (outlier obtained from I^2 estimates) consists of 8 studies of which 6 are smaller studies which received increased weight in the HR analysis compared to the OR analysis while the two larger studies received smaller weights; this affected both the individual HR estimates that have moved closer to each other and the relevant weights of the studies as presented in Fig. 4.

In 34% of the outlying meta-analyses, the individual study estimates and the corresponding weights were affected by a combination of differing event probability across study arms, differences in between-study heterogeneity or increased within-study variability on the OR relative to the HR scale. In the presence of a limited amount of studies in the meta-analyses this was even more evident. Additional examples of forest plots indicating the discrepancies among the results are shown in Additional file 1.

Results for "OEV" data

In the Cochrane database, 157 meta-analyses were originally analysed using the "O-E" and "V" statistics on a HR scale. The median number of meta-analyses per review was 2 with IQR (2, 3). We observed that analysing time-to-event outcomes as HRs is restricted to very few medical specialties (Tables 2). For the "OEV" data, "Cancer" was still the most frequent medical specialty for both outcomes as observed in "binary" data (Table 2).

Table 3 provides the percentages of significant and non-significant meta-analyses for each outcome for

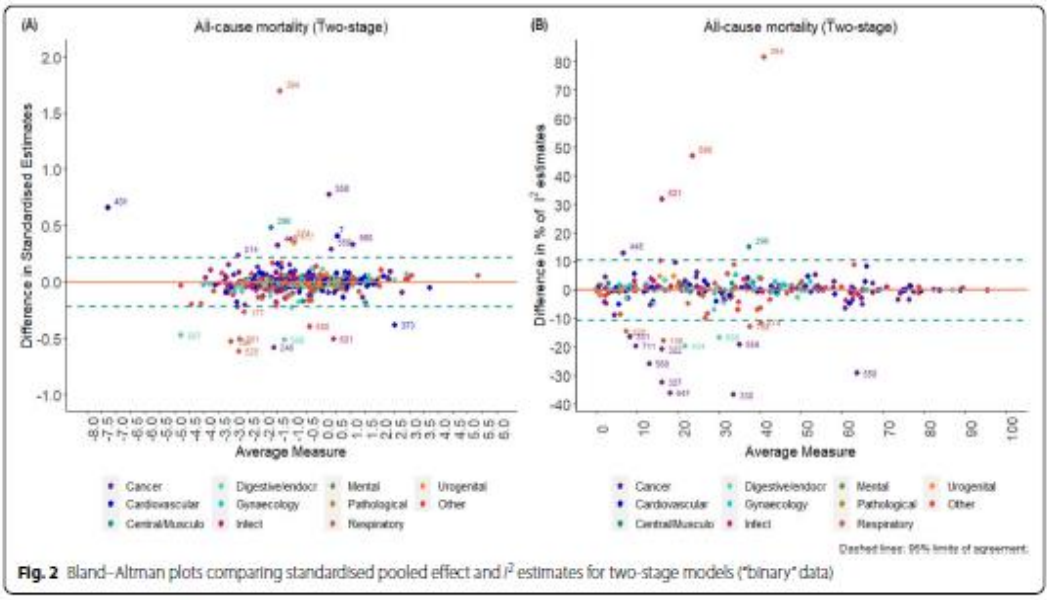


Fig. 2 Bland–Altman plots comparing standardised pooled effect and I^2 estimates for two-stage models ("binary" data)

two-stage models, indicating that discrepancies are more prevalent in the "OEV" data compared to the "binary" data; additionally the amount of discrepancies observed in statistical significance from the comparison of OR and HR obtained from the clog-log link was smaller than the amount of discrepancies observed between the OR and HR analyses.

Bland–Altman plots produced for "OEV" data indicated that the average difference between each pair of methods is larger than those obtained from the "binary" data (Figs. 5 and 6). For example, for overall survival, the average difference between the two methods for the standardised pooled effect estimates was 0.2 units (-1.8 units, 2.1 units) for OR versus HR and 0.2 units (-2.2 units, 2.5 units) for HR using clog-log versus HR; however, for OR vs HR clog-log differences the average bias was 0 units (-2.6 units, 2.7 units) indicating that clog-log is a closer approximation to OR rather than HR analyses

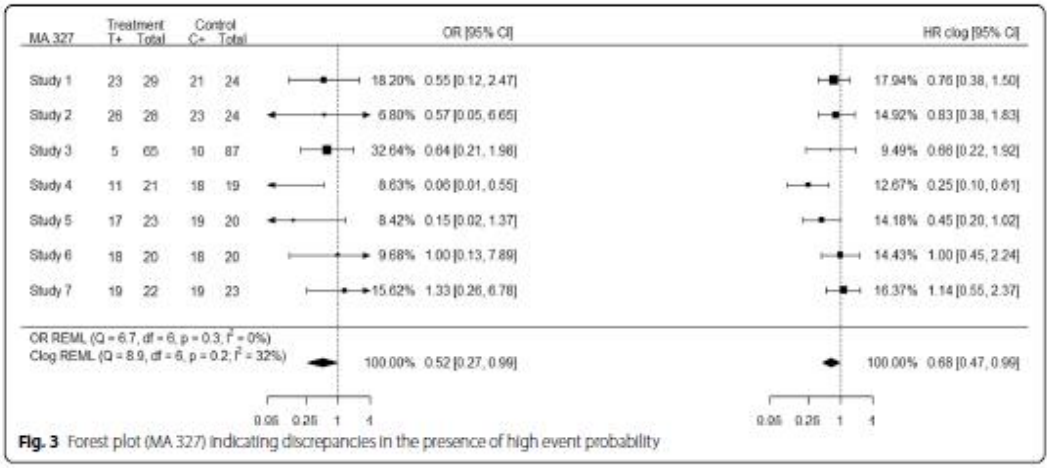


Fig. 3 Forest plot (MA 327) indicating discrepancies in the presence of high event probability

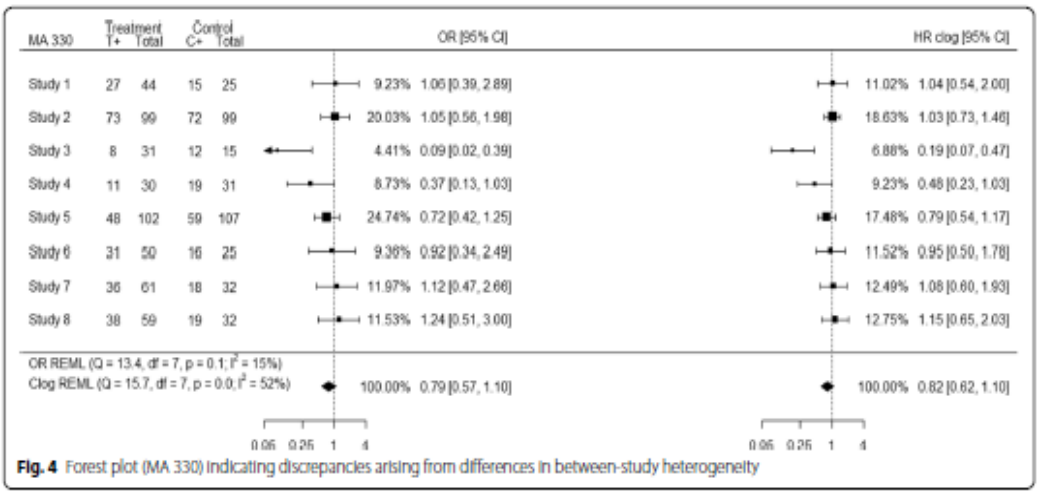


Fig. 4 Forest plot (MA 330) indicating discrepancies arising from differences in between-study heterogeneity

(Fig. 5). For the estimation of I^2, the average difference between the methods is -6% (-41%, 29%) for OR versus HR, -6% (-42%, 31%) for HR using clog-log versus HR, and 0% (-21%, 21%) for OR vs HR clog-log differences; similarly the clog-log seems a closer approximation to OR analyses rather than HR analyses (Fig. 6). The corresponding results for the outcome of progression/disease free survival are shown in Additional file 1.

Outliers were considered 28% of the "OEV" meta-analyses. Of these, 57% were from IPD rather than non-IPD and 54% of them were for the outcome of overall survival. In 50% of the outliers a high event probability (defined here as probability greater than 0.7) was observed, suggesting that this may be an important factor associated with differences among the scales used. For example, meta-analysis 45 (outlier obtained from standardised estimates) consists of 7 studies for which the event probability was greater than 0.7 for all the studies; consequently high event probability affected substantially the differences in the individual study estimates between the OR and HR analyses, leading to different allocated relative weights for the studies, and discrepancies in the pooled effect estimates as shown in Fig. 7. Even though the individual HR clog-log estimates were closer to the individual OR estimates the final pooled effect estimate was closer to the pooled HR estimate; this was not though the case for all meta-analyses.

Increased within-study variability on the OR scale relative to the HR scale may affect the weighting more than the actual estimates in the studies, for example for meta-analysis 17 (Additional file 1; outlier obtained from standardised estimates), producing differences in the pooled effect estimates between the two scales. Similarly, even though the individual study estimates and weights of OR and HR clog-log were closer to each other, the HR clog-log pooled effect estimate was closer to the pooled HR estimate; however, this was not the case for all meta-analyses. Important differences in between-study heterogeneity between the HR and OR analyses were observed in meta-analyses such as 42, 90. For example, meta-analysis 90 (outlier obtained from I^2 estimates) consists of 11 studies out of which 8 are smaller studies and 3 are larger studies. Smaller studies received increased weight in the HR analysis compared to the OR analysis, while larger studies received smaller weights in the HR scale compared to OR scale. However, this was not the case on the HR clog-log scale as presented in Fig. 8.

In 46% of the outlying meta-analyses, the individual study estimates, and the corresponding weights were affected by a combination of differing event probability across study arms, differences in between-study heterogeneity or increased within-study variability on the OR relative to the HR scale. In the presence of a limited amount of studies in the meta-analyses this was even more evident. Additional forest plots indicating the discrepancies among the results are shown in Additional file 1.

Overall, using the "OEV" data, a mixed pattern was observed. In 39% ($n=11$) of outlying meta-analyses the OR pooled effect estimate was closer to HR pooled effect estimate; however in 4 out of 11 outlying meta-analyses the individual study estimates obtained from the HR clog-log link were a closer approximation to the individual study HR estimates. Similarly, even though in 61% ($n=17$) of the outlying meta-analyses the HR clog-log pooled effect estimate was closer to the pooled

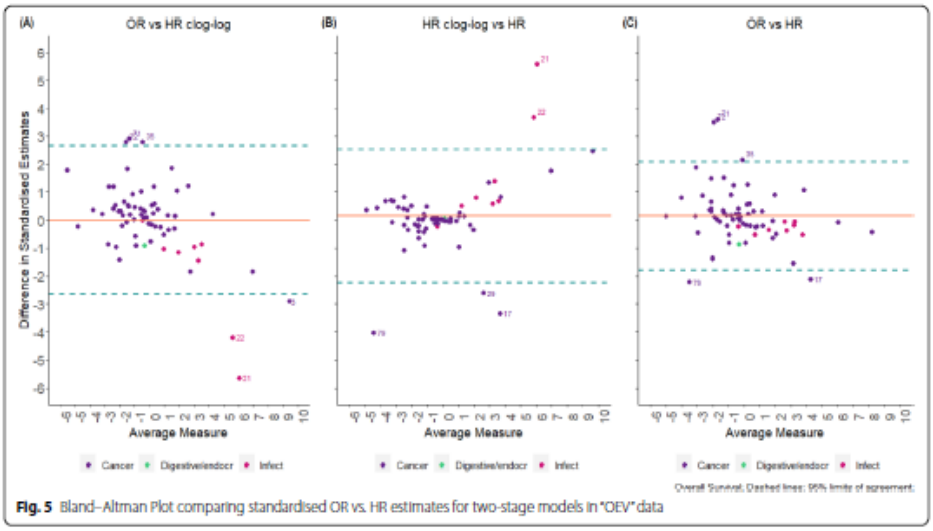


Fig. 5 Bland–Altman Plot comparing standardised OR vs. HR estimates for two-stage models in "OEV" data

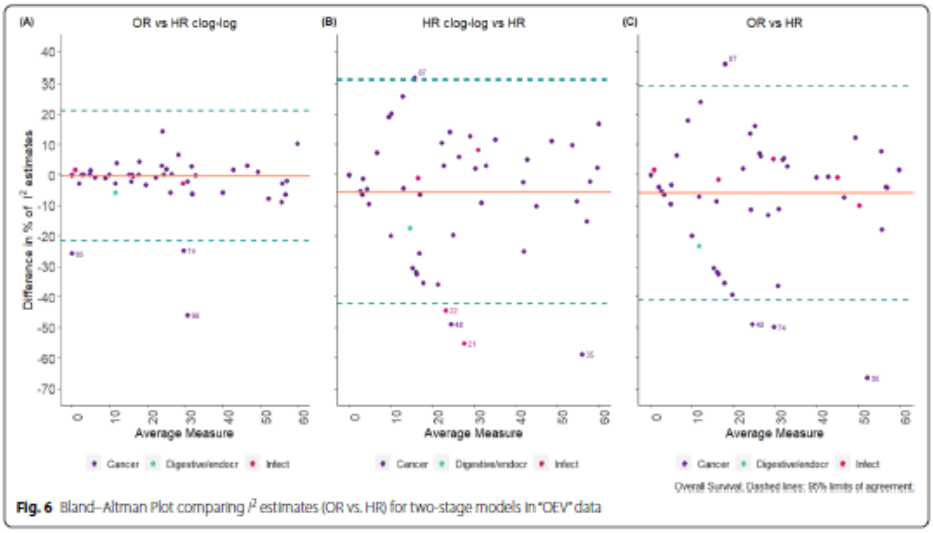


Fig. 6 Bland–Altman Plot comparing I^2 estimates (OR vs. HR) for two-stage models in "OEV" data

HR estimate, 3 of outlying meta-analyses provided individual study OR estimates closer to individual study HR estimates, and another 3 individual study HR clog-log estimates were closer to individual study OR estimates.

Discussion

Using meta-analysis data from the CDSR of 2008, we investigated how time-to-event outcomes are treated within meta-analysis; we explored the differences that occur when data are analysed as binary as opposed to

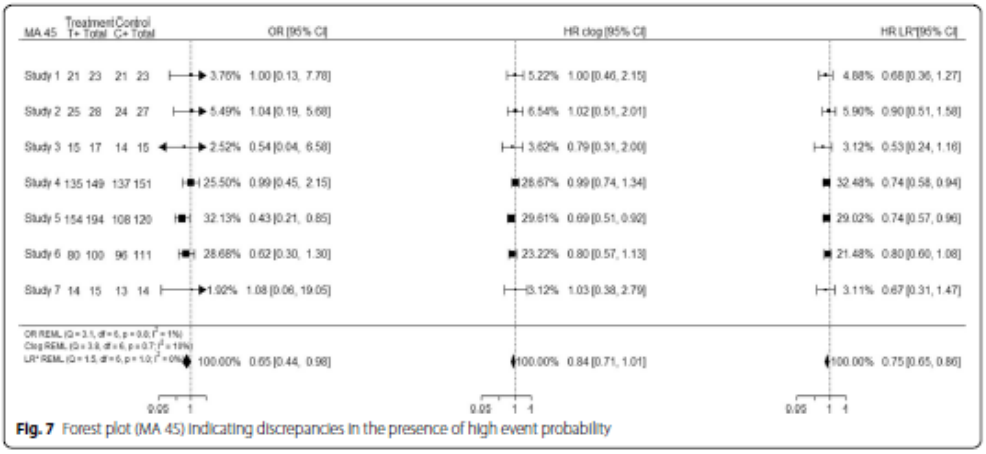


Fig. 7 Forest plot (MA 45) indicating discrepancies in the presence of high event probability

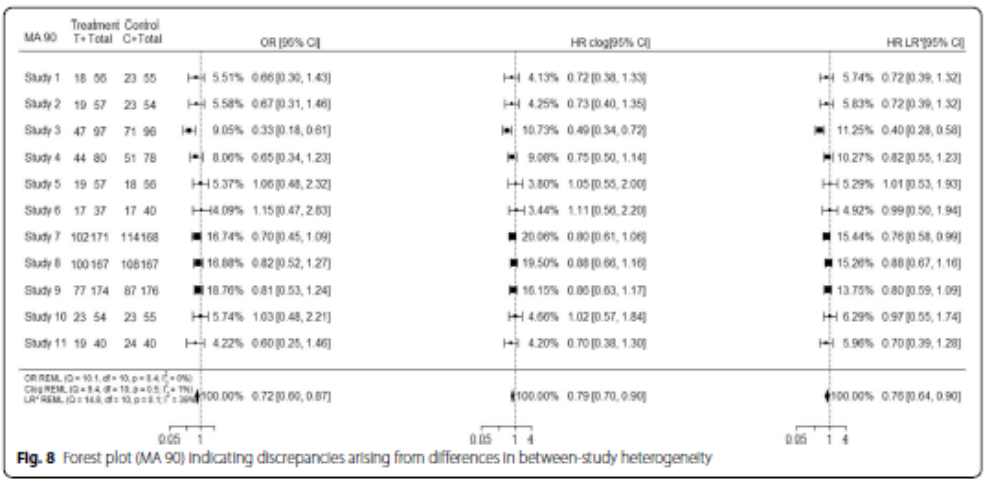


Fig. 8 Forest plot (MA 90) indicating discrepancies arising from differences in between-study heterogeneity

analysing the data using the complementary log–log link or using the "O-E" and "V" statistics where interpretation is conducted on a HR scale. For both datasets, we identified important reasons associated with discordance among the results, indicating that the correct choice of the method does matter and may affect the interpretation and conclusions drawn from the results. Our analyses highlighted that high event probability was an important factor associated with discordant effect estimates; changes to between and within-study variation were important mechanisms producing differences in the results as well. However, there were occasions where there was no clear single factor driving the differences, since there was a combination of reasons affecting the individual study estimates and corresponding weights. Regarding method selection, based on the "OEV" data we identified that a mixed pattern was observed and there was no clear indication under which exact conditions the clog-log link outperforms logit link on an OR scale and vice versa.

While most of the meta-analyses within the database were analysed originally as binary, with an outcome classification of all-cause mortality it is worth mentioning that these meta-analyses could include the outcome of short-term mortality (e.g. 30 days) or longer-term mortality (e.g. 5 years); therefore some of these meta-analyses with short follow-up may have been appropriately analysed as binary. The outcome classification of all-cause mortality was considered a representative sample of survival meta-analysis up to 2008, however results might be different for other outcomes and results might have changed in later reviews where more information on methodology was available. The data used for the comparison of OR/HR scale in the "OEV" data were slightly different; we used the number of events and non-events for the OR and HR clog-log calculation (as in "binary" data) and calculated a HR based on "O-E" and V statistics. Therefore, there is a possibility for some cases that the two data sets entered by Cochrane reviewers may not completely correspond to each other.

We did not assess other reasons for differences between the results due to lack of information on censoring and follow-up times. Interpretation of the results was conducted with caution as we are interpreting the results based on known factors, without excluding other unknown factors that may have affected the results. We were not able to examine whether current practice of analysing time-to-event data has changed and whether methodological choices have improved since 2008. Further work examining the differences observed between analyses on the OR and HR scales in the presence of IPD is necessary.

The model used to analyse time-to-event data as binary is the conventional approach widely used by many systematic reviewers and meta-analysts [19]. It is quick, inexpensive and study results are obtained from appropriately synthesized study publications or by contacting study authors [20]. This approach to analysis ignores censored observations [21] and treats them as missing and has also been criticised for the within-study normality assumptions required [20].

The use of a clog-log link function, facilitating the results' interpretation in a HR scale for both "binary" and "OEV" data, was the best alternative approach enabling us to make comparisons between the scales used if only binary summaries are available. In the past, the clog-log link has been proven to provide a close approximation to Cox regression invoking a proportional hazards assumption, rather than a proportional odds assumption [6]. However, due to lack on information on "O-E" and "V" statistics for "binary" data only, we were not able to assess whether the HR obtained from the clog-log link is a close approximation to the true HR; therefore this magnifies the importance of extracting appropriate information when conducting time-to-event meta-analysis. For the "OEV" data, "O-E and V" data provide the best method to analyse aggregate data and facilitate results' interpretation on the HR scale but in the absence of IPD important biases may occur when large treatment effects and unbalanced data are present [22]. Additionally, we were not able to identify a clear pattern under which the complementary log–log link could be employed since there were circumstances under which it performed better or worse than an OR analysis; therefore we were not able to identify whether the clog-log approach is useful when a MA includes binary summaries alongside OEV or HR summaries. IPD and simulation studies are required to assess in more detail the conditions determining where this method would be acceptable.

For the "binary" data, we also used a one-stage random-effects model with fixed study-specific effects describing the baseline risk probability of the event in each study. These models use exact binomial likelihoods and may therefore be more accurate, especially with sparse data [14]. The fixed study-specific effects cause difficulties in estimation since the number of parameters increases with the number of studies, but maximum likelihood theory requires the number of parameters to remain stable as the sample size increases. A random-effects model with random study-specific effects could be applied, however based on simulation studies this model performed better than others without any serious biases present [14]. We were not able to make comparisons using one-stage models in the "OEV" data. We would be able to apply one-stage models when the data were analysed as binary, but we did not have the IPD required to fit one-stage models on the HR scale.

To our knowledge, no research has been conducted using such a large database assessing the differences between a) analysing the data as binary and interpreting the results in an OR scale and b) analysing the data either using the clog-log link or log-rank "O-E" and V statistics facilitating interpretation on the HR scale.

We have demonstrated the impact of reanalysing meta-analyses ("binary" or "OEV" datasets) within the Cochrane Database on a different scale, identifying the main drivers influencing discrepancies between the meta-analytic results. Our findings provide useful insights into changes to meta-analytical results and indicate that choice of method used in meta-analysis of survival data does matter, especially in the presence of high event probabilities.

Conclusions

In conclusion, our findings indicate that time-to-event data should be ideally analysed accounting for their natural properties, as it is possible for important discrepancies to be observed and conclusions from the meta-analysis to be altered. We identified that dichotomising time-to-event outcomes may be adequate for low event probabilities but not for high event probabilities. In meta-analyses where only binary data are available, the complementary log–log link may be a useful alternative when analysing time-to-event outcomes as binary, however the exact conditions need further exploration. These findings provide guidance on the appropriate methodology that should be used when conducting such meta-analyses.

Abbreviations

CDSR: Cochrane Database of Systematic Reviews; HR(s): Hazard Ratio(s); IQR: Interquartile Range; IPD: Individual Participant Data; MA(s): Meta-analysis(es); OEV: Observed minus Expected and Variance statistics; OR(s): Odds Ratio(s); REML: Restricted Maximum Likelihood; RR(s): Risk Ratio(s) or Relative Risk(s); SR(s): Systematic Review(s).

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12874-022-01541-9.

Additional file 1: Section 1. Fitting one-stage random-effects models for "binary" data. **Section 2.** Number (%) of (non-)significant meta-analyses under different scales for one-stage models ("binary" data). **Section 3.** Bland-Altman plots comparing standardised pooled effect and I^2 estimates for one-stage models ("binary" data). **Section 4.** Forest plots for example MAs considered as outliers in our analyses ("binary" data). **Section 5.** Bland-Altman Plot comparing standardised OR vs. HR estimates for two-stage models in "OEV" data. **Section 6.** Forest plot for example MAs considered as outliers in our analyses ("OEV" data). Section 7: R Code

Acknowledgements

We are grateful to the Nordic Cochrane Centre and the Cochrane Collaboration Steering Group for providing us with access to the Cochrane Database of Systematic Reviews. We would like to thank James Carpenter for the valuable suggestions and discussions we had during the preparation of this project. We would also like to thank Larysa Rydzewska for providing us with results obtained from the MRC Clinical Trials Unit's Survey of Collaborative Review Groups.

Authors' contributions

RMT proposed the study. TS performed the statistical analyses and drafted the manuscript. TS, RMT, DF, JFT and IRW jointly contributed to interpreting the results and to revising the manuscript. All authors approved the final manuscript.

Funding

TS received a Doctoral Training Grant from the UK Medical Research Council. RMT, DF, JFT and IRW were supported by the Medical Research Council Programme MC_UU_00004/06. The funders had no direct role in the writing of the manuscript or decision to submit it for publication.

Availability of data and materials

Data are available upon reasonable request, if permission is obtained from Cochrane.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

All authors have approved the manuscript for publication.

Competing interests

The authors declare that they have no competing interests.

Received: 11 June 2021 Accepted: 27 January 2022
Published online: 20 March 2022

References

1. Tierney JF, Stewart LA, Ghersi D, Burdett S, Sydes MR. Practical methods for incorporating summary time-to-event data into meta-analysis. Trials [Electronic Resource]. 2007;8:16.
2. Higgins JP, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, et al. Cochrane handbook for systematic reviews of interventions: John Wiley & Sons; 2019.
3. Tierney J, Rydzewska L. Improving the quality of the analysis of time-to-event outcomes in Cochrane reviews. [Unpublished]. In press 2008.
4. Green MS, Symons MJ. A comparison of the logistic risk function and the proportional hazards model in prospective epidemiologic studies. J Chronic Dis. 1983;36(10):715–23.
5. Ingram DD, Kleinman JC. Empirical comparisons of proportional hazards and logistic regression models. Stat Med. 1989;8(5):525–38.
6. Singer JD, Willett JB, Willett JB. Applied longitudinal data analysis: Modeling change and event occurrence: Oxford university press; 2003.
7. Hedeker D, Siddiqui O, Hu FB. Random-effects regression analysis of correlated grouped-time survival data. Stat Methods Med Res. 2000;9(2):161–79.
8. Peduzzi P, Holford T, Detre K, Chan Y-K. Comparison of the logistic and Cox regression models when outcome is determined in all patients after a fixed period of time. J Chronic Dis. 1987;40(8):761–7.
9. Annesi I, Moreau T, Lellouch J. Efficiency of the logistic regression and Cox proportional hazards models in longitudinal studies. Stat Med. 1989;8(12):1515–21.
10. Stare J, Maucort-Boulch D. Odds ratio, hazard ratio and relative risk. Metodoloski zvezki. 2016;13(1):59.
11. Callas PW, Pastides H, Hosmer DW. Empirical comparisons of proportional hazards, poisson, and logistic regression modeling of occupational cohort data. Am J Ind Med. 1998;33(1):33–47.
12. Davey J, Turner RM, Clarke MJ, Higgins JPT. Characteristics of meta-analyses and their component studies in the Cochrane Database of Systematic Reviews: a cross-sectional, descriptive analysis. BMC Med Res Methodol. 2011;11(1):160.
13. Whitehead A, Whitehead J. A general parametric approach to the meta-analysis of randomized clinical trials. Stat Med. 1991;10(11):1665–77.
14. 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):1059–85.
15. Sweeting JM, Sutton AC, Lambert P. What to add to nothing? Use and avoidance of continuity corrections in meta-analysis of sparse data. Statistics in medicine. 2004;23(9):1351–75.
16. Veroniki AA, Jackson D, Viechtbauer W, Bender R, Bowden J, Knapp G, et al. Methods to estimate the between-study variance and its uncertainty in meta-analysis. Research synthesis methods. 2016;7(1):55–79.
17. Bland JM, Altman DG. The logrank test. BMJ. 2004;328(7447):1073.
18. Bland JM, Altman D. Statistical methods for assessing agreement between two methods of clinical measurement. The lancet. 1986;327(8476):307–10.
19. Simmonds MC, Higgins JP. A general framework for the use of logistic regression models in meta-analysis. Stat Methods Med Res. 2016;25(6):2858–77.

20. 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):855–75.
21. Holzhauer B. Meta-analysis of aggregate data on medical events. Stat Med. 2017;36(5):723–37.
22. Greenland S, Salvan A. Bias in the one-step method for pooling study results. Stat Med. 1990;9(3):247–52.

Publisher's Note