The analysis and reporting of time to event data in randomised controlled trials: impact on evidence synthesis

Thesis submitted in accordance with the requirements of the University of Liverpool for the degree of Doctor of Philosophy by

Ashma Krishan

April 2021

Abstract

Thesis Title: The analysis and reporting of time to event data in randomised controlled trials: impact on evidence synthesis

Author: Ashma Krishan

Introduction and aims

The most commonly used approaches for the analysis of time-to-event (TTE) outcomes impose an assumption of proportional hazards (PH), such that the hazard ratio (HR) is assumed to be constant over time. Meta-analysis of TTE data is most commonly based on extracting or estimating the HR from individual trials, and so again assumes PH. Methods are available for assessing the validity of the PH assumption, however, the assumption is not always checked or reported for validity. This is a problem for meta-analysis, where different assumptions may have been made in the analysis of each included study. The aim of this thesis is to investigate how often the PH assumption is assessed within Randomised Controlled Trials (RCTs) and meta-analysis, including understanding the impact of non-PH on meta-analyses. This is of particular importance as current research has focused on alternative methodology, without knowing what impact non-PH may have on results.

Methods

The thesis summarises the results from a novel systematic review (SR) of the reporting of meta-analysis of TTE outcomes that have assumed PH, and how often the results of the PH assumption were reported. Two further SRs of PH assumption reporting within RCTs and Single Technology Assessments were also performed. A survey was also conducted targeted at the UKCRC network of registered clinical trials units, to understand what is done routinely rather than what is reported within RCTs. A simulation study was undertaken to assess the suitability of different modelling approaches for meta-analysis of TTE data in situations where PH is valid and invalid.

Results

All of the reviews on reporting of the PH assumption within SRs and RCTs, highlighted the poor reporting of the PH assumption. Only 33 out of 123 (27%) SRs and only 12 out of 106 (11%) RCTs reported the PH assumption. For the simulation study, meta-analytic datasets were simulated for twelve scenarios. Across scenarios, parameters controlling the Weibull distribution, the censoring level (25% and 75%), the time-dependent log HR (None, 0.1 and 0.5) and whether the treatment effect across studies is homogeneous or heterogenous were varied. The simulated datasets were analysed using Cox, Weibull, Accelerated Failure Time, and Flexible Parametric models. In situations where PH is valid, all models performed well as expected. However, as soon as the time-dependent log HR of 0.1 was introduced, the Cox and Weibull model could not cope. The best performing model in all cases was the Flexible Parametric Model.

Conclusions

The work of this thesis has provided a detailed insight into the poor reporting of the methods used to assess the PH assumption as well as the lack of reporting of the results of assumption checking. The work of this thesis highlighted the lack of reporting guidelines as there is no mention of the PH assumption in the CONSORT or PRISMA guidelines, Cochrane handbook or the ICH E9 guidelines. Recommendations that can be used by trialists, reviewers and ‘consumers’ of reviews on how to approach the PH assumption have been provided in the thesis.

Acknowledgements

Firstly, I would like to thank my supervisors Professor Catrin Tudur Smith, Professor Nicky Welton, Dr Maria Sudell and Dr Kerry Dwan for their excellent guidance and continuous support throughout my PhD.

Thank you to the funders of the research, the Medical Research Council as part of the North West Hub for Trial Methodological Research.

My sincere thanks must also go to my many colleagues across the research groups I have worked with; the Liverpool Reviews and Implementation Group and the Cochrane Cystic Fibrosis Disorders and Genetic Disorders Group. The completion of this PhD would not have been possible without the support, both academic and social of my fellow PhD students who made the last six years enjoyable and some of whom became my maternity buddies too!!

I would also like to thank my new employers, the Centre for Biostatistics at the University of Manchester and in particular Dr Chris Sutton as completion of this PhD would not be possible without his support.

I am eternally grateful to my family for always encouraging and supporting me and being my biggest cheerleaders in life! Particular thanks to my parents for pushing me to be the best I can be – I am honoured to submit this work in my maiden name. A special mention to my amazing nephew Mayek who proof read my thesis and saved his uncle from doing it, just before starting revision for his AS levels! When I started my PhD I never thought I would be going through two pregnancies over my PhD journey but I couldn't be more thankful for my two babies: Ava, thank you for bringing the greatest joy to my life and for reminding me there was more to life than the PhD! I especially looked forward to our Fridays together which were just about "us". Aarna, you were the sunshine that found its way to me at a time when the

world was a scary place…you gave me the push I needed to get my thesis written. Just would have been a slightly more happier person doing it if you slept a bit more!!

Lastly, and most importantly, thank you to my husband Amit, for picking up the slack financially and in terms of childcare duties over the last few years. Thank you for being by my side along every twist and turn of this journey. You could have done the washing though, just saying!

Table of Contents

List of tables

List of figures

Abbreviations

AC	Appraisal Committee
ACD	Appraisal Consultation Document
AFT	Accelerated Failure Time
AHR	Average Hazard Ratio
AIC	Akaike Information Criterion
BIC	Bayesian Information Criterion
CI	Confidence Interval
CONSORT	Consolidated standards of reporting trials
CTU	Clinical Trial Unit
DGM	Data Generating Mechanism
DSU	Decision Support Unit
ERG	Evidence Review Group
FAD	Final Appraisal Determination
FDA	Food and Drug Administration
HR	Hazard Ratio
HTA	Health Technology Assessment
IPD	Individual Patient Data
K-M	Kaplan-Meier
LRIG	Liverpool Reviews and Implementation Group
MCSE	Monte Carlo Standard Error

MRC CTU	Medical Research Council Clinical Trial Unit
MSE	Mean Squared Error
MTA	Multiple Technology Appraisal
NHS	National Health Service
NICE	National Institute for Health and Care Excellence
NMA	Network Meta-Analysis
OR	Odds Ratio
OS	Overall Survival
PFS	Progression Free Survival
PH	Proportional Hazards
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RCT	Randomised Controlled Trial
RMST	Restricted Mean Survival Time
RMSTD	Restricted Mean Survival Time Difference
RR	Risk Ratio
SD	Standard Deviation
SE	Standard Error
SR	Systematic Review
STA	Single Technology Appraisal
TA	Technology Appraisal
TR	Time Ratio

TSD	Technical Support Document
TTE	Time-To-Event
UK	United Kingdom
US	United States

Initials of researchers involved in the work in this thesis

AK	Ashma Krishan
CTS	Catrin Tudur Smith

1 Introduction

Randomised Controlled Trials (RCTs) are regarded as the gold standard of study designs to evaluate the effectiveness of a treatment in medical research in humans[1]. RCTs are prearranged experiments, that involve participants being randomly assigned to interventions. If RCTs are designed and conducted appropriately, then randomisation lowers the chance of any bias, by giving investigators the chance to control for factors, both known and unknown, that could otherwise influence the results[2].

Data from a RCT can be either qualitative or quantitative and equally within each of these categories there are different types of data e.g. examples of quantitative data include binary, continuous and time to event.

The main focus of this thesis is on time to event (TTE) data so the next section begins with an introduction to TTE in Section 1.1, followed by an overview of methods for analysing TTE including Cox proportional hazards (PH) model and underlying assumption of PH in Section 1.2. Section 1.3 introduces techniques of evidence synthesis including systematic reviews and meta-analysis. Section 1.4 explains what happens when the PH assumption is invalid as well as detailing previous work carried out on the importance of PH assumption checking. Finally, the objectives of the PhD and structure are explained in Section 1.5.

1.1 Time to event

Time to event data occur when interest is focussed on the time until a particular event occurs. Such data is often described as survival data as the event of interest is death, but other events are possible e.g. time until tumour recurrence.[3] The focus of this PhD is on the application of TTE analysis to data relating to any events occurring regardless of type of event; therefore, the discussion will be phrased in terms of TTE outcomes rather than only survival outcomes.

If the event of interest has not been observed for an individual, then the TTE is censored. Possible reasons for a TTE outcome being censored include data from a study being analysed at a particular time point, where the individuals have not experienced the event of interest. Alternatively, the individual could be lost to follow up, i.e. the event of interest status at the time of analysis might not be known and there may be no way to find out either, i.e. hospital or GP records etc. In this situation, the only information available on the TTE outcome is the last date on which the individual was known to be event free.

The survival and hazard functions are used in the statistical analysis of TTE data. The hazard function $h(t)$ is defined as the risk or hazard of an event occurring at time t, given that an individual has been event free until time t. $H(t)$ is the cumulative hazard function, defined as the sum of instantaneous hazards up to time t. The survival function $S(t)$ is the probability of being event-free up to time t. The functions can be directly linked as follows[3]:

$$S(t) = \Pr(T > t) = exp\{-H(t)\} = \exp\left\{-\int_0^t h(u)du\right\},$$

where T is the event time. The survival function is commonly expressed in plot form known as a Kaplan-Meier (K-M) plot. The K-M survival plot, which plots the K-M survival probability against time can provide a useful summary of the data that can be used to estimate measures such as median survival time. The large skew as seen in the distribution of most survival data is the reason why mean survival is not reported.[4]

A non-parametric approach that is the most popular method for comparing the survival of two or more intervention arms, which takes the full follow-up period into account is known as the log-rank test. The log-rank test is used to test the null hypothesis that there is no difference between the intervention groups in the probability of an event (such as death or relapse) occurring at any time point. Further details on the log-rank test are provided in Chapter 2.

The nature of the hazard function makes it more flexible for modelling than the survival function[3]. The hazard function can be estimated using a few different statistical modelling methods. All of the statistical modelling methods available follow different distributional assumptions including non-parametric procedures, semi-parametric models and parametric models. A detailed guide to modelling of TTE data in parametric, semi-parametric and non-parametric settings are provided in Chapter 2. In this chapter, the focus will be on the semi-parametric model, the Cox PH model and in particular on the PH assumption. Further details on the Cox PH model are given in Section 1.2.

1.2 Cox Proportional hazards model and assumption checking

1.2.1 Cox Proportional Hazards model

The Cox PH model[5], is the most commonly used method for the analysis of TTE outcomes in RCTs. The Cox PH model is a semi-parametric model as it does not require making any assumptions about the shape of the baseline hazard function. The Cox PH model describes the relationship between the hazard function and a set of of p covariates $(x_1, x_2, \dots, x_p)$, as follows for i individuals, with common baseline hazard $h_0(t)$, and β_1,…, β_p individual coefficients[5],

$$h_i(t) = h_0(t)exp\{\beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi}\}, \tag{1}$$

In the model as seen in equation (1), $h_0(t)$ is the baseline hazard that is likely to vary over time t, which equals the hazard function if all of the covariate terms included in the model take the value zero (i.e. exp(0) is equal to 1). The term $\exp(\beta_p)$ is referred to as a hazard ratio (HR). The Cox model is in essence a multiple linear regression of the logarithm of the hazard on the set of covariates x_i with the log hazard function being an ‘intercept’ term, which varies over time.

1.2.2 Cox PH Assumption

Although the Cox PH model makes no assumptions about the underlying statistical distribution of the event time, it does assume that the hazard rates for the groups are proportional over time and so the survival curves do not cross. However, if the survival curves cross and thus the PH assumption is not valid then the HR is not constant and is said to vary over time in which case the Cox PH model is not an appropriate method to use in such cases. Examples where the PH assumption is valid and an example where the curves cross and PH does not hold can be seen in Figure 1[6] and Figure 2[7] respectively.

Figure 1: K-M plot of adjuvant chemotherapy vs Control arm

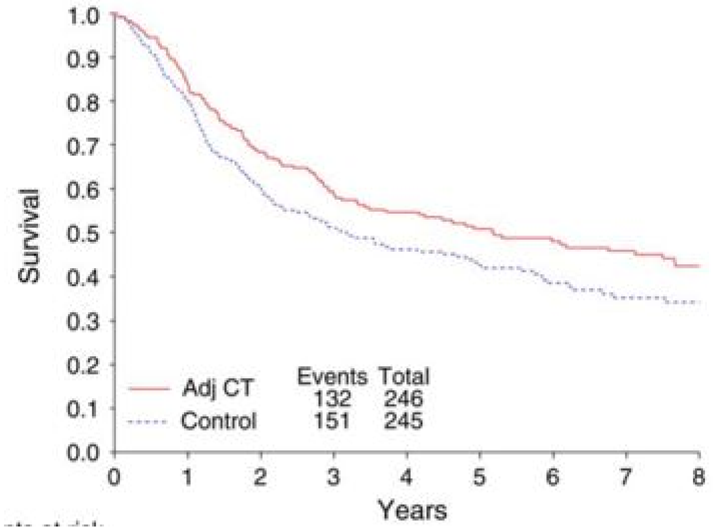
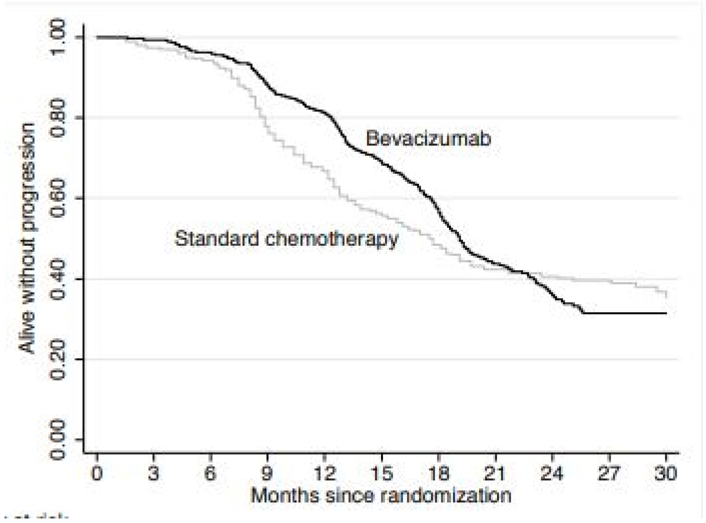


Figure 2: K-M plot of Bevacizumab vs Standard chemotherapy



A HR above one suggests that the covariate is positively related to the hazard rate but negatively related to the length of survival[8]. For example, a HR of 1.20 and where treatment arm equals one and the control arm equals zero, suggests that there may be a 20% greater risk of dying in the treatment group compared to the control group at any time during follow-up.

It is imperative that the assumption of PH is verified to ensure it holds. The PH assumption is violated if the coefficients of one or more of the covariates in the model vary with time or there are covariates that are time-dependent. Methods for assessing whether covariates are time-dependent will be further discussed in Chapter 2.

1.3 Evidence Synthesis

Evidence synthesis is a broad term used to define methods for combining sources of quantitative and/or qualitative evidence. The preparation of the research question of interest in a clinical setting for any evidence synthesis requires careful attention; a research question must be specific enough for results to be clinically useful but not too specific so that inadequate amounts of evidence are available[9]. A commonly applied analogy to this decision is the choice of whether to 'lump' or to 'split'[10]; in other words, whether to take a general approach to a wide variety of settings and participant groups or whether to constrict a research question into a homogenous evidence base[11].

In a clinical setting, where interventions and treatment effects are of importance, clinical assumptions underlying a synthesis must be considered as closely as statistical assumptions[12]. It is doubtful that a treatment effect would be replicated identically in two clinical studies given variations in participant populations and settings. However, if an intervention does provide true benefit over another then, the direction of effect can be expected to be the same in a range of heterogeneous situations[13]. This true direction of treatment effect is more likely to be noticeable in a synthesis when a number of studies are considered together.

The evidence synthesis methods of relevance to this thesis are systematic reviews, meta-analysis and network meta-analysis, which are introduced in the following sections.

1.3.1 Systematic reviews and meta-analysis

Systematic reviews are generally used as an approach to summarising the results of all independent sources of evidence, which focus on the same or similar questions in a systematic manner[14 15]. Systematic reviews of RCTs are widely known to provide the highest quality of results in evidence based medicine[16]. Nevertheless, the quality of the systematic review or any synthesis is dependent on the quality and completeness of the evidence[9].

Meta-analysis is a statistical approach used to combine the results of each study included in the systematic review, in order to obtain a single pooled result, which gives an overall relative treatment effect of one treatment compared to another[17]. Using this technique has many advantages namely, increases sample size and may increase power and precision, all whilst reducing the likelihood of a chance result. In this way, conducting a meta-analysis provides more information about the treatment effects which single studies do not have the power to detect[15 18].

Meta-analysis can be conducted using a fixed-effects or random-effects approach. The fixed effects model assumes that all the studies share a common effect, denoted θ. Hence, the differences in the observed effect sizes are all due to sampling error. The fixed effects model can be written:

$$Y_i = \theta + \varepsilon_i, \quad \varepsilon_i \sim \mathrm{N}(0, {\sigma_i}^2)$$

where Y_i are the observed effect sizes, θ is the common effect and ε_i are the random errors for studies i=1,…,k where k is the total number of studies. The random errors ε_i are assumed to be normally distributed with mean zero and variance ${\sigma_i}^2 > 0$, where the random errors are independent of each other[19].

The random-effects model assumes that a common effect across all studies is unlikely to be true as the differences between the observed effect sizes could be large and this may not be fully explained by the sampling errors. Causes of such heterogeneity in meta-analysis include differences in study design and methodology, participant characteristics and clinical settings[20 21]. In particular, for TTE data sources of heterogeneity include time-dependent treatment effects and differences in length of follow-up time across trials[22]. Further details on heterogeneity are given in Chapter 2, Section 2.2.4. The random effects model can be written:

$$Y_i = \mu + \delta_i + \varepsilon_i, \qquad \delta_i \sim N(0, \tau^2) \qquad \varepsilon_i \sim N(0, {\sigma_i}^2)$$

where Y_i are the observed effect sizes, μ is the mean effect, δ_i are deviations of the study-specific effect sizes from the mean effect and ε_i are the random errors. The deviations δ_i are assumed to be independent and identically distributed from a normal distribution with mean zero and variance τ^2 which is referred to as the between-study variance. The random errors, ε_i are the same as for the fixed effects model[19].

1.3.1.1 Aggregate data and Individual patient data

The most common method for conducting quantitative synthesis is using aggregate data; an approach where summary statistics such as mean differences, event counts, odds ratios, hazard ratios etc, are extracted from published literature and in some cases may be supplemented by unpublished information from the original trialists. The fixed effects approaches: inverse-variance weighting[15], Mantel-Haenszel[23] or Peto[18], or random effect approaches: Dersimonian and Laird[21] method could then be used to pool estimates from multiple studies. Further details on all of these methods are given in Chapter 2, Section 2.2.1.

An alternative approach which is regarded as the gold standard approach to the synthesis of study results uses individual patient data (IPD) retrieved from the original trialist which is then reanalysed.

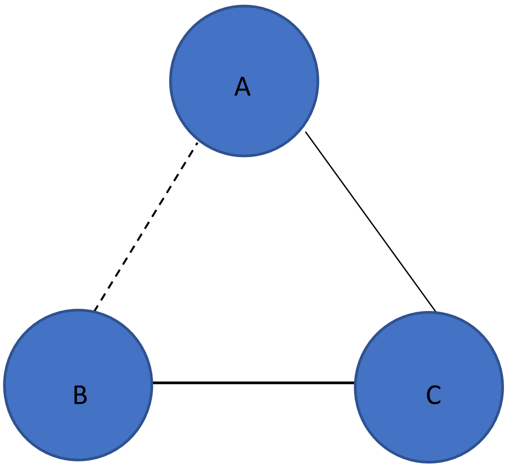
The availability of IPD provides many advantages including the opportunity to examine the data thoroughly, the chance to produce reliable analyses across studies, avoiding issues of within-study selective reporting and the chance to conduct further analyses such as treatment-covariate interactions[24 25]. Further details on methods available for performing an aggregate data or IPD meta-analysis are given in Chapter 2.

1.3.2 Network meta-analysis

Within traditional meta-analysis, usually two interventions (or classes of interventions) are compared head-to-head. However, within clinical settings where a large range of interventions are available, some interventions may never have been compared directly within a clinical trial. In these cases, traditional pair-wise meta-analysis cannot be used as it does not provide a suitable estimate of the relative effectiveness of all interventions of interest in order to support medical decision making[26].

Network meta-analysis, also known as mixed treatment comparison or multiple treatment meta-analysis, provides a framework for the synthesis of direct evidence and indirect evidence. A simple example is shown in Figure 3 where direct evidence for interventions A and B is unavailable but using indirect evidence for interventions A and B available from the direct comparison with a common intervention C, an indirect estimate for interventions A and B can be calculated.

Figure 3. Direct and Indirect evidence from the network of interventions A, B and C



The National Institute for Health and Care Excellence (NICE) is an independent organization responsible for providing national guidance to the NHS in England on a range of clinical and public health issues, including the appraisal of new health technologies. Network meta-analysis (NMA) is often conducted within Single Technology Appraisals (STAs) which is a process specifically designed by NICE for the appraisal of a single health technology for a single indication, where most of the relevant evidence lies with one manufacturer or sponsor and typically covers new technologies shortly after UK market authorization is granted[27]. Further details on NMA of TTE data and in particular STAs including TTE data and why the PH assumption is an issue is given in Chapter 6.

1.4 Proportional hazards assumption and previous work

1.4.1 Proportional hazards assumption

There are many methods available for assessing the PH assumption as will be explained in Chapter 2, however many of these methods are developed for use within RCTs where the IPD would be available. Meta-analysis of TTE data is commonly based on extracting or estimating the HR from individual trials, and so again assumes PH. Although there are methods for meta-analysis that do not impose this assumption[28], these methods are complex to apply without statistical expertise. IPD are needed to fully explore the assumption although this can also be done by reconstructing data from published survival curves provided curves are of adequate quality and other supplementary data have been published[29]. Further details on digitising K-M curves are given in Chapter 2. Due to complexity of methods and insufficient published data, meta-analysis and resulting clinical decisions, may be based on inappropriate methods. However, the impact of this is unclear.

Due to the level of uncertainty around the assessment of the PH assumption, a review of the reporting of the PH assumption within systematic reviews followed by a review of the reporting of the PH assumption within RCTs are required. These reviews will allow us to understand how often the PH assumption is being assessed and whether it is a cause for concern or not. It is also unclear what impact using inappropriate methods can have on meta-analysis and clinical decisions, hence why a simulation study is required to assess the impact of the violation of the PH assumption and investigate whether there is enough impact to alter overall conclusions. The next section summarises previous work on reporting of TTE outcomes and in particular on assessing the PH assumption.

1.4.2 Summary of previous work on TTE outcomes

The first known review to investigate the reporting of TTE outcomes within individual studies was carried out by Altman *et al*[30] in the 1990s. The review includes 132 publications with TTE outcomes published in five oncology journals. The aim was to review the reporting of study design, data handling, design and presentation in clinical oncology journals. Altman *et al*[30] report that 9 out of 132 (7%) individual studies did not state how many patients were included in the analysis, almost half of the publications (45%) did not provide a summary of length of follow-up and in 62% of publications at least one endpoint was not clearly defined. The majority of the results were reported as p-values and it was rare for survival probabilities or HRs to be presented. The log-rank test was the preferred method of univariate analysis with 75% of publications reporting it. Multivariate analyses such as the Cox PH model were conducted in 47 publications, with only 25 out of 47 (53%) publications reporting a TTE estimate of effect size such as HR or odds ratio (OR) and even less (34%) reporting a measure of precision such as a standard error or Confidence Interval (CI). The PH assumption was only investigated in two publications, with one assessing via a log cumulative hazard plot and the other by comparing the Cox regression estimates to those from a fully parametric model.

Survival curves were presented in 95% of the individual studies. The quality of the survival plots was considered poor in 43 out of 117 (37%) published curves, for reasons including censored observations not marked, poor or unhelpful numerical axis, inadequate or no legend given, survival curves of two or more groups not clearly distinguished and inconsistency between curves. Altman *et al*[30] conclude that in only 28 out of 132 (21%) individual studies, was the presentations of analyses and graphs considered adequate. The authors provided a set of guidelines in the appendix for the presentation of survival analyses.

Another systematic review by Abraira *et al*[31] published in 2013 compared TTE outcomes for single studies published in 1991 to those published in 2007 in 13 high-impact medical journals and showed a large rise in the number of published analyses from 104 (17%) publications in 1991 to 240 (33.5%) publications in 2007. The objective of this systematic review was to review how survival analyses within single studies are reported across medical journals and to evaluate changes in reporting over time and between journals. The review highlighted there had been little improvement in the quality of reporting of these analyses during this time period. Abraira *et al*[31] highlight the lack of publications reporting the number of events with 30 out of 104 (28.8%) publications in 1991 compared to only 60 out of 240 (25%) publications in 2007. The authors also emphasise the lack of reporting of key elements needed to interpret survival analysis results including description of censored cases, sample size estimation and assessment of the PH assumption. The results demonstrate that the PH assumption was only assessed in 5 out of 104 (10.6%) publications in 1991 whilst in 2007 the assumption was assessed in 47 out of 240 (26.3%) publications. Abraira *et al*[31] conclude that a high proportion of publications are deficient in their reporting of survival analysis methods and results and there was little improvement over the 16 year time period. Similar to Altman *et al*[30], the authors also present a list in the appendix of minimum requirements for the reporting of survival analyses.

Although the Altman *et al*[30] review was conducted and published prior to the introduction of the CONSORT statement for improving the quality of reporting of RCTs (first published in 1996[32], then revised in 2001[33], before being updated in 2010[34]), most of the recent reviews published post 2008[31 35], illustrate that reporting levels are similar to those reported over 10 years earlier. There does not seem to be considerable improvement in the assessment of the PH assumption, with the number of publications assessing the assumption still at less than 10%. Although the CONSORT statement doesn't specifically mention the PH assumption, a set of "minimum requirements" for reporting survival analyses have been published in

Altman *et al* in 1995[30] which include "When Cox regression analyses are performed, describe the criteria used to select the variables in the initial model, the procedure to specify the final model and describe any methods used to assess the model assumptions." In Abraira *et al* in 2013[31] the set of "minimum requirements" were slightly updated which state "When using regression models, report the method used and results of model assumptions checking (e.g., the proportional hazards assumption in Cox models or distributional form in parametric models)." However, in spite of some "minimum requirements" being published further work is still required to ensure authors are not only assessing the PH assumption and mentioning it in the publication, but are also explaining how they assessed the assumption, what the results were, and what appropriate action was taken dependent on the results.

In a more recent review, published in 2016 by Batson *et al*[35], 32 RCTs with TTE outcomes published in five oncology journals between April and July 2015 were included. The objective of this review was to review the methods and reporting of survival analyses and to access the suitability and relevance of survival data reported in RCTs for the inclusion into meta-analysis.

Batson *et al*[35] showed that the number of events were reported in 23 out of 32 (72%) publications, all of the publications reported the analysis of overall survival (OS) and 90% of publications reported a measure of follow-up time. The Cox PH model was reported in 28 out of 32 (88%) publications as either a univariate or multivariable analysis. No details were reported in the studies on the strategy for model building, the goodness of fit of the final model or final model validation. The results from the PH model were presented as HR and 95% CI. Assessment of the PH assumption was only reported in two out of 28 (7%) publications where the Cox PH model was used for analysis. Graphical methods were used for assessing proportionality including the log cumulative hazards plot and the plot of Schoenfeld residuals. None of these publications presented the plots but reported that the PH assumption was considered reasonable. Batson *et al*[35] report that an additional study stated that "Because the Cox proportional hazards model is the most commonly used approach to analyse time to event endpoints and because the two curves do not cross in this negative study, no tests for proportionality were done." However, the review authors state that after assessing the K-M curve it was clear the survival curves did cross. Batson et al[35] conducted a review of the survival curves published in the 28 individual studies using the Cox model and found that the survival curves crossed in 20 of the studies, implying that the PH assumption was invalid.

Survival curves were presented in all 32 publications. The review states that poor resolution and the use of relatively thick lines were the main limitations of the K-M curves as it made it difficult to separate points where treatments had very similar survival probabilities. Batson *et al*[35] state that further work is needed for assessing the impact of the PH assumption violation on the interpretation of study results and any subsequent meta-analyses and NMA.

Rulli et al[36] conducted a systematic review in 2018 where they included phase II and phase III RCTs published between January 2004 and January 2015. 115 studies were included that met the eligibility criteria. The authors report that only four (3%) trials assessed the PH assumption. The methods used to assess the PH assumption include scaled Schoenfeld residuals, adding time-dependent variables in the model, plotting the log-log survival functions and informally assessing the K-M curves. Rulli et al[36] conducted their own testing of the PH assumption and found that in 12 trials the PH assumption was violated. The authors explain why they have also found that there is a relationship between the type of treatments being compared and non-PH being present. New oncological treatments are regularly being compared to conventional treatments with a different method of action. This is often reflected in the disease progression and could explain why when treatments with different mechanisms of action are compared, the hazards are not proportional. Since treatments with different methods of action are being increasingly examined, further attention needs to be given to the statistical methods being used in such circumstances.

In 2014, Royston and Parmar[37] also provided reasons why they felt non-PH was being detected more frequently now compared to the past. Their reasons included that Phase III trials were larger now, so have more power to detect non-PH if it is present. Also, with a biological revolution, many more new therapies or treatments are being investigated. These are often given to a patient for a certain period of time and then stopped in which case the effect of the intervention may be seen during the treatment period but may diminish gradually afterwards. Another reason why non-PH is being detected could be that patients are selected to participate in trials so event rates are low initially for both arms as the patients need to be fit enough and need to meet the inclusion/exclusion criteria in order to be recruited. Then after the initial period differences between treatments are seen but these differences will tail off after some time as patients progress and move onto the next line of treatment. The reasons explained in this section for non-PH being increasingly identified highlight why the planned simulation study is so important.

To my knowledge, there have not been any reviews conducted on the reporting of the PH assumption within meta-analyses. There have been many methodology reviews explaining how summary TTE data can be incorporated into meta-analysis[14 22 38 39], with only the Williamson et al[22] review explaining how to assess the PH assumption. The Williamson et al[22] review was published in 2002 and focuses on how to extract the log(HR) and variance from aggregate data before showing how to assess the PH assumption using aggregate data. The review authors use a few methods including using overall log(HR) estimate for each study, the log cumulative hazard plot and the log(HR) for different time intervals. Williamson et al[22] do suggest that if the PH assumption is valid for one particular study then it is often reasonable to expect that the assumption will be valid for all studies, if the studies are similar enough with regard to other design and methodological features.

Similarly, there have been many reviews of IPD meta-analysis of TTE data[40-43], with Simmonds et al[43] and de Jong et al[42] mentioning the PH assumption within the reviews. Simmonds et al[43] published in 2011 conducted a simulation study where they compared three methods for estimating a treatment effect in a randomised trial collecting TTE data: a hypergeometric proportional odds model, a Cox PH model and an interval-censored logistic model which could be either proportional hazards or proportional odds depending on the link function chosen. The event times were simulated from a Weibull distribution. The main findings from the simulation study were that PH methods "will be biased when the hazards are not proportional…A greater awareness of the proportionality assumptions of the analysis methods is needed in meta-analyses, and investigation and testing of proportional hazards or odds assumptions should be a standard part of meta-analyses if interpretation of the findings is to be appropriate."

De Jong et al[42] published in 2020 conducted a literature review of IPD meta-analysis of TTE data which included 128 reviews. The literature review resulted in 10 key recommendations that the review authors have summarised including "consider non-PH models", "account for clustering in one-stage models, preferably by stratification of the baseline" and "apply one-stage models if trials are very small or the outcome very rare". The review authors also described what could be done if the PH assumption was found to be invalid. They explained that if non-proportional hazards were present then an interaction effect could be included between the intervention effect (or a covariate) and time in the one-stage approach or first part of the two-stage approach. Additionally, fractional polynomials or splines could also be applied but these approaches could complicate the interpretation of the regression parameters. The authors explained that another approach could be to use methods not reliant on the PH assumption such as restricted mean survival time (RMST)[44]. The percentile ratio[45] was another approach suggested by the review authors which is defined as the expected ratio for the time at which a certain amount of participants will have an event in the intervention group compared to the control group. Two-stage meta-analysis methods have been developed for the percentile ratio as suggested by Barrett et al[46].

However, the literature does highlight that there have been no reviews conducted on the reporting of the PH assumption within meta-analyses, which is why it is so important to understand what happens within meta-analyses.

1.5 Thesis objective and structure

The Cox PH model is the most commonly used method for analysing TTE outcomes, but is based on the validity of the assumption of PH. Although, methods are available for assessing the validity of the PH assumption, the summary of previous work in Section 1.4.2 highlights that assessment of the validity of the assumption is seldom reported. This is a bigger issue when meta-analysis is performed, where different assumptions may have been made in the analysis of each included study.

The aim of this thesis is to investigate how often the PH assumption is assessed within RCTs and meta-analysis, including understanding the impact of non-PH on meta-analyses. This is of particular importance as current research has focused on alternative methodology, without knowing what impact non-PH may have on overall results.

Chapter 2 provides a more detailed summary of methods for assessing the PH assumption, introduces alternative methods that do not depend on the PH assumption, and methods for analysing TTE outcomes within meta-analyses.

In Chapter 3, I conduct a systematic review of the reporting of the PH assumption within RCTs, and a survey of registered Clinical Trial Units (CTUs) in the UK to identify current practice for the analysis of TTE outcomes in clinical trials, focussing in particular on the assessment of the PH assumption.

In Chapter 4, I present a second systematic review of the reporting of the PH assumption but this time focussed on practice within meta-analyses.

In Chapter 5, I present an assessment of the results from digitising the K-M plots included in Chapter 4, including three worked examples focusing on digitising K-M curves to investigate non-PH in individual trials and in meta-analysis.

In Chapter 6, I summarise a further systematic review of the reporting of the PH assumption but this time focussed on practice within Single Technology Appraisals with emphasis on both clinical and cost-effectiveness results.

In Chapter 7, I conduct a large simulation study, informed by work in preceding chapters, to understand the impact of the violation of the PH assumption. The simulation study assesses the suitability of different modelling approaches for meta-analysis of TTE data in situations where PH is valid and invalid.

The final chapter summarises the findings of the previous chapters, reflects upon the implications for both clinical practice and research, and provides discussion of further research needed.

2 Literature Review

In this chapter a detailed overview of methods for modelling time to event data is given, along with details on methods for assessing the PH assumption and a description of alternative methods that do not require the PH assumption. Further details are also given on how to analyse aggregate data meta-analysis, individual patient data meta-analysis and what to do when there is a mix of both types of data.

2.1 Modelling time to event data in individual studies

2.1.1 Non-parametric models

Non-parametric methods do not require any specific assumptions to be made about the underlying distribution of the survival times or the shape of the survival curve. Two non-parametric procedures for comparing two or more groups of survival times, are the Log-rank test and the Wilcoxon test. For the simple case in a study comparing two intervention groups, the observed number of events in each intervention group along with the expected number of events are calculated under the null hypothesis of no difference between the groups. The log-rank test and the K-M survival curve are based on the same assumptions[47], which include that censoring is not related to prognosis, the survival probabilities are equal for individuals recruited early and late within a study, and the events happened at the times mentioned. The log-rank test is used to test the null hypothesis that there is no difference between the populations in the probability of an event at any time point. The log-rank test is more likely to detect a difference between groups when the risk of an event is consistently greater for one group than another. It is unlikely to detect a difference when survival curves cross, as can happen when comparing a medical with a surgical intervention.[48]

Similarly, the Wilcoxon test, also known as the Breslow test, is also used to test the null hypothesis that there is equality in the survival functions of two intervention groups. By definition, the Wilcoxon test seems similar to the log-rank test so the reasons for choosing one test over the other are mentioned next. If the hazard rate at any given time for an individual in one intervention group is proportional to the hazard rate at that time for a similar individual in the other intervention group so in other words, the assumption of proportional hazards is valid and the survival curves in the K-M plot do not cross then the log-rank test should be used.[49] In all other cases, the Wilcoxon test should be used to test the hypothesis of no difference in the survival time of the two intervention groups.[3] Therefore, it is vital to check the assumption of proportional hazards when using the log-rank test to assess whether the assumption holds.

The non-parametric methods are useful when comparing two or more groups of survival times but limited when supplementary information on explanatory variables is required.

2.1.2 Semi-parametric models

The main semi-parametric approach that this section will focus on is the Cox PH model[5], which represents the most commonly used method for the analysis of TTE outcomes in RCTs.

2.1.3 Methods for testing PH assumption

2.1.3.1 Kaplan-Meier survival plot

The K-M plot displays the survival probabilities for one or more groups against time. The plot can be inspected visually to assess whether there is crossing of the plotted survival curves as seen in Figure 1 and Figure 2. This method provides a subjective view on the validity of the assumption. Additionally, the K-M plot could suggest that the curves do not cross or only slightly cross, allowing one to make a decision on the validity when the actual data could be suggesting the opposite. K-M curves could also cross due to limited sample size and events making it again hard to judge. Another factor that is of importance when assessing the assumption is the quality of the K-M curve as a poor-quality curve could lead to an inaccurate decision being made. Hence, even though it could be helpful to assess the K-M plot visually to understand the data, an additional method should also be used when making a decision on the proportionality of the hazards.

2.1.3.2 Log-cumulative hazard plot

The log cumulative plot is another graphical tool used for the assessment of the PH assumption. If the PH assumption is valid, then a plot of the logarithm of the cumulative hazard function in each covariate against the logarithm of time, should produce lines that are parallel. The plot is also known as the log(-log(survival)) plot as the cumulative hazard is equal to the negative logarithm of the survival proportion. To assess the assumption using this plot, the survival data are firstly grouped according to the levels of one or more covariates. If continuous variables are included in the analysis, then their values need to be categorised into groups before use.[3] Then the K-M estimate of the survivor function of the data in each covariate is obtained, before producing the log cumulative hazard plot. Again, this approach also requires a subjective assessment. Although, this method is informative, convergent and divergent lines could be due to either the PH assumption being invalid or an important covariate not being included. Therefore, the plot would need to be assessed carefully.

2.1.3.3 Schoenfeld residuals

The Schoenfeld residuals is another graphical assessment of the PH assumption, but a graphical summary that tests the covariates for time-dependence. The Schoenfeld residuals[50] do not take just one value of residual from each individual but instead take a set of values, so one set for each covariate included in the fitted Cox regression model. However, Grambsch and Therneau[51] have developed scaled Schoenfeld residuals which is one of the most powerful diagnostic tools for proportionality according to Ng'andu[52] as the scaled Schoenfeld residuals assess the association between residuals and time hence highlighting if a time-dependent covariate exists and also provide a test statistic for formally testing for proportionality. The expected value of the *i*th scaled Schoenfeld residuals $i = 1, 2, \dots, p$, for the *j*th covariate in the model, X_j, $j = 1, 2, \dots, p$, denoted r^*_{Sji}, is given by:

$$E(r^*_{Sji}) \approx \beta_j(t_i) - \widehat{\beta}_j \tag{2}$$

where $\beta_j(t)$ is the time-varying coefficient of X_j, $\beta_j(t_i)$ is the value of this coefficient at the survival time of the *i*th individual, t_i, and $\hat{\beta}_j$ is the estimated value of β_j in the fitted Cox model. This equation suggests that a plot of the values of $r^*_{Sji} + \hat{\beta}_j$, or even just the scaled Schoenfeld residuals, r^*_{Sji}, against the observed survival times should provide details on the form of the time-dependent coefficient of X_j, $\beta_j(t)$. In particular, a horizontal line will imply that the coefficient of X_j is constant and that the PH assumption is valid. An overall or global test of the PH assumption across all the *p* covariates included in the Cox regression model provides a test statistic known as the *Grambsch and Therneau test of proportional hazards*.

2.1.3.4 Linear correlation test

Another method for testing the association between residuals and time is the linear correlation test developed by Harrell[52 53]. This is a simple test of the PH assumption based on Schoenfeld's residuals of the model. This test is conducted using Fisher's *z*-transform of the Pearson correlation between the residuals and the rank order of failure time. The residuals do not depend on time and they do not contain the estimated baseline hazard function which simplifies their asymptotic distribution. The test statistic for testing the validity of the PH assumption, is calculated using the following formula:

$$Z = \rho\sqrt{(n_u - 2)/(1 - \rho^2)} \tag{3}$$

where ρ is the correlation between the residuals and failure time order and n_u is the total number of uncensored observations. This test statistic is likely to be positive if the HR for high values of the covariates increases over time, and negative if the HR decreases over time. This method does not require categorization of the time variable or the covariate.[52]

2.1.3.5 Lee and Pirie method

The Lee and Pirie[54] graphical method is more often known as the H-H plot where the cumulative hazards for the intervention groups are plotted at the same time points. The basis of this method is that events may occur more, less or equally rapidly with increasing time in one series than in another, such as patients may have greater survival rates in one intervention group than another. The H-H plot is another graphical form for assessing the PH assumption by assessing if the trend follows a linear pattern. The plot contains a line of intercept which starts at the origin and then the cumulative hazards are plotted, if the data follows the line of intercept then the hazards are proportional. However, if the hazards are below the line then the PH assumption is invalid as the cumulative hazard for the treatment on the y-axis is getting proportionately further from the treatment on the x-axis over time. If the data sits above the line, then again, the PH assumption is invalid as the hazard for the intervention becomes proportionately smaller over time as the hazard for the comparator decreases more quickly than the intervention hazard. The Lee and Pirie method is similar to other graphical methods for assessing the PH assumption as again it is a subjective assessment of the assumption. However, an extension to the Lee and Pirie method exists which conducts a formal test to test if the intercept for the linear trend is significantly different from zero. The test is based on the logic that cumulative hazards will always start at zero (H[t]=-In[1]), so the linear trend of the data will always start at the origin. However, if the intercept of the H-H plot is significantly different from zero, there is evidence to suggest that the trend is not linear, and that PH does not hold across the time period. It is also likely that the linear trend could go through the origin but then the hazards are above the line for the first half of the data and then below the line for the second half, in which case PH would be invalid. Therefore, it is necessary to assess the plots and conduct the formal test to fully understand the validity of the assumption.

2.1.3.6 Time-dependent covariates

Another way of assessing departures from PH is by introducing a time-dependent covariate to the Cox regression model. A PH model for the hazard function of the *ith* individual in a study is

$$h_i(t) = exp\{\beta_1 x_{1i}\} h_0(t),$$

where x_{1i} is the value of an indicator value X_1 that is zero for the control arm and unity for the new treatment. The relative hazard of the event at any time for a patient on the new treatment, relative to one on the control arm is then e^{β_1}, which is independent of the survival time. A time-dependent covariate X_2, where X_2 = $X_1 t$ is added to the model and the hazard of the event at time t for the *ith* individual becomes,

$$h_i(t) = exp\{\beta_1 x_{1i} + \beta_2 x_{2i}\} h_0(t), \qquad (4)$$

where $x_{2i} = x_{1i}t$ is the value of $X_1 t$ for the *ith* individual. The relative hazard at time t is then: $\exp(\beta_1 + \beta_2 t)$ since X_2= t under the new treatment, and zero otherwise. This hazard ratio depends on t, and the model in equation (4) is no longer a PH model. In particular, if β_2< 0, the relative hazard decreases with time which means that the hazard of the event occurring on the new treatment, relative to that on the control arm, decreases with time. If β_1< 0, the superiority of the new treatment becomes more apparent as time goes on. In cases where β_2> 0, the relative hazard of the event occurring on the new treatment increases with time, reflecting an increasing risk of the event occurring on the new treatment compared to the control arm.[3]

To test the null hypothesis that the hazard is constant so PH is invalid, the likelihood ratio test statistic can be conducted by comparing the change in the value of the $-2log\hat{L}$ statistic to percentage points of the chi-squared distribution to test the significance of the covariate. This is therefore a formal test of PH.[3]

The methods described above explain the many ways available for assessing the PH assumption. Once the validity of the assumption has been checked, a decision can be made on the assumption. If it is valid then nothing further needs doing, however if the assumption does not hold then alternative methods will need to be considered which do not depend on the PH assumption being valid.

All of the above approaches can be used in individual studies and within a meta-analysis setting provided the raw data is available. However, if the raw data isn't available and the study authors have been approached to obtain the data but had no success then there is one approach that can be used to obtain pseudo IPD data as described in Section 2.1.3.7.

2.1.3.7 Digitising K-M curves

This approach uses the published K-M plot to obtain pseudo IPD data as suggested by Guyot et al[29], by digitising the K-M curves using software such as DigitiseIt (http://www.digitizeit.de/). Using this software, reconstructed IPD is obtained, which can then be used to conduct secondary analyses of survival data, whether for efficacy analyses or cost-effectiveness analyses. Another way in which this method can be used is to assess the PH assumption by using the reconstructed IPD to conduct further testing using any of the previously described methods. The method of reconstructing the K-M data has been validated and it has been suggested that reproducibility and accuracy of reconstructed statistics such as survival probabilities and median survival times was excellent as mentioned by Guyot et al[29]. However, for obtaining the HRs, reasonable accuracy can only be obtained if numbers at risk are provided otherwise further assumptions will need to be made. This method has been applied to obtain survival statistics which have been tested for accuracy and reproducibility[29], as well as some initial work being done for assessing the PH assumption[55 56]. Further details on how this is carried out in practice are given in Chapter 5.

2.1.4 Parametric models

There are three distributions that are most commonly assumed for the analysis of survival data using parametric models: exponential, Weibull and Gompertz. All three distributions will be explained in detail with particular focus on the Weibull distribution as it is the most commonly chosen parametric model. Once the distributional model for survival times is specified the hazard function, survival function and probability density functions can be determined:

$$S(t) = exp\{-H(t)\} \tag{5}$$

and

$$f(t) = h(t)S(t) = -\frac{dS(t)}{dt}, \tag{6}$$

where $H(t) = \int_0^t h(u)du$ is the integrated hazard function and $f(t)$ is the probability density function of the survival time.[3]

2.1.4.1 Exponential distribution

The exponential distribution is the simplest model for the hazard function as it assumes the hazard is constant over time. The hazard of an event at any time after the time origin of the study is then the same, regardless of the time that has elapsed.[3] Hence, the hazard function is written as

$$h(t) = \lambda,$$

for $0 \leq t < \infty$. The parameter λ is a positive constant that can be calculated by fitting the model to observed survival data. The survival function can be written as:

$$S(t) = exp\left\{-\int_0^t \lambda \, du\right\},$$

$$= \ e^{-\lambda t},$$

and thus the corresponding probability density function of the survival times is:

$$f(t) = \lambda e^{-\lambda t},$$

for $0 \leq t < \infty$. The median of the exponential distribution, t(50) is such that $S\{t(50)\} = 0.5$ that is,

$$exp\{-\lambda t(50)\} = 0.5$$

so that

$$t(50) = \frac{1}{\lambda} log2$$

2.1.4.2 Gompertz distribution

The Gompertz model was introduced by Gompertz in 1825, as a model for human mortality[3]. The hazard function is written as:

$$h(t) = \lambda e^{\theta t},$$

for $0 \leq t < \infty$, and $\lambda > 0$. If $\theta = 0$, then the hazard function has a constant value, λ, and the survival times then follow an exponential distribution. The parameter θ determines the shape of the hazard function, with positive values resulting in a hazard function that increases with time. The survivor function can be written as:

$$S(t) = exp\left\{\frac{\lambda}{\theta}(1 - e^{\theta t}\right\},$$

with the corresponding probability density function given as:

$$f(t) = \lambda e^{\theta t} exp\left\{\frac{\lambda}{\theta}(1 - e^{\theta t}\right\},$$

The median survival time is written as:

$$t(50) = \frac{1}{\theta} log\left\{1 + \frac{\theta}{\lambda} log2\right\}$$

2.1.4.3 Weibull distribution

Unfortunately, the assumption of a constant hazard function is rarely plausible. Thus, a more broad form of the hazard function can be written as:

$$h(t) = \lambda\gamma t^{\gamma-1},$$

for $0 \leq t < \infty$, a function that depends on two parameters, λ the scale parameter and γ the shape parameter, both which are greater than 0. In the case where $\gamma = 1$, then the hazard function is equal to λ, and the survival times follow an exponential distribution. For all other values of γ, the hazard function increases or decreases monotonically, therefore does not change direction. The survivor function can be written as:

$$S(t) = exp\left\{-\int_0^t \lambda\gamma u^{\gamma-1} du\right\}$$

$$= \exp(-\lambda t^{\gamma})$$

with the corresponding probability density function given as:

$$f(t) = \lambda\gamma t^{\gamma-1}\exp(-\lambda t^{\gamma}),$$

for $0 \leq t < \infty$. The Weibull distribution is denoted as $W(\lambda, \gamma)$ with scale parameter λ and shape parameter γ. The right-hand tail of this distribution tends to be longer than the left-hand tail, hence why this distribution is known to be positively skewed.[3] The median survival time is known as:

$$t(50) = \left\{\frac{1}{\lambda} log2\right\}^{1/\gamma}$$

As the Weibull hazard function can take a range of different forms, based on the value of the shape parameter, γ and as suitable summary statistics can easily be acquired, this distribution is widely used in the parametric analysis of survival data.

2.1.5 Comparing parametric and semi-parametric approaches

As with any parametric approach, there is a need to identify the most appropriate distribution for the data which is not always straightforward. However, when a suitable distribution is found, the parametric model is more informative than the Cox model. It is simple to estimate the hazard function and to obtain predicted survival times in order to extrapolate the data and make projections into the future, which isn't possible within the Cox model. In addition, the parametric models yield more efficient and precise estimates (smaller standard errors). The results from the Cox or parametric models can be compared directly, as in essence the different model types are simply different approaches for estimating the same result.[8]

2.1.5.1 Assessing PH Assumption for parametric models

Before fitting a model based on an assumed parametric form for the hazard function, the validity of the PH assumption should be checked. If the hazard function is reasonably constant over time, this would suggest that the exponential distribution is the most appropriate model for the data. However, if the hazard function is increasing or decreasing monotonically with increasing survival time, then a Weibull distribution may be the preferred option.

Nevertheless, a more informative method for assessing whether a particular distribution for the survival times is reasonable is by comparing the survivor function for the data with the survivor function from the chosen model. An ideal approach for implementing this is by transforming the survivor function to produce a plot that gives a straight line if the assumed model is suitable.

An example of how to perform this is by taking a single sample of survival data, and using say a Weibull distribution for the survival times. As the Weibull distribution contains a scale parameter, λ and shape parameter γ, written as:

$$S(t) = exp\{-\lambda t^{\gamma}\}$$

Taking the logarithm of the survivor function, $S(t)$ multiplying by -1 and then taking logarithms again gives:

$$log\{-logS(t)\} = \log\lambda + \gamma\log t$$

This is the log-cumulative hazard plot which was mentioned earlier in this chapter. If the log-cumulative hazard plot gives a straight line, the plot can be used to deliver a rough estimate of the two parameters in the Weibull distribution. However, the intercept and slope of the straight line are now $\log\lambda$ and γ, thus the slope of the line provides an estimate of the shape parameter and the exponent of the intercept provides an estimate of the scale parameter. If the slope of the log-cumulative hazard plot is close to unity, it is possible to model the survival times using the much simpler exponential distribution.

Details have been given on how to use the log-cumulative hazards model to assess if the correct parametric model has been used to model the survival times, but the log-cumulative hazards models can also be used to assess the PH assumption. For the assumption to be valid the hazard rates needs to be constant at any time point for the two intervention groups. To explain how the assumption is checked, the Weibull distribution will be used to demonstrate so the two intervention groups are labelled as intervention A and intervention B. The indicator variable will be labelled X, which takes a value of zero if an individual is receiving intervention A and value of unity if an individual is receiving intervention B. Under the PH model, the hazard of an event occurring at time t for the *i*th individual is given by:

$$h_i(t) = e^{\beta x_i} h_0(t)$$

where x_i is the value of X for an *i*th individual. Therefore, the hazard at time t for an individual in intervention group A is $h_0(t)$, and for an individual in intervention group B is $\psi h_0(t)$, where $\psi = \exp(\beta)$. The value β is the logarithm of the ratio of the hazard for an individual in intervention group B, compared to that of an individual in intervention group A. Therefore in this situation, a single sample of survival times from intervention group A will follow a Weibull distribution $W(\lambda, \gamma)$, where the log-cumulative hazard plot will give a straight line with intercept $\log \lambda$ and slope γ. A single sample of survival times from intervention group B will also follow a Weibull distribution $W(\psi\lambda, \gamma)$, where the log-cumulative hazard plot will give a straight line with slope γ but with intercept $\log \psi + \log \lambda$. If the estimated log-cumulative hazard function is plotted against the logarithm of the survival times for individuals from the two intervention groups, then parallel straight lines would suggest that the PH assumption and Weibull survival times are reasonable. If however, the two lines in the plot are in principal straight, but not parallel, then this would suggest that the shape parameter, γ is different for the two groups, and the hazards are not proportional. Additionally, this would also suggest that the Weibull model may not be appropriate to use to model the survival times and an alternative distribution may be more suitable. However, if the curves seem reasonably parallel, then this would suggest that the PH assumption is valid and that the cox regression model may be more appropriate to use.[3]

2.1.6 Accelerated Failure Time model

The accelerated failure time (AFT) model allows a wider range of survival distributions including the Weibull distribution, log-logistic distribution, log-normal, gamma and inverse Gaussian distributions which allow increased flexibility. The parametric AFT model assumes the covariates measured for an individual act multiplicatively on the time-scale, so in other words say for the covariate treatment, the length of survival is either increasing or decreasing in the new treatment group compared to the standard treatment group. For a group of individuals with covariates $(x_1, x_2, \dots, x_p)$, the model is written as:

$$S(t) = S_0(\varphi t)$$

where $S_0(t)$ is the baseline survivor function and φ is the 'acceleration factor', which is an unknown positive constant. The AFT model is sometimes referred to as:

$$\log T = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p + \varepsilon$$

where ε is the measure of variability in the survival times. Here the survival times are multiplied by a constant effect, and the exponentiated coefficients, $\exp(b_i)$ are referred to as time ratios, where a time ratio of greater than one for the covariate suggests that the time to the event has been extended whereas a time ratio of less than one suggests that the event is likely to occur earlier.

Where the survival times are modelled using the Weibull distribution, it can be shown that the AFT and PH models are equal.[8] In this situation, the AFT model only differs from the PH model in terms of interpretation of the effect size as the AFT model is measured in terms of time ratios (TRs) compared to HRs for PH models. Similar to the parametric PH models, the AFT model can also be used for survival probability projections which is often used for cost-effectiveness analyses. Additionally, there are underlying assumptions of the AFT model that need to be considered. These include a suitable choice of probability distribution for survival times and the covariate effects are expected to be constant and multiplicative on the time-scale.[3 8]

2.1.7 The Royston and Parmar method

It is sometimes difficult to determine which probability distribution should be used to model the survival times. In this situation, the Royston and Parmar[57] method can be considered, which models the underlying baseline hazard parametrically but allows the model to have greater flexibility than is possible with fully parametric models. This method begins by fitting a Weibull model for the hazard of an event occurring at time t, where:

$$h_i(t) = \exp(\beta' x_i)\, h_0(t),$$

where $h_0(t) = \lambda\gamma t^{\gamma-1}$, λ and γ are the scale and shape parameters of the Weibull distribution, and x_i is the vector of values of *p* covariates for the *ith* individual. The cumulative hazard function can be written as:

$$H_i(t) = \int_0^t h_i(u)du = \exp(\beta' x_i)\lambda t^{\gamma},$$

and the log-cumulative hazard function becomes:

$$\log H_i\,(t) = \beta' x_i + \log\lambda + \ \gamma \log t$$

If we set $\eta_i = \beta' x_i$, and let $\gamma_0 = \log\lambda$, $\gamma_1 = \gamma$ and $y = \log t$, then the log-cumulative hazard function for the Weibull model becomes:

$$\log H_i(t) = \gamma_0 + \gamma_1 y + \ \eta_i$$

This rearrangement shows that the log-cumulative hazard function is linear in $y = \log t$.

The next step in fitting the Royston and Parmar model is to simplify the linear term in *y* to a natural cubic spline in *y*. To define this, the series of values of *y* are separated into a number of intervals, where the boundary between each interval is called a knot. A simple example of this would be for say 3 knots to take the smallest and largest y-values so k_{min} and k_{max} and then divide the range into two so the knot in the middle becomes k_1. Thus, there are two knots at the boundaries and one interval knot such that this cubic expression in y is defined as $y \epsilon (k_{min}, k_1)$ and $y \epsilon (k_1, k_{max})$. These two cubic expressions are then forced to have a smooth join at the interval knot k_1 to give a cubic spline. The flexibility of the parametric model for $\log H_i(t)$ can be increased by adding in more internal knots. However, the higher the number of knots, the more complex the curve becomes. The log-cumulative hazard function can be extended for a model with m interval knots, which means the survival times no longer follow a Weibull distribution, or any other known distribution, although the model still assumes PH between the covariates.

For fitting the model, the maximum likelihood method can be used. Firstly, the Akaike information criterion (AIC) statistic can be used to select the covariates to include in the model, before using the statistic to determine how many knots should be fitted to the model, where a model with zero knots is a standard Weibull model.[3]

Although the flexible parametric model gives appropriate estimates of treatment effects under PH and non-PH with the addition of time-dependent effects, they are complex to fit and interpret. An alternative effect estimate to the HR is described next.

2.1.8 HR Interpretation under non-PH

In the presence of non-PH the interpretation of HR is an average HR (AHR) over the observed follow-up. A recent paper by Schemper et al[58] clarified and explored many definitions of "average" and concluded in favour of a definition proposed by Kalbfleisch and Prentice[59] based on weighted Cox regression:

$$AHR = \frac{\int \left[{h_1(t)}/{h(t)} \right] w(t)f(t)dt}{\int \left[{h_0(t)}/{h(t)} \right] w(t)f(t)dt},$$

where $h_0(t) and\ h_1(t)$ are the hazard functions in the two treatment groups, h(t) = $h_1(t) + h_0(t)$ and w(t) is the weight function to be chosen by the user. F(t) is the density function but Royston and Parmar[60] note the uncertainty about what the distribution of the density function is. Royston and Parmar[60] also argue against the proposal by Schemper et al[58] that an overall estimate of the HR can be regarded as an average of time-dependent HRs over the event times. The issue with the average HR is that it is uninterpretable especially if the hazards cross so that the HR switches from being greater than one to less than one over the entire period.

Royston and Parmar[44] discuss recently reported trials that observed crossing survival curves, but inspite of which the HR was still reported. Although the PH assumption should be checked, a recent review[35] has highlighted the poor reporting of the PH assumption and this is further explored in Chapter 3.

Additionally, early stopping rules that assume PH can be making inappropriate decisions in particular in cases where the HR later changes considerably. Unfortunately, no single summary of HR or risk difference can appropriately describe situations in which the treatment effect changes direction as follow-up increases. Even in instances where PH is valid, the HR is not as clinically meaningful as some other types of effect estimates such as difference in average survival time or proportion say alive at a fixed time point, concealing the absolute difference between the treatments and failing to communicate the clinical value of a treatment.[44]

2.1.9 Definition of restricted mean survival time

Due to the difficulties with interpretation of HR, particularly under non-PH, the RMST, a measure of average survival from time 0 to a specified time point, was first introduced in 1949[61] and then included in some seminal work of Per Kragh Anderson and others[62-65]. The difference in RMST has been regarded as an alternative to the HR as a summary measure of the treatment effect.

The RMST, $\mu(t^*)$ say, of a random variable T, is the mean of the survival time X=min(T,t*) limited to some horizon t*>0. It equals the area under the survival curve S(t) up to t*:

$$\begin{aligned}\mu(t^*) &= E(X) \qquad (7)\\ &= E[\min(T, t^*)]\\ &= \int_0^{t^*} S(t)dt\end{aligned}$$

When T is time to death, we may interpret $\mu(t^*)$ as the 't*-year life expectancy.' For example, a patient might be told that 'your life expectancy with X treatment and Z disease over the next 18 months is 9 months', or 'treatment A increases your life expectancy during the next 18 months by 2 months, compared with treatment B'. The measure $\mu(t^*)$ increases monotonically with the t* as equation (7) gives a non-negative, increasing function of t*. The integral is not in general evaluable due to the almost universal right-censoring of the time to event.[44]

2.1.9.1 Methods for estimating RMST

In 2011, Royston and Parmar[60] described three approaches for estimating the difference in RMST, which will all be explained in detail below. These methods use pseudo-values, flexible parametric model and integration of the Kaplan-Meier estimate.

2.1.9.1.1 Pseudo-values

Using pseudo-values the RMST for individuals can be estimated by a non-parametric jack-knife method (leave-one-out approach)[63]. Suppose the focus is on the parameter, θ. The first step is to estimate θ based on the whole sample with observations for each individual i (i=1,2,…,n). Then estimate θ again using the leave-one-out approach so based on a subsample by omitting an observation, i say. The pseudo-value $\widehat{\theta_{\imath}}$ for observation i is calculated as the difference between the two estimates of θ, and is defined as:

$$\widehat{\theta_{\imath}} = n\hat{\theta} - (n-1)\widehat{\theta_{-\imath}} \tag{8}$$

where $\hat{\theta}$ is the estimate based on the full sample and $\widehat{\theta_{-\imath}}$ is the estimate based on the subset of the sample without observation i. The average of pseudo-values across all observations is given by

$$\hat{\theta}_{peudo} = \frac{1}{n}\sum_{i=1}^{n}\widehat{\theta_{\imath}}, \tag{9}$$

From this equation, there is

$$E(\hat{\theta}_{peudo}) = E(\hat{\theta}_i),$$

suggesting $E(\hat{\theta}_{peudo}) = \theta$, if $E(\hat{\theta}_i) = \theta$. Additionally, based on the definition of an individual's pseudo-value in (8), there is $E(\hat{\theta}_i) = \theta$ if $E(\hat{\theta}) = \theta$. Hence, the use of the unbiased estimator $\hat{\theta}$ is vital for $\hat{\theta}_{peudo}$ to be unbiased for θ.

Therefore, the pseudo-values for the RMST[60 63] are defined as

$$\hat{\mu}_i^* = \int_0^{t^*} \hat{S}_i(t)dt$$

$$= n\int_0^{t^*} \hat{S}\,(t)\,dt - (n-1)\int_0^{t^*} \hat{S}_{-i}\,(t)\,dt$$

where the survival function $\widehat{S(t)}$ can be replaced by a Kaplan-Meier estimate

$$\hat{S}(t) = \prod_{u\leq t}\left(1 - \frac{d_u}{n_u}\right) \qquad (10)$$

where d_u denotes the total number of failures from time origin to time *u* and n_u is the total number of individuals still at risk prior to time *u*. The pseudo-values estimator for the RMST is then defined as

$$\hat{\mu}_{pseudo}^* = \frac{1}{n}\sum_{i=1}^{n} \hat{\mu}_i^*$$

The pseudo-values and Kaplan-Meier estimates are both non-parametric. Therefore, when combined they provide a non-parametric estimate of the RMST.[66]

2.1.9.1.2 Flexible parametric model

A parametric method for calculating the RMST is using the flexible parametric survival model. As explained in Section 2.1.7, Royston and Parmar suggest approximating the log of the cumulative baseline hazard $H_0(t)$ using a function of the log of time

$$lnH_0(t) = \gamma_0 + \gamma_1 \ln t + \gamma_2 \nu_1(\ln t) + \cdots + \gamma_{K_0+1}\nu_{K_0}(\ln t) \qquad (11)$$

where $\gamma_i$are regression parameters with i=0,1,…K+1 and $\nu_i$is the *i*th spline with i=1,2,…,K_0). Here, K_0 represents the number of distinct internal knots which again have been defined in Section 2.1.7. The RMST in equation (7) can be rewritten as

$$\mu^* = \int_0^{t^*} S(t)\, dt = \int_0^{t^*} \exp\left(-H(t)\right)\, dt \qquad (12)$$

Then setting $lnH_0(t) = s(\ln t\,|\gamma, K_0)$, the log cumulative hazard function can be specified as

$$\ln H(t) = s(\ln t\,|\gamma, K_0) + s(\ln t\,|\delta, K_1)x + \beta x$$

where x is the treatment arm. The interaction term $s(\ln t\,|\delta, K_1)x$ is added into the model to account for the non-PH hazards. If the number of knots increase, then the complexity of the model also increases. It is known that the estimates of RMST are similar when the degrees of freedom for the baseline distribution is 3 or more (i.e. K_0=2) Hence, the default number of degrees of freedom in statistical software tends to be 3. Flexible parametric models can be calculated in Stata, R and SaS.

2.1.9.1.3 Integration of survival functions

Another method that is also often used for estimating RMST is directly integrating the Kaplan-Meier estimate of the survivor function from time 0 to t^*. To calculate the RMST, simply integrate the Kaplan-Meier estimate $\hat{S}_j(t) of\ S_j(t)$ on (0, t^*) for each treatment group separately, where $S_j(t) = S_0(t)^{\exp(\widehat{\beta_J})}$, $S_0(t)$is the baseline survival function and $\widehat{\beta_J}$ is the log HR for treatment j compared to the control group. This is equal to using survival estimates from a cox model stratified by treatment group.[66]

2.1.9.2 Calculating the difference in RMST and its variance

The difference between the RMSTs for two treatment groups of a trial is defined as

$$\Delta^* = \mu_1^* - \mu_0^*$$

This measure is known as the RMST difference (rmstD). The rmstD measures the quantity by which the treatment group changes the survival time on average up to time t^* compared to the control group. The interpretation of rmstD is quite simple, e.g. patients in the treatment group have Δ^* more years, say gained/lost in life expectancy from 0 to t^*compared to patients in the control group.

The variance of the RMST is defined as:

$$Var(X) = RSDST^2 = E(X^2) - [E(X)]^2$$

$$= 2\int_0^{t^*} tS(t)\, dt - \left[\int_0^{t^*} S(t)\, dt\right]^2$$

where RSDST is the restricted standard deviation of survival time. The RSDST is then $\sqrt{Var(X)}$.[44]

2.2 Meta-Analysis

If the objective is to conduct a meta-analysis of TTE data from multiple studies, an aggregate data approach would usually focus on extracting the HR, or suitable data to approximate it[14], and measure of precision from each individual study to input into the calculation of the pooled overall estimate. However, noting the issues outlined in previous sections, with poor reporting of TTE data in primary studies, and the lack of adequate investigation and reporting of the underlying assumptions, it is conceivable that meta-analyses based on aggregate data may be problematic. Therefore, carrying out IPD meta-analysis has many advantages including the opportunity to examine the data thoroughly, the chance to produce reliable analyses across studies, avoiding issues of within-study selective reporting[67] and the chance to conduct further analyses such as treatment-covariate interactions.

2.2.1 Aggregate data meta-analysis

As it is not always possible to obtain IPD due to restrictions gaining access to data or restrictions with resources, performing an aggregate data meta-analysis may be the only option, and will often be a pre-requisite to more detailed IPD approaches.

The basic data required for an aggregate data meta-analysis is the log HR and variance of the log HR from each study. If summary statistics are presented, the following three approaches can be used to obtain estimates of HR and a measure of uncertainty from the study reports for inclusion in a meta-analysis.

1. The simplest approach is that if the trialists have analysed the data using a Cox PH model then estimates of the log HR and its variance can be extracted. If the HR is reported alongside the confidence interval or p-value then an estimate of the variance can be obtained using these values as shown in the Cochrane Handbook[68].

2. Another approach is to estimate the HR approximately, using statistics computed during the log-rank analysis. The log HR is estimated by (O-E)/V and standard error is equal to $1/\sqrt{V}$, where O is the observed number of events in the treatment arm, E is the log-rank expected number of events in the treatment arm, O-E is the log-rank statistic and V is the variance of the log-rank statistic. This approach is described in more detail by Parmar et al[14].

3. The third approach is to estimate the log HR and its variance from survival curves. This approach was described in Parmar et al[14] in 1998 by manually estimating survival probabilities at appropriate time points by reading off the published curves. However a more efficient and accurate way of extracting data from the survival curves is to reconstruct approximate IPD from published curves using specific software as suggested by Guyot et al[29]. This approach allows a re-analysis of the data to estimate the HR and its variance.

The methodology of aggregate data meta-analysis is well developed with many different approaches being used including fixed effect approaches such as inverse-variance weighting[15], the Mantel-Haenszel method[23] or the Peto method[18], and random effect approaches such as Dersimonian and Laird[21].

The inverse-variance[15] method is where the weight given to each study is chosen to be the inverse of the variance of the effect estimate. The DerSimonian and Laird[21] method is a variation of the inverse-variance method which incorporates the assumption that the different studies are estimating different, yet related, intervention effects.

The Mantel-Haenszel[23] method is usually the default fixed-effect method of meta-analysis programmed within software packages such as Stata[69] and R[70]. This method is used when data is sparse, either in terms of event rates being low or study size being small or the estimates of the standard errors of the effect estimates that are used in the inverse variance methods being poor. The Mantel-Haenszel method uses a different weighting system that depends upon which effect measure is being used.

Another fixed-effects method is the Peto[18] method which uses the number of observed and expected events to calculate the pooled log HR. Here, O is the observed number of events and E is an expected number of events in the experimental intervention group of each study. The data required for the analysis is derived from the number of events and the individual times to event on the research arm of each trial.

2.2.2 IPD Meta-analysis

For the IPD meta-analyses, two different statistical approaches can be considered, a one-stage and a two-stage model. One-stage models simultaneously analyse IPD from all studies, while accounting for the separate studies. Such models, while more computationally complex, offer additional flexibility to incorporate covariates, interaction terms or heterogeneity parameters[41 71 72]. A series of one-stage Cox models for the meta-analysis of TTE data where IPD are available, has been described previously[72].

Alternatively, a two-stage approach would fit separate models to the data from each study individually as the first stage and then study level results (logHR and standard error (SE)) from these separate models pooled together using general meta-analytic models in the second stage.

2.2.3 IPD and Aggregate data

If IPD is available for some included studies but not all, and if suitable aggregate data is available in those studies without IPD then the different types of data can be pooled together, using a two-stage approach as described above.

2.2.4 Heterogeneity

Heterogeneity is defined as any variability between studies included in the meta-analysis[73], and can be split up into methodological or statistical heterogeneity (difference in study designs or differences in the study effect sizes) or clinical heterogeneity (difference in study population or difference in interventions). Heterogeneity is measured using various statistics including the I^2 statistics, τ^2 or Cochrane's Q [73]. It is crucial to account for heterogeneity within meta-analysis, as otherwise the results could be misleading. If there is no evidence of heterogeneity between the included studies, then this could suggest that the test doesn't have enough power to detect the heterogeneity but in such a case a fixed effects model could be used to pool the studies together. If there is some heterogeneity between studies, this needs to be factored into the analysis, such as by using a random-effects model[21]. If there is a large amount of heterogeneity present between the studies, it may be inappropriate to pool the results from different studies, hence where a narrative summary of results may be more appropriate.

3 Review of reporting of the Proportional Hazards assumption within Randomised Controlled Trials

3.1 Introduction

Most of the previous methodology reviews[30 31] conducted have focused on reporting of TTE outcomes in RCTs with very little mention of the PH assumption and hardly any reporting of the methods used for assessing the PH assumption. A recent methodological review by Batson *et al*[35] was the first known review to not only report on how often the PH assumption was assessed but also to report what methods were used to assess the PH assumption. The main limitations for this review were that only 32 publications were included and all were from an oncology setting.

The objectives of this chapter are to (i) assess the frequency and approaches used for exploring the PH assumption within individual RCTs, and (ii) assess which methods of analysis are used in current practice. This is achieved by conducting an in-depth review of a sample of published RCTs that have included the analysis of a TTE outcome using a method assuming PH, and undertaking a survey of current practice targeted at the UKCRC network of registered CTUs, who are regularly involved with conducting and analysing clinical trials with TTE outcomes across a wide range of disease areas.

3.2 Methods of the in-depth review

3.2.1 Identification of RCTs

To identify the RCTs to include in the review, a list of systematic reviews of RCTs that are known to have analysed TTE outcomes using methods assuming PH (to be described in full in Chapter 4) was examined. From the list of 123 included systematic reviews, a focused sample of 20 (16%) systematic reviews were selected to provide a pragmatic but representative sample of RCTs that included:

- A selection of reviews including different types of data so that a mix of aggregate data reviews, IPD reviews and reviews including aggregate data and IPD were included.
- A varied number of RCTs included in the systematic review ranging from 2 included studies up to 39 included studies.

It is worth noting that for the selection of studies, the studies chosen were sent to all supervisors for review. Screening of studies and data extraction were all completed by myself.

3.2.2 Eligibility criteria

3.2.2.1 Inclusion criteria

- RCTs using an analysis approach that assumes PH including Cox PH regression and log-rank test
- RCTs that have a comparative element of two treatments
- RCTs are phase II/III studies

3.2.2.2 Exclusion criteria

- RCTs analysed using methods not assuming PH
- No TTE analysis conducted
- RCTs where TTE data is treated dichotomously so OR were reported so no HR reported.
- RCTs where additional data has been obtained from the study authors to conduct TTE analysis in systematic review

- Not possible to access the RCT publication
- Full-text not available in English
- Abstracts

3.2.3 Screening of studies

All RCTs included in the sample of systematic reviews chosen for inclusion were screened for eligibility. For consistency, the same TTE outcome that is used later in Chapter 4 was used.

3.2.4 Data Extraction

Data extraction was performed by creating a pilot database of all extracted data in Microsoft Excel. The content of the data extraction database was based on guidelines for the reporting of survival outcomes and analyses in Batson *et al*[35]. The first version of the database was piloted using three RCTs to ensure sufficient data was being captured. A screenshot of the database is provided in Appendix 1, but in brief included:

- Review title
- RCT title
- TTE outcome (Only one endpoint per RCT has been reported here as the main focus has been on reporting the same survival outcome as used in Chapter 4)
- Method of analysis (i.e. Cox PH model, log-rank test etc)
- Sample size
- Clinical Area
- Length of follow-up
- Level of censoring observed
- HR (95% CI)

- Whether the PH assumption was assessed
- Method used to assess the PH assumption
- Result of PH assumption
- If PH assumption was not valid was an alternative method used instead. If so, what method was used.
- Graphical plots such as Kaplan-Meier (K-M) plot presented
- Number at risk/Number of events reported

3.2.5 Clinical Trials Unit Survey

A survey was conducted to understand what methods were used by statisticians to analyse TTE outcomes in RCTs with particular focus on the assessment of PH. To conduct this survey all UKCRC registered CTUs in the UK were contacted, to supplement the review of RCTs described in Section 3.2.4, which could be prone to selective reporting of the information about the PH assumption. Hence, by conducting a survey of practice it provides a better indication of what is done routinely rather than what is reported. The survey consisted of three questions regarding the methods used for the analysis of TTE outcomes, what methods they used to assess the PH assumption when using a method that assumes PH and what they would do if the PH assumption was invalid. Statisticians at each CTU were approached and were asked to provide responses on behalf of the CTU, so they could discuss the survey with colleagues before completing the survey. An online link was sent to the Registered CTU co-ordinator who then sent it to all CTUs. One email reminder was sent out after three weeks. I received email notification on completion of the survey. A glossary of all methods mentioned in the survey was also sent out. A copy of the survey is included in Appendix 2.

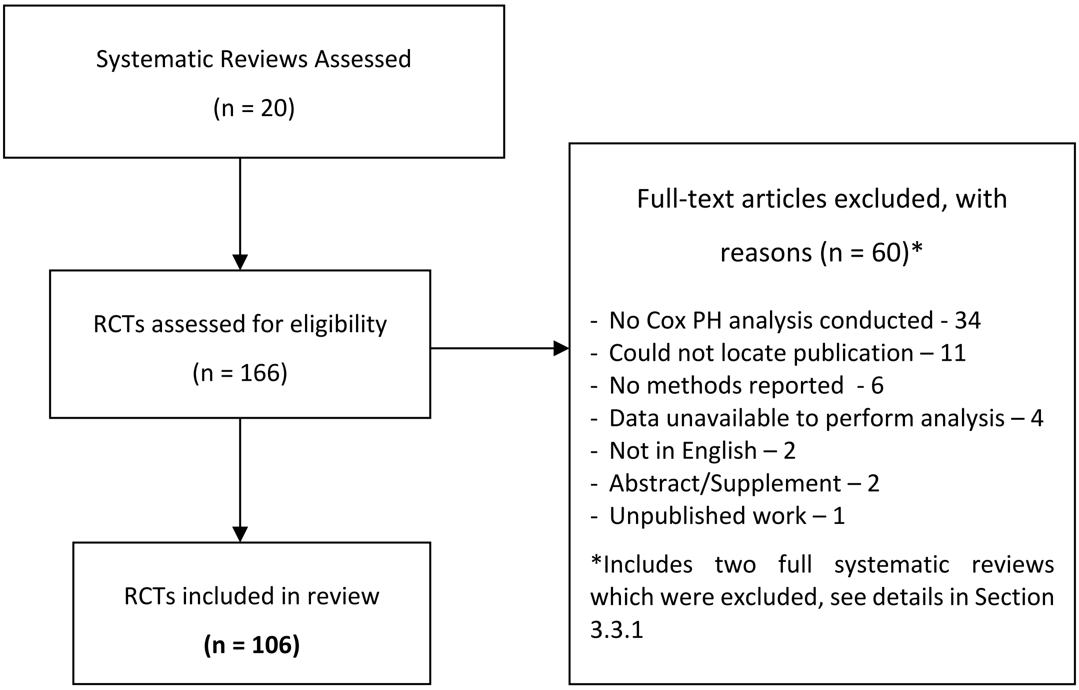
3.3 Results

3.3.1 Characteristics of included RCTs in the review

166 RCTs that had been included in 20 systematic reviews were assessed for eligibility. In total 106 out of 166 (64%) publications of Phase III RCTs were included from across 18 reviews. Two full reviews had to be excluded as for one review, four of the publications could not be located and one study was in Japanese and for the other review none of the publications conducted any Cox PH analysis. The study flow diagram is shown in Figure 4. The references for all 106 RCTs are given in Appendix 3. It is important to note that no duplicate studies were identified.

The publications detailed RCTs in a range of clinical areas including Cardiology, Neurology, Oncology and Epilepsy. Total sample sizes in the RCTs ranged from 26 to 33,357 patients with a median sample size of 408 patients. The year of publication of the RCTs ranged from 1985 to 2013.

Figure 4: PRISMA flow diagram for RCT review


Systematic Reviews Assessed
(n = 20)
RCTs assessed for eligibility
(n = 166)
RCTs included in review
(n = 106)
Full-text articles excluded, with reasons (n = 60)*
- No Cox PH analysis conducted - 34
- Could not locate publication – 11
- No methods reported - 6
- Data unavailable to perform analysis – 4
- Not in English – 2
- Abstract/Supplement – 2
- Unpublished work – 1
*Includes two full systematic reviews which were excluded, see details in Section 3.3.1

For 56 out of 106 (53%) of the RCTs, the TTE outcome was overall survival and a further 14 RCTS (11%) reported results for a composite outcome such as a combined endpoint of fatal stroke, fatal myocardial infarction, and other fatal cardiovascular disease. For the other category, endpoints such as time-to-treatment failure, event-free survival and graft loss were included. (Table 1)

Overall survival was consistently defined as time-to-death from any cause, and progression free survival (PFS) was consistently defined as time to progression or death from any cause.

Table 1: Survival Endpoints of Included RCTs

Survival Endpoints	**N (%)**
Overall Survival	56 (53)
Composite outcome	14 (13)
Progression-free Survival	4 (4)
Time to event[1]	4 (4)
Other[2]	28 (26)

Time to event[1] = time to first seizure, time to first diabetic event etc. Other[2]= time-to-treatment failure, event-free survival, graft loss and risk of major events.

3.3.2 Statistical Methods

As shown in Table 2, the most common statistical method used for comparing treatment groups was the Cox PH model, which was used in 68 out of 106 (64%) RCTs that had used an analysis method that assumes PH. Around 31% of RCTs reported the log-rank test to compare treatment groups and generate p-values. 101 (95%) RCTs reported univariate analyses, whilst only five RCTs reported using multivariate analyses.

Table 2: Statistical methods used across RCTs

Statistical methods used across RCTs	**n (%)**
Cox PH	68 (64)
Log rank test	33 (31)
Stratified Cox PH	5 (5)

Presentation of HR

The Cox PH model including stratified Cox PH model was reported in 73 (69%) publications as either a univariate or multivariable analysis. 57 out of 73 (78%) RCTs presented results from the Cox PH models as HR and associated 95% CIs. Around 17 out of 106 (16%) RCTs presented results as relative risk or relative risk reduction alongside a 95% confidence interval and/or p-value. For seven out of 106 (7%) RCTs only the HR and p-value or only the p-value was reported.

Assumption of PH

Testing of the PH assumption was only reported in 12 out of 106 (11%) RCTs that conducted a survival analysis using Cox PH model and log-rank test. Graphical methods were the most popular approach for assessing the PH assumption including log-cumulative hazard versus log (time) plots, cumulative hazard versus time plots and Schoenfeld residual plots as seen in 7[7 74-79] out of 12 (58%) RCTs as detailed in Table 3. Two of the publications used more than one method to assess proportionality: K-M curves to visually assess the assumption followed by using a more formal method: Schoenfeld residuals and another RCT used log-cumulative hazard vs log(time) plots to visually assess the proportionality assumption followed by including a time-dependent covariate in the model. Three (25%) RCTs included time-dependent covariates in the Cox PH model to investigate proportionality and two RCTs mentioned they assessed the PH assumption but do not provide any details on methods used for assessing the assumption.

Table 3: Methods used for assessing the PH assumption

Methods for assessing the PH assumption	**n (%)**
Use time-dependent covariate	3 (25)
Log-cumulative hazard vs log (time) plots	3 (25)
Log-cumulative hazard vs log (time) plots and time-dependent covariate	1 (8)
Cumulative hazard vs time plots	1 (8)
Schoenfeld residuals	1 (8)
K-M curves and Schoenfeld residuals	1 (8)
No details given	2 (17)

In terms of results, only five out of the 12 (42%) publications mentioned whether the PH assumption was valid or not. Of these, three publications reported that the PH assumption was considered reasonable, one publication mentioned that the PH assumption was not valid so they reported the RMST difference in the paper, and one did not mention that they had tested for proportionality but included a time-dependent covariate in the model and mentioned that as p-value=0.54, there is no change in the results. The remaining seven RCTs did not mention the results of the PH assumption checking.

Changes over time

Figure 5 reports on whether the practice of assessing non-PH within RCTs has improved over time.

Figure 5: Histogram of the assessment of the PH assumption over time in RCTs

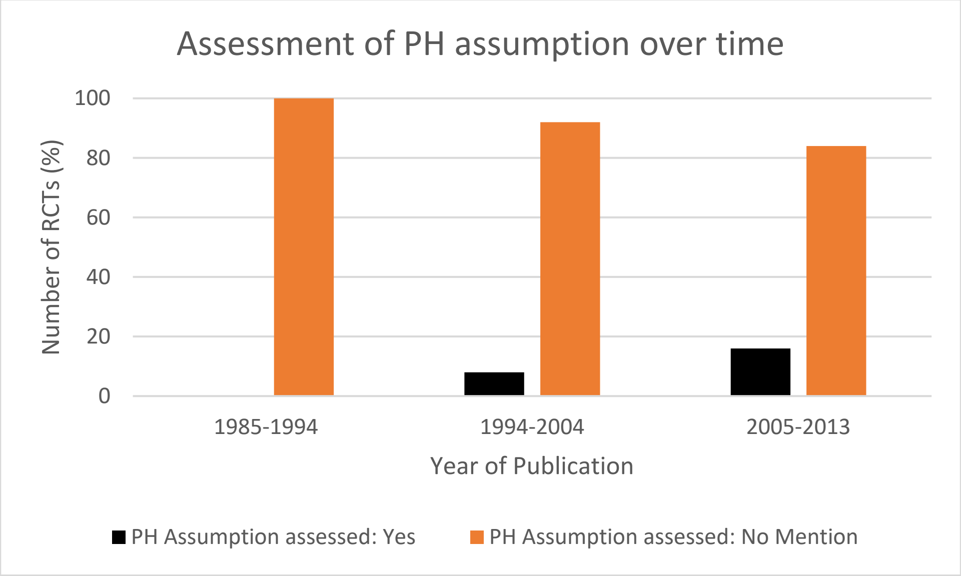


Figure 5 explains that the levels of reporting of the PH assumption have improved over time reflecting signs of an association between time of publication and reporting standards. However, as only 12 (11%) RCTs have assessed the PH assumption the numbers are very low to draw any conclusions, but the numbers do suggest a positive trend.

Graphical display

Only 75 out of 106 (71%) RCTs reported survival curves in the publications. The method of survival curve calculation across all publications was the K-M method, although this was not always clearly stated. The patient numbers at risk were reported in 38 (36%) of the RCTs and the number of events was reported in 8 out of 106 (7.5%) RCTS. All of the publications clearly differentiated between treatments in the survival curves. For the majority of the K-M curves, the quality and visibility of the curves was clear, however there were a few publications that had used relatively thick lines which made it difficult to distinguish points where treatments had very similar survival probabilities.

3.4 CTU Survey results – analysis of TTE outcomes at the trial level

As of January 2018, there was 51 UKCRC registered CTUs in the UK. In total 31 (61%) CTUs completed the survey over a five week period. A reminder email was sent out three weeks after the initial email for any further responses. Table 4 shows the results from the first question on the methods used by the CTU for analysing TTE outcomes. The majority of the CTUs use the Cox PH regression, K-M method and log-rank test for analysing TTE outcomes. Over 50% of CTUs said that they either sometimes or frequently used parametric PH models and flexible parametric models for analyses. There were some other methods suggested by CTUs but these were methods used by individual CTUs.

Table 4: Methods used for analysing TTE outcomes

What method do you use for the analysis of TTE outcomes?	**Frequently**	**Sometimes**	**Never**
- Cox PH regression	28 (90.3%)	2 (6.5%)	1 (3.2%)
- Kaplan-Meier method	25 (80.6%)	5 (16.1%)	1 (3.2%)
- Log-rank test	23 (74.2%)	7 (22.6%)	1 (3.2%)
- Parametric PH model	4 (12.9%)	14 (45.2%)	13 (41.9%)
- Accelerated Failure Time model	1 (3.2%)	9 (29.0%)	21 (67.7%)
- Flexible parametric model	1 (3.2%)	17 (54.8%)	13 (41.9%)
Other:	**N (%)**		
Restricted Mean Survival Time (RMST) or Combined Test only if PH invalid	2 (6.4%)		
Competing risks e.g. Fine and Gray model	1 (3.2%)		
Frailty modelling	1 (3.2%)		
Multi-level PH model	1 (3.2%)		
Parametric non-PH model such as time-dependent logistic model	1 (3.2%)		

The results in Table 5 highlight that all CTUs did assess the PH assumption when using a method that assumed PH, with majority using Schoenfeld residuals, K-M plots, log-cumulative hazard plots and time-dependent covariates. The majority of the methods being used by the CTUs are graphical approaches.

Table 5: Methods used most commonly for assessing the PH assumption

If using a method that assumes PH which methods are used most commonly to assess the PH assumption	**Number of CTUs (%)**	
	N/A	**Yes**
(a) Assumption not assessed	31 (100)	0 (0)
(b) Schoenfeld residuals	3 (9.7)	28 (90.3)
(c) Kaplan-Meier plots	11 (35.5)	20 (64.5)
(d) Lee and Pirie method	31 (100)	0 (0)
(e) Log cumulative hazard	9 (29.0)	22 (71.0)
(f) Time dependent covariates	12 (38.7)	19 (61.3)
Other	**N (%)**	
Deviance residuals	1 (3.2)	
Kolmogorov-type supremum test (SAS test for PH assumption	1 (3.2)	

The results in Table 6 are by number of responses rather than number of CTUs as in the previous two questions, as some CTUs listed multiple alternative approaches for analysis. Three CTUs reported that the method of analysis that assumes PH would still be used if the PH assumption was invalid. The rest of the 28 CTUs suggested that alternative approaches they would take for analysing the data, such as using time-dependent covariates in the model, using flexible parametric models and using RMST. However, some of the CTUs along with listing what they would do also mentioned that the approach used would also depend on other factors including degree of non-proportionality, the importance of the outcome, if the survival curves cross and if there is a total absence of valid interpretation. One CTU also mentioned that they have never found PH to be invalid so not had to use an alternative approach.

Table 6: Methods used if the PH assumption is invalid

If the PH assumption is invalid, what approach would you take for the analysis	**Number of CTUs (%)**
- Continue to use the method assuming PH	3 (7.7)
- Use an alternative method that does not assume PH (Please specify)*	
Parametric models	2 (5.1)
Include time-dependent covariates in model	14 (35.9)
Accelerated failure time model	2 (5.1)
Flexible parametric model	6 (15.4)
Piecewise models	2 (5.1)
Restricted Mean Survival Time	5 (12.8)
Shared frailty	1 (2.6)
Test statistics using "Combined test" which combines Cox and permutation test based on RMST	1 (2.6)
Landmark analysis	2 (5.1)
Subgroup analyses	1 (2.6)

*CTUs could select multiple responses

3.5 Discussion

This review highlights that although there is much better reporting of TTE outcomes in RCTs compared to the 1990s when the first known review of TTE outcomes by Altman *et al*[30] was published, the reporting of the PH assumption is still a significant issue. The use of suitable statistical methods are of key importance within evidence-based medicine, with huge potential to impact decision making with inappropriate method use.

3.5.1 Summary of key findings and implications

This review considers 106 RCTs identified from 18 systematic reviews, published between 1985 and 2013. The most common disease areas included Cardiology, Neurology, Oncology and Epilepsy.

The most commonly used statistical method for comparing treatment groups that assumes PH was the Cox PH model reported in 73 (69%) RCTs, followed by the log rank test reported in 33 (31%) RCTs. Although both of these approaches rely on the validity of the PH assumption, testing of this assumption was only reported in 12 out of 106 (11%) RCTs. There was variation in the approach used with 7 (58%) RCTs using graphical methods, 3 (25%) publications including time-dependent covariates in the Cox PH model and 2 (17%) RCTs not providing any details on what method was used.

Only 75 out of 106 (71%) RCTs reported survival curves in the publications. An approach suggested by Guyot et al[29] demonstrates how the survival data can be reconstructed from published Kaplan-Meier survival curves by digitising the curves and obtaining pseudo-IPD. Although a reconstruction algorithm has been written as a R function by Guyot et al[29], extensions to this algorithm have been made by Wei et al[80], which requires the published survival-curves and the number of patients at risk reported in the trial publication. This extension to the algorithm is more user-friendly and easy to use as explained in Chapter 5. However, with only around 36% of RCTs reporting the number of patients at risk in this review, there is very little scope for additional checks to be carried out if the necessary information is not available within publications.

The CTU survey results also highlight that the Cox PH regression method is the most common method for analysing TTE data, being used by over 90% of UKCRC registered CTUs. The results also highlighted that more complex methods like parametric PH models and flexible parametric models are also being used by more than 50% of CTUs. All of the CTUs mentioned that they do assess the PH assumption when using a method that assumes PH. It is unclear whether the CTUs mentioned this as they are telling me what they think I want to hear about what the correct approach is whether this is what they are doing in practice. Unfortunately, as the studies included in this review are only until 2013, and more recent publications are not included it is difficult to know what CTUs are doing in practice.

Similar to the IPD reviews, majority of the CTUs are also using graphical approaches for assessing the PH assumption. Although, most of the CTUs suggested alternative methods they would use to analyse TTE data if the PH assumption was invalid, many CTUs mentioned that the approach used would also depend on other factors including degree of non-proportionality, the importance of the outcome, if the survival curves cross and if there is a total absence of valid interpretation. These comments suggest that even when the CTUs are aware that the PH assumption is invalid, the decision to perform an alternative analysis is dependent on other factors despite whether the results are biased and based on incorrect assumptions. Three CTUs reported that they would use a method that assumes PH even if the PH assumption was invalid. It remains unclear whether CTUs are performing this in addition to using an alternative method not dependent on the PH assumption. Within CTUs, clinicians usually prefer reporting the HR as it's the most commonly reported effect measure for TTE data. As previously mentioned in Section 2.1.8, in the presence of non-PH the "average" HR could be reported but provided it's clear that the effect is changing over time in which case a time-dependent effect could be included in the model. This would be similar to reporting a very heterogeneous meta-analysis where a pooled effect would be reported but which needs to be interpreted with the heterogeneity statistics as well as exploring why the effect is changing.

In summary, findings of this review again demonstrate the poor reporting of the PH assumption despite previous recommendations[30 31] to encourage authors to improve the reporting of the PH assumption in journal articles.

3.5.2 Strengths and limitations

Previous reviews[30 31 35] have either included RCTs and observational studies or considered the reporting of TTE data within Oncology and using specific journals. To my knowledge, this is the first review to consider the reporting of Phase II and III RCTs where no journal or disease area restriction has been made.

Due to time constraints, only a selection of systematic reviews were chosen at random from which 106 RCTs were identified to be included in this review. Therefore, it is possible the rate of PH assumption reporting may differ in a larger, or different, sample.

In this review, Phase IV RCTs were not considered as from a quick search it was clear that it was difficult to identify Phase IV RCTs that met the inclusion/exclusion criteria. Therefore, it is unlikely that by not including Phase IV studies in the inclusion criteria, there will be much impact on the number of included studies.

3.5.3 Comparison to previous work

Chapter 1, Section 1.4.2 summarises the findings of previous reviews of TTE outcomes. It should be noted that previous reviews have varied in characteristics, inclusion criteria and objective, therefore all comparisons made between previous reviews and this review are informal and narrative and results of each should be interpreted within the context and objective of the review.

Results from this review are in agreement with previous work[35] about the Cox PH model being the most commonly used method for analysing TTE outcomes. The RCTs included in this review included better quality of survival curves than seen in previous reviews with Altman *et al*[30] and Abraira *et al*[31] reporting the poor quality of survival curves. In Altman *et al*[30], the included studies were research papers published between October and December 1991 and in Abraira *et al*[31], included studies that were published in 1991 compared to 2007. Out of the 75 RCTs that presented survival curves, majority of them were of good quality suggesting that there has been considerable improvement in the quality of the survival curves over time.

Similar to the results of this review, Batson *et al*[35] highlight the poor reporting of the PH assumption in 7% (2 out of 28) of RCTs, and Altman *et al*[30] reported that the PH assumption was assessed in even fewer, only 5% (2 out of 132) of publications, whilst Abraira *et al*[31] reports that in 1991 the PH assumption was assessed in 10.6% (5 out of 104) of publications whereas in 2007 the PH assumption was assessed in 26.3% (47 out of 240) of publications. The results illustrate that although the 'assessment of model assumptions' is a criteria included in the suggested guidelines checklist published by Altman *et al*[30] in 1995 and then also by Abraira *et al*[31] in 2013, further work is still required to ensure authors are not only assessing the PH assumption and mentioning it in the publication, but are also explaining how they assessed the assumption and what the results were, and what appropriate action was taken dependent on the results.

All previous reviews[30 31 35] have concluded that an improvement is needed in the level of reporting of TTE outcomes in journal publications for RCTs. Many of the guidelines that are available focus on the presentation of TTE outcomes but it seems until the impact of the PH assumption violation is known and further recommendations can be made, we may see little improvement in the reporting of the PH assumption over time.

3.5.4 Concluding remarks

The reporting of the PH assumption in RCTs is an area that has been mentioned in the past[30 31 35], especially how poor the reporting of the PH assumption is but it seems no steps have been taken to tackle the poor reporting. Further work to explore the reporting of the PH assumption in STAs is included in Chapter 6 to allow us to further understand whether the reporting of the PH assumption is also an issue within cost-effectiveness analysis or whether it only affects clinical effectiveness analysis.

A simulation study which helps understand the impact of the violation of the PH assumption on results and conclusions is presented in Chapter 7.

4 Review of the reporting of the Proportional Hazards Assumption within Meta-Analyses

4.1 Introduction

Previous work has demonstrated that, although there are alternative methods available to analyse TTE data that do not require the assumption of proportional hazards (such as AFT Model or Flexible Parametric Model), trialists continue to use PH approaches such as the Cox PH model due to its widespread use and to aid comparability with results from other trials.[8]

Meta-analysis of TTE data is usually performed using the HR from each study, and so the implicit assumption of a constant HR over time, i.e. that the PH assumption is valid, is being made. Where IPD is available, the PH assumption can be assessed using approaches such as the Schoenfeld residuals. However, when only study level data is available, it is difficult to assess the PH assumption. When the PH assumption is violated, both the results from the study and meta-analyses will be biased[35]. It is currently unknown what impact including studies, where the PH assumption is violated can have on meta-analysis.[35]

Aggregate data meta-analyses make up the vast majority of the applied meta-analysis literature. Previous work has indicated that up to the year 2004, less than 10% of published meta-analyses per year used IPD (in fact, for most years the figure was less than 5%).[81]

To my knowledge, there have not been any reviews conducted on the reporting of the PH assumption within meta-analyses. The reporting of the PH assumption continues to be a problem within publications of randomised trials as seen in Chapter 3, never mind within meta-analysis. Simmonds et al[43] explained in 2011 that “a greater awareness of the proportionality assumptions of the analysis methods is needed in meta-analyses, and investigation and testing of proportional hazards or odds assumptions should be a standard part of meta-analyses if interpretation of the findings is to be appropriate.”

Therefore, the objectives of this chapter are to (i) review the reporting of analyses of TTE outcomes within systematic reviews; and (ii) to assess the reporting of the PH assumption within systematic reviews. It is also of interest to review whether the practice of assessing non-PH within reviews has improved over time or not.

4.2 Methods of the review

4.2.1 Systematic search

In order to identify all eligible systematic reviews, a systematic search of the following databases was carried out (see Appendix 4 for the search strategy):

- Cochrane Database of Systematic Reviews
- Database of Abstracts of Reviews of Effects
- MEDLINE (via Ovid)
- MEDLINE In-Process (via Ovid)
- EMBASE (via Ovid)
- PubMed

The first two databases form part of the Cochrane Library[82].

4.2.2 Eligibility Criteria

The eligibility criteria has been applied at the review level rather than RCT level. Therefore, the methodology used within the review and the data included within the review was used to decide if the review was included or not rather than the individual RCTs included in the review.

4.2.2.1 Inclusion Criteria

- Systematic reviews including meta-analyses of RCTs of adults and/or children of all parallel designs (e.g. superiority, non-inferiority, equivalence etc.) reported in a full-text journal article or Cochrane database.
- Systematic review includes RCTs that are phase II/III studies
- Systematic review includes a meta-analysis of TTE outcome data that has been analysed using methods assuming PH including Cox PH model and log-rank test
- Systematic reviews published between 2005 and 2015 in order to capture the most recent methods and allow the review to be manageable

4.2.2.2 Exclusion Criteria

- Systematic review includes non-randomised, observational or cohort studies or report such as letters, reviews, editorials, comments on journal articles etc.
- Systematic reviews of prognostic factors
- Systematic review includes TTE data that has not been analysed using methods that assume PH
- No TTE analysis conducted
- Systematic reviews where TTE data is treated dichotomously so risk ratios (RR) or OR were reported rather than HR being reported.
- Narrative systematic reviews (those without any included meta-analysis).
- Provisional/conference abstract.

- No comparative treatments analysed (single treatment group)
- Full-text not originally published in English to allow for an assessment of outcome and statistical reporting by English speaking reviewers.

It is important to note that provided a review included TTE data and used methods that assume PH to analyse the data, then it was included in the review regardless of whether some RCTs used methods that assume PH and some used methods that do not assume PH.

4.2.3 Screening of studies

All studies identified in the systematic search of the electronic databases were screened for eligibility. To begin with, the title and abstract was screened followed by full-text screening. If a full-text manuscript of an abstract could not be found, it was first requested from the library and where that was not possible or the article was not available, the abstract was excluded. Any uncertainty over inclusion/exclusion of studies was discussed with CTS and a decision was made whether to include the review or not. No references were excluded during the screening of the title and abstract.

4.2.4 Data Extraction

Data extraction was performed using a database created in Microsoft Excel (included in Appendix 5). The content of the database was based on guidelines for the reporting of survival outcomes and analyses in Altman *et al*[30] and Abraira *et al*[31]. The first version of the database was piloted using a few systematic reviews to ensure sufficient data was being captured to allow us to understand what the authors have done and what the results are.

Data extraction was performed on all studies by AK. The first stage of data extraction was performed on all eligible studies identified in the search. At this stage the definition of the outcome was noted and whether more than one TTE outcome was included in the review. Due to time constraints and high numbers of included reviews only one outcome per review was included. The preference was to use the primary TTE outcome unless there was no primary TTE outcome in which case a secondary TTE outcome was included. If multiple secondary outcomes were specified then the first listed TTE outcome was included.

Once details concerning the definition of the outcome had been extracted, the next stage was to extract information on the following:

- Review title and date of publication
- Name of journal
- Clinical Area
- Details on time to event outcome: primary or secondary outcome; what the outcome is and the number of TTE outcomes included
- Number of included studies in systematic review
- Number of patients included in systematic review
- Whether IPD or aggregate data used or both
- Method of meta-analysis of TTE outcome: whether one- or two-stage method, inverse-variance weighting or DerSimonian and Laird random effects model etc.
- Additional details provided on method of analysis
- Graphical plots such as K-M plot presented
- Whether K-M plot presented for pooled data or individual studies
- Number at risk if K-M plot presented
- Whether PH assumption was assessed in the systematic review
- Method used to assess the PH assumption in the systematic review

- Result of PH assumption
- If PH assumption was not valid was an alternative method used instead. If so, what method was used.

4.2.5 Data Analysis and presentation of results

Numerical results are presented as numbers and percentages and medians with the minimum and maximum, where specified. No formal statistical analyses were conducted.

For the purposes of reporting results and as the method used for PH assumption checking can vary according to the type of data included in a systematic review, each of the systematic reviews were classified according to the type of data used for analysis: (i) systematic reviews of aggregate data, (ii) systematic reviews of IPD and (iii) systematic reviews involving both aggregate data and IPD. For aggregate data meta-analyses, the systematic reviews were categorised according to the method of meta-analysis used: the inverse-variance weighted method[15] (fixed effect), a Mantel-Haenszel method[23] (fixed effect) or DerSimonian and Laird method[21] (random effect) (further details are given in Chapter 2, Section 2.2.1). For the IPD meta-analysis, two distinct statistical approaches were considered, whether the data had been analysed using a one-stage or a two-stage method (see Chapter 2, Section 2.2.2 for further details). For reviews involving both IPD and aggregate data, the following statistical approaches were considered, whether a one-stage or two-stage method was used for analysing IPD followed by whether a fixed or random effects statistical model as listed under aggregate data was used for pooling the IPD and aggregate data together (see Chapter 2, Section 0 and 2.2.3 for further details on these methods).

However, some systematic reviews did not include sufficient details to understand which method had been used, with mainly details on whether a fixed effects or random effects model had been used not being specified. Therefore, in these cases the method of analysis has been reported as unclear.

4.3 Results

4.3.1 Search strategy for Systematic reviews

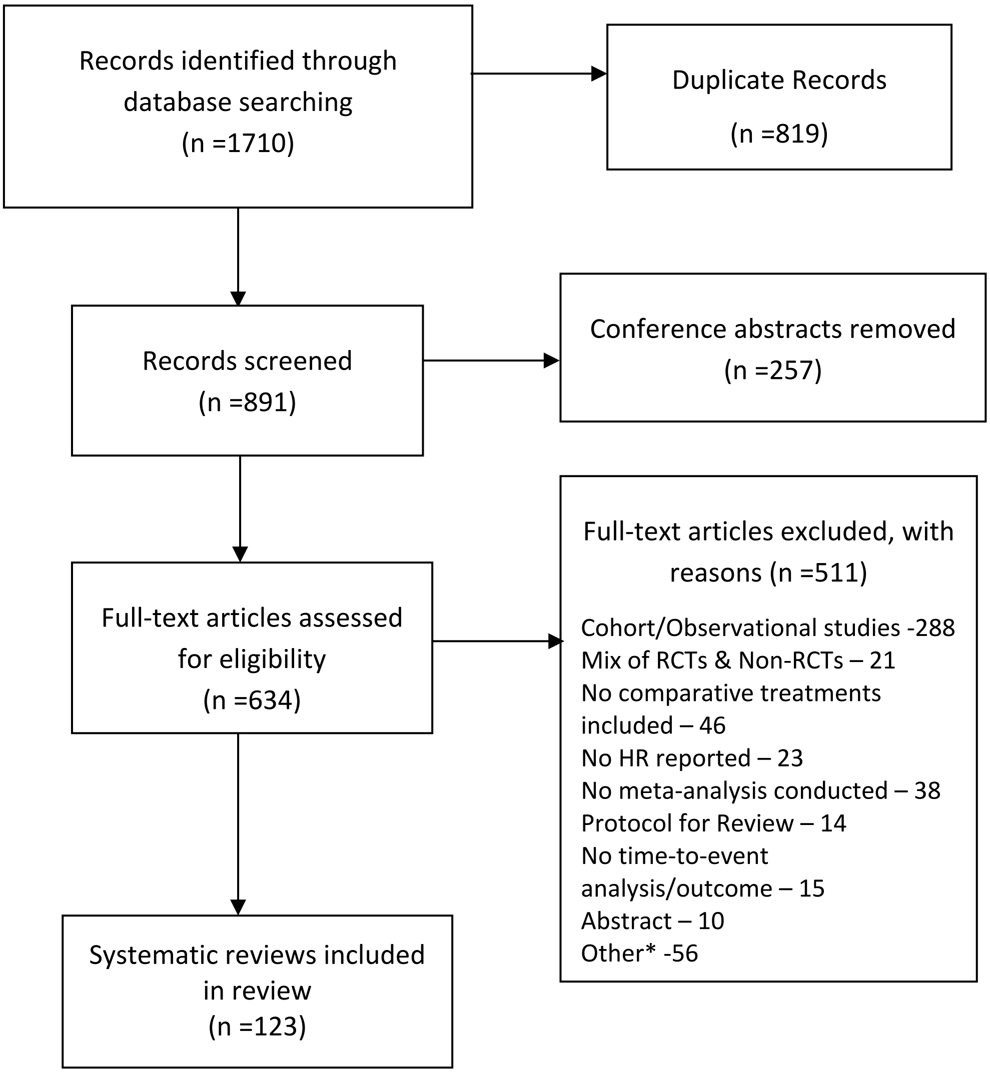
The electronic search to identify the systematic reviews outlined in Section 4.2.1, identified a total of 1710 references, which were downloaded into Endnote software. After removing 819 duplicate reviews, the inclusion and exclusion criteria was applied to the remaining 891 references. At this stage, 257 conference abstracts were excluded, although every effort was made to find the full-text articles they were linked to. A further 511 full-text articles were excluded (for reasons such as the included studies being cohort/observational studies, no meta-analysis being conducted and no HR being reported) resulting in 123 full-text systematic reviews being eligible for inclusion in the review.

Some rapid review approaches were used when performing this review including:

- Search limit: limited by date and language
- Screening: title/abstract and full-text screening performed by one reviewer only
- Data extraction: one person extracted the data

The PRISMA diagram in Figure 6 describes the screening process, including reasons for exclusion at each screening stage. See Appendix 6 for a reference list of the 123 included reviews.

Figure 6: PRISMA flow diagram for Systematic reviews


Records identified through database searching (n =1710)
Duplicate Records (n =819)
Records screened (n =891)
Conference abstracts removed (n =257)
Full-text articles assessed for eligibility (n =634)
Full-text articles excluded, with reasons (n =511)
Cohort/Observational studies -288
Mix of RCTs & Non-RCTs – 21
No comparative treatments included – 46
No HR reported – 23
No meta-analysis conducted – 38
Protocol for Review – 14
No time-to-event analysis/outcome – 15
Abstract – 10
Other* -56
Systematic reviews included in review (n =123)

HR=Hazard Ratio; RCT=Randomised Controlled Trial

*Other = Reasons include methodology papers, reviews of cost-effectiveness analysis, review of phase IV study, dose response study, multi-arm trials and systematic reviews of network meta-analysis (NMA), diagnostic test accuracy (DTA) and overview of reviews.

Data extraction on the assessment of the PH assumption was not always straightforward. In some cases, a simple search for ‘proportional’ or ‘PH’ was sufficient but in some cases manual reading was necessary. For some reviews it was only stated in the methods section that the assumption was assessed but no results from assumption checking were provided so in these situations it was necessary to manually read the results and discussion section to see if any results or information were provided.

4.3.2 Study and participant characteristics

The study and participant characteristics are presented in Table 7. The 123 eligible reviews were published between January 2005 and August 2015 and included 956 studies. OS was the most common outcome with 46% of reviews including it as an outcome. The majority of the systematic reviews had included studies in the two major disease areas, Oncology (48%) and Cardiology (22%). 22% (27 out of 123) of the reviews were published in the Cochrane database of systematic reviews. The number of studies included in the systematic reviews ranged between 2 and 53.

Table 7: Study and participant characteristics

Characteristics of Included reviews	Number of reviews (% of 123 reviews)
Clinical Area	
Oncology	59 (48)
Cardiovascular	27 (22)
Epilepsy	5 (4)
Chronic obstructive pulmonary disease	4 (3)
Surgery	4 (3)
Other[1]	24 (20)
Review Outcome	
Overall Survival	56 (46)
Time to first event	17 (14)
Composite of death	7 (6)
Progression-free survival	6 (5)
Disease-free survival	4 (3)
Time to withdrawal	4 (3)
Other[2]	29 (24)
Review Type	
Cochrane Review	27 (22)
Journal	96 (78)
Publication Year	
2005-2008	33 (27)
2009-2012	50 (41)
2013-2015	40 (32)
Number of included studies	
Median (Min, Max)	5 (2, 53)

Other[1] = Reasons include Arthritis, Diabetes, Fractures, Stroke, Outpatient Care and Depression. Other[2] = Reasons include Major adverse cardiac events (MACE), duration of hospital stay, event-free survival, risk of event, time to treatment failure and time-to-treatment progression.

4.3.3 Results from the Systematic reviews

Of the 123 reviews that fulfilled the inclusion criteria, 35 reviews (28%) included only aggregate data, 81 (66%) included IPD and 7 (6%) included both IPD and aggregate data. Four (3%) reviews included studies with IPD and aggregate data but for analyses purposes only used the IPD and ignored the aggregate data. For reporting we included these four studies in the IPD category.

All of the systematic reviews included in this review had either analysed the data using Cox PH regression (as they had IPD (88 (72%) reviews) or extracted results from a RCT that had analysed data using a Cox regression model (35 (28%) reviews). As specified by the inclusion criteria of this review, all of the included reviews had summarised the treatment effect across studies using a pooled HR and therefore had made the implicit assumption, that the HR is constant over time. However, only 27% (33 out of 123 reviews) examined whether this was a reasonable assumption to make. A few examples of how authors assessed the assumptions include “The assumption of proportional hazards was explored graphically and by carrying out a test for proportionality of the interaction between variables included in the model and the logarithm of time.” and “The appropriateness of the proportional hazards assumption was evaluated visually using Kaplan–Meier plots of the survival curves for each treatment.” None of the remaining 90 reviews mentioned the assumption in the systematic review. Further details on the methods used for assessing the PH assumption are given by data type in Table 9.

Changes over time

Figure 7 reports on whether the practice of assessing non-PH at the review level has improved over time or not.

Figure 7: Histogram of the assessment of the PH assumption over time in Systematic Reviews

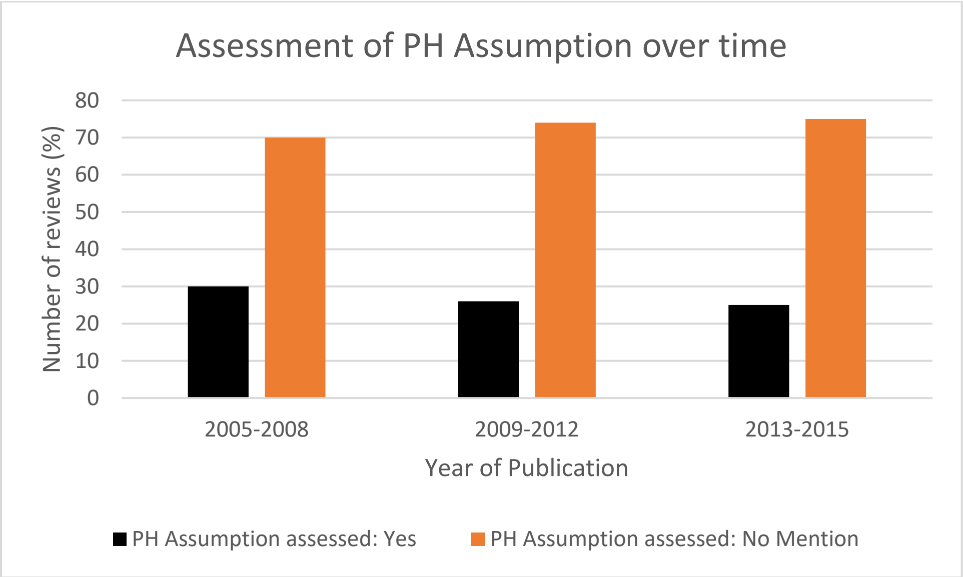


Figure 7 highlights that the levels of reporting of the PH assumption have not improved over the past decade with only around 25-30% of reviews reporting the PH assumption in systematic reviews and this remaining fairly constant. Prior to 2005 the Altman et al[30] review had mentioned the poor reporting of the PH assumption and subsequently included a set of "minimum requirements" for reporting TTE outcomes which includes "When Cox regression analyses are performed, describe the criteria used to select the variables in the initial model, the procedure to specify the final model and describe any methods used to assess the model assumptions." In 2013, Abraira *et al*[31] also found the reporting of the PH assumption to be poor and therefore updated the set of "minimum requirements" outlined by Altman et al[30] which state "When using regression models, report the method used and results of model assumptions checking (e.g., the proportional hazards assumption in Cox models or distributional form in parametric models)." There has been a publication focused on how to assess the PH assumption within TTE outcomes and in particular within meta-analysis in Williamson et al[22] in 2002 and then in 2011 Simmonds et al[43] concluded with the help of a simulation study that the PH methods will be biased when the hazards are not proportional and that investigation and testing of the PH assumption should be a standard part of meta-analyses if interpretation of the findings is to be appropriate. Since 2015 there have been a few more publications on the reporting of the PH assumption as seen in Batson et al[35] in 2016 as well as what to do if the PH assumption is invalid as seen in De Jong et al[42] in 2020. Figure 6 shows that even though there has been a few publications highlighting the poor reporting of the PH assumption as well as publications that show how to assess the PH assumption, there hasn't been a surge in review authors reviewing the assumption over the years.

Currently there is no mention of how to assess the PH assumption at the review level or details on what to do if the PH assumption in invalid in the Cochrane Handbook[82]. The NICE have a technical support document (TSD) on survival analysis[83] which was published in 2011[83]. In the document, they focus on TTE modelling methods at the trial level with a strong focus on the PH assumption. The document includes details on how to assess the PH assumption and alternative methods that can be used if the assumption does not hold. An updated TSD was published in 2020[84] focusing mainly on methods not dependent on the PH assumption. Further details on these documents are included in Chapter 6, Section 6.1.1.

4.3.4 Statistical Methods used within the Systematic Reviews

Table 8: Statistical approach used for estimating the pooled relative HR according to the data type used for meta-analysis

Statistical Approach	**Number of reviews (%)**
Aggregate Data	**35**
DerSimonian and Laird	21 (60)
Mantel-Haenszel	6 (17)
Inverse-Variance Weighting	3 (9)
Mantel-Haenszel/Inverse-Variance Weighting	4 (11)
Unclear	1 (3)
Individual Patient Data	**81**
One-stage method	20
- one-stage cox model stratified by trial	8 (40)
- one-stage cox model stratified by covariate	10 (50)
- one-stage cox model (unclear if stratified by trial or trial indicator)	2 (10)
Two-stage method	61
- cox PH model fitted to each trial seperately followed by pooling trial results (two stage) using a fixed effect approach	19 (31)
- cox PH model fitted to each trial seperately followed by pooling trial results (two stage) using a DerSimonian and Laird random effect approach	5 (8)
- cox PH model fitted to each trial seperately followed by pooling trial results (two stage) but unclear if used fixed or random effects model	37 (61)
Individual Patient Data & Aggregate Data	**7**
- cox PH model fitted to each IPD trial separately followed by pooling IPD and aggregate data using a DerSimonian and Laird random effect approach	3 (43)
- cox PH model fitted to each IPD trial separately followed by pooling IPD and aggregate data using an IV weighting fixed effect approach	1 (14)
- one-stage cox model stratified by trial followed by pooling IPD pooled effect and aggregate data using a DerSimonian and Laird random effect approach	2 (29)
- one-stage cox model stratified by trial followed by pooling IPD pooled effect and aggregate data but unclear if used fixed or random effects model	1 (14)

4.3.4.1 Aggregate Data

Methods used to analyse data

Table 8 highlights that for the 35 systematic reviews that included only aggregate data the most common statistical approach used for the meta-analysis (second stage) was the DerSimonian and Laird (random effects) method being used in 60% (21 out of 35 reviews) of reviews, whilst 13 of the reviews (37%) used fixed effects methods.

PH assumption checking

None of the systematic reviews that were based on aggregate data mentioned the assessment of PH of the included trials or at the review level.

4.3.4.2 IPD

Methods used to analyse data

Table 8 reports that for the 81 systematic reviews that included IPD, 75% (61 out of 81) of reviews used a two-stage approach and 20 (25%) of reviews used a one-stage approach. For the two-stage approach, 19 of the reviews (31%) used a cox PH model fitted to each trial separately followed by pooling trial results (two stage) using a fixed effects model (either a Mantel Haenszel approach or inverse-variance weighting approach) whilst five (8%) reviews used a DerSimonian and Laird random effects approach to pool the trial results. For 37 (61%) reviews, it was unclear whether a fixed effects or random effects model had been used for pooling the results. For the one-stage approach, eight (40%) of reviews used a one-stage cox model stratified by study, whilst 10 (50%) used a one-stage cox model stratified by trial indicators. These results seem to be in line with reviews of methods used to perform IPD meta-analysis in current practice published by Simmonds et al in 2005[71] and more recently in 2015[85], where 76% of reviews with TTE data used two-stage methods. It is likely that two-stage methods are used more often than one-stage as one-stage analyses of TTE data are less well developed and little software is available[41 42].

PH assumption checking

Table 9 highlights the methods used for assessing the PH assumption in the systematic reviews including IPD. Testing of the PH assumption was reported in 30 out of 81 (37%) IPD reviews. The remaining 51 (63%) reviews did not mention assessing the PH assumption. 10 out of 30 (33%) reviews assessing the PH assumption used two methods for assumption checking, with a graphical method being used such as Kaplan-Meier plots or log-cumulative hazard plots to visually assess the assumption followed by a formal assessment of the assumption such as including time-varying covariates. The most common approach was assessing the PH assumption using time-varying covariates as seen in seven (23%) reviews. The CTU survey also highlighted that around 60% of CTUs used time-varying covariates to assess the assumption. Similarly, Schoenfeld residuals was another popular approach in the CTU survey for assessing the PH assumption but only used within four (13%) reviews. In three of the reviews, the review authors only visually examined the Kaplan-Meier plots to assess the validity of the assumption. Since visually examining plots is a subjective approach for assessing the assumption, there is no guarantee that it is accurate as it is wholly dependent on the quality of the curves. For five (17%) of the IPD reviews, no details were given on the method used for assessing the method used other than stating that the assumption was assessed. It is important to note that 28 out of the 30 (93%) reviews explored the assumption in each individual trial separately. For two of the reviews, the authors mentioned that for one of the studies the PH assumption was not met. One of the reviews did not elaborate any further whilst the other review stated that overall the PH assumption was met. This review is in agreement with Williamson et al[22] who also suggested that if the PH assumption is valid for one particular study then it is expected that the assumption will be valid for all studies. Another two reviews did not provide any detail on assumption checking in order to understand how they tested the assumption for the meta-analysis.

Table 9: Methods used to assess the PH assumption in the Systematic reviews with IPD

Method for assessing PH assumption	Number of reviews (%)	Number of reviews that presented results (%)
Time-varying covariates	7 (23)	4 (57)
Log-cumulative hazard plot & time-varying covariates	5 (17)	4 (80)
Schoenfeld residuals	4 (13)	3 (75)
Time-varying covariates & Kaplan-Meier curves	3 (10)	1 (33)
Examining Kaplan-Meier curves	3 (10)	1 (33)
Time-varying covariates & Schoenfeld residuals	1 (3)	0 (0)
Log-cumulative hazard plot & Schoenfeld residuals	1 (3)	1 (100)
Kolmogorov-type supremum test	1 (3)	1 (100)
Unclear	5 (17)	2 (40)

The conclusion from the assessment of PH assumption was only reported in 57% (17 out of 30) of the reviews that mentioned exploring it. In 13 (76%) of these reviews, the PH assumption was reported to be valid, two reviews reported the assumption was invalid and alternative methods including a piecewise cox regression model and a log-rank test were used for analysis. It is unclear why a log-rank test was used as an alternative method given that a log-rank test is also dependent on the PH assumption. For two of the systematic reviews, the authors explained that the PH assumption is valid for the majority of the included RCTs but is invalid for one of the RCTs, but do not assess whether the PH assumption is still valid when the data from the individual studies are pooled together and re-analysed. It is unclear whether the assessment of the PH assumption was not reported in the remaining

14 reviews as it was valid, but the review authors didn't think it was important to report in the publication or whether it was invalid but the review authors ignored the assumption or whether the methodology to investigate the assumption was reported but the review authors didn't assess the assumption.

4.3.4.3 IPD and Aggregate data

Methods used to analyse data

For systematic reviews including RCTs with IPD and aggregate data in the analysis, 71% (5 out of 7) of reviews used a cox PH model fitted to each trial separately followed by pooling IPD and aggregate data using a DerSimonian and Laird random effect approach. Only one review used the inverse variance weighting fixed effect approach.

PH assumption checking

The PH assumption was assessed in three out of seven (43%) of the systematic reviews[86-88] with no mention of the PH assumption in the remaining four (57%) reviews. One of the reviews used a log-cumulative hazard plot to visually assess the assumption, whilst another review used a gamma-frailty model to test departures from assumption. One of the reviews first used a log-cumulative hazard plot to visually assess the assumption before using time-varying covariates to formally test the assumption. The results from assessing the PH assumption were reported in two of the reviews with both reporting that the assumption was valid.

4.4 Discussion

This chapter systematically examined systematic reviews of TTE outcomes in relation to the reporting of the PH assumption within meta-analyses.

4.4.1 Summary of key results and implications

This review identified 123 systematic reviews with 956 RCTs included, published between 2005 and 2015, in a range of medical journals and the Cochrane database of systematic reviews. The two most common disease areas for the systematic reviews were Oncology (48%) and Cardiology (22%). Seven composite outcomes were identified compared to 116 non-composite outcomes. No patterns were identified in the assessment of the PH assumption in terms of type of outcomes explored.

All of the systematic reviews included in this review analysed the data using Cox PH regression or obtained results from a RCT that conducted analysis using Cox regression. 73% (90 out of 123) of systematic reviews failed to adequately describe methods and results of appropriate approaches to assess the validity of the PH assumption. The results showed that the practice of assessing the PH assumption at the review level has not improved over time. Currently, the Cochrane Handbook[68] doesn't include any recommendations on how to assess the PH assumption at the review level or on what to do if the PH assumption is invalid. NICE[27] published a TSD[83] in 2011 which includes some details on how to assess the PH assumption and methods that can be used if the assumption is invalid on an individual trial level. An updated TSD was published in 2020[84] which focuses on methods not dependent on the PH assumption. However, it is unclear whether these TSDs are being used only within Evidence Review Groups (ERG) or whether these documents are being used widely. Another approach which can be used to assess the PH assumption is digitisation of published K-M curves to obtain reconstructed pseudo-IPD which can then be used to perform TTE analysis and to assess the PH assumption. This approach has been used at the trial level since 2012, however there hasn't been widespread use of this approach at the review level so far. In order to perform the digitisation approach at the review level, the trial level K-M curves would need to be digitised as part of the data extraction process. Current research conducted on how to assess the PH assumption suggests that there isn't any clear recommendations on how to assess the PH assumption at the review level.

For aggregate data meta-analysis, 35 reviews were identified with 60% of reviews using the DerSimonian and Laird method to estimate the treatment effect. None of the reviews reported assessing the PH assumption.

The number of IPD meta-analyses is growing steadily[85 89 90] with 66% (81 out of 123) of systematic reviews included in our inclusion criteria using IPD. Since meta-analyses of TTE data are challenging if based on published summary data[14] it is not surprising to find that more IPD analyses were conducted in the published literature. For IPD meta-analyses, two-stage methods were used in 75% of reviews. These results are in agreement with current practice as reported by Simmonds et al in 2005[71] where two-thirds of the reviews used two-stage methods and more recently in 2015[85] where 76% of reviews used two-stage methods. The Simmonds et al[85] review published in 2015 included reviews published between 2008 and end of 2014, so there is overlap between the Simmonds et al[85] review and this review which includes systematic reviews published between 2005 and 2015. The biggest difference is that the Simmonds et al[85] review includes any IPD meta-analysis review whilst this review only includes systematic reviews including TTE data.

Testing of the PH assumption was reported in 37% (30 out of 81) of reviews, with seven (23%) reviews using the time-varying covariate method for assessing proportionality. 33% of reviews reported using a graphical method to visually assess the assumption followed by a statistical method for formally assessing the assumption. Graphical methods are subjective so using statistical methods to test proportionality are recommended[52 91]. The results from testing the PH assumption were reported in 57% (17 out of 30) of reviews, with the assumption being valid in 13 (76%) reviews and for two reviews where the assumption did not hold, the review authors used alternative methods.

For reviews with IPD and aggregate data, 6% (7 out of 123) of systematic reviews were identified and the PH assumption was reported in 43% (3 out of 7) of reviews with results being reported for 2 of the reviews.

The results for meta-analyses using IPD, aggregate data and a mix of IPD and aggregate data demonstrate that with access to IPD further analyses and assessments can be carried out which is not possible with summary data. However, there are methods that can be used to investigate the PH assumption using aggregate data as shown by Williamson et al[22], where the log-cumulative hazard plot using survival probabilities can be read off K-M curves, which seems like a simple but effective approach for assumption checking. None of the included reviews used this approach. Although, the method itself is not tricky to use, the quality and clarity of the K-M graph are key to the success of this approach as if curves are too close together or if there is a lot of censoring then reading the survival probabilities accurately can be difficult. Williamson et al[22] suggest that they do not feel the method is particularly useful and could lead to incorrect interpretations.

In summary, findings of this review demonstrate the poor reporting of the PH assumption at the review level. To date there has been no recommendations published on how to assess the PH assumption within meta-analyses as well as what alternative methods can be used when PH is invalid. It is also unknown what impact the PH assumption violation can have on meta-analysis which needs to be further investigated by conducting simulation studies as suggested by Batson et al[35]. Such simulation studies are presented in Chapter 7. An IPD meta-analysis of TTE data was conducted in 2011 by Simmonds et al[43] as described in Section 1.4.2

where they concluded that that PH methods "will be biased when the hazards are not proportional…A greater awareness of the proportionality assumptions of the analysis methods is needed in meta-analyses, and investigation and testing of proportional hazards or odds assumptions should be a standard part of meta-analyses if interpretation of the findings is to be appropriate."

This review also highlighted that it is unclear whether review authors are aware of methods for assessing the PH assumption and choose to ignore the assumption as they do not think violation of the assumption could impact the results or whether authors genuinely do not know how to check the assumption. Hence, further work is needed to understand the impact of the violated of the assumption and also understanding why review authors do not assess the assumption.

4.4.2 Strengths and limitations

To my knowledge, this is the first systematic review of reviews of TTE outcomes to consider the reporting of TTE outcomes in terms of the potential implications for meta-analysis and in particular assessing the reporting of the PH assumption.

Due to time constraints, only multiple electronic databases were searched but did not perform any hand searching of grey literature which is a potential limitation, so might have missed a relevant review however as an extensive search was carried out, it is unlikely it will impact the number of included studies to a great extent. Also only full-text reviews published in English were included due to time constraints, however it is unlikely that this would have impacted the included studies.

A further limitation of this review has been using some rapid review approaches such as screening and data extraction only being performed by myself. Although, the screening and data extraction have been double checked by myself, there is no guarantee to say that there are no errors. However, due to time constraints it was important to use some rapid review approaches.

4.4.3 Concluding remarks

In conclusion, in line with all previous work conducted in this area, the current systematic review has shown concerning reporting inadequacies relating to the reporting of the PH assumption. Current research suggests that there aren't any published recommendations on how to assess the PH assumption at the review level and this could be the reason why reviewers are not reporting the PH assumption results as it may be that they are not assessing the assumption. This confirms that further work is necessary on recommendations on assessing the PH assumption at the review level which could be shared with experts in the field and developed more formally into a guidance document as carried out by Gamble et al[92] for the statistical analysis plan guidance. Additionally, future work can also entail interviewing reviewers to find out why they didn't assess the methods to find out if it is due to lack of awareness of the methods in which case a guidance document would be the correct approach to take in tackling this.

Simmonds et al[43] mentioned that the PH methods can be biased if hazards are not proportional in 2011, however it is unclear how much notice has been taken given the level of assumption checking being performed. The results of this review are sufficient to confirm that further work is necessary to understand the impact of the PH assumption violation on RCTs and meta-analyses. This further work will allow us to identify whether further guidance is needed on why the PH assumption needs to be checked and what to do when the assumption of PH is invalid.

5 Digitising K-M curves

5.1 Previous work

Parmar et al[14] introduced an approach for extracting information from published K-M curves to approximate the log HR and variance from individual studies to enable the study to contribute to meta-analysis. This approach was then extended in 2002 by Williamson et al[22] to include patient numbers at risk in the approximation of the log HR and variance. Later in 2007, Tierney et al[39] presented a summary paper describing these approaches to a clinical audience including what to do if published K-M curves are available. These earlier approaches focus on manual extraction of data from published curves but a more efficient and accurate approach is to reconstruct approximate IPD from published curves using specific software as suggested by Guyot et al[29] in 2012. Guyot et al[29] explained that in the earlier work survival probabilities were extracted from published curves or the text to approximate aggregate data, but not all of the reported information was being used to identify the censoring pattern. The authors mention that in order to obtain consistent results and use all of the available published information they used "iterative numerical methods to solve the inverted K-M equations"[29]. Guyot et al[29] describe how they use digital software such as DigitiseIt (http://www.digitizeit.de/) to read in x and y coordinates of the K-M curves from the published graphs as well as using information on patients number at risk which is often published under the x-axis of the K-M graph to reconstruct the K-M data for each arm. If data on number of events is available then this can also be included in order to obtain even more accurate pseudo IPD. The paper primarily focused on the reproducibility and accuracy of the reconstructed statistics (see Section 2.1.3.7), by comparing the summary measures reported in the original publications to the ones obtained by the analysis of the reconstructed K-M data on six sets of K-M curves reconstructed by three observers each repeated twice. The reproducibility and accuracy of reconstructed results was excellent for median survival and probability of survival. However, the accuracy and reproducibility wasn't as good for HRs due to the following reasons: the HR is a weighted average of ratios along the entire risk period so the algorithm needs to make assumptions about the level of censoring within each segment and these assumptions have an impact on the relative weighting of different sections of the curve. The results for an example presented in the paper are as follows: the original publication presented the survival rate at year one to be 55, the median duration to be 14.9 and the HR (95% CI) to be 0.68 (0.52, 0.89). The reconstructed results were 56.1 for survival rate at year one, 14.9 for the median duration and 0.73 (0.57, 0.94) for the HR (95% CI). The results reflect the level of accuracy and reproducibility that is present.

In 2017, Wei and Royston published a paper[80] in the Stata journal on how to reconstruct TTE data from published K-M curves using Stata commands. This algorithm was based on Guyot et al[29] as an R function but with further improvements such as being able to use survival probabilities, survival percentages, failure probabilities or percentages as data input. The authors describe that once the data is extracted using software such as DigitiseIt, the Stata package ipdfc[80] can be used for fitting the Cox PH model and for assessing the PH assumption or even for performing secondary or alternative analyses such as RMST.

More recently, in 2016 and 2018 there has been published research where reconstructed TTE data has been used for performing analyses and for assessing the PH assumption. In 2016, Trinquart et al[56] used this approach on 54 trials in total and found evidence of non-proportionality in 18 (33%) trials. More recently in 2018, Liang et al[55] also reported using published K-M curves alongside numbers at risk and total number of events if available to reconstruct pseudo IPD for each arm. The authors found evidence of non-proportionality in seven (28%) out of 25 trials. The authors found that the published HRs and the HRs based on reconstructed data were very close to each other. These two are only a few of the published papers that are using this reconstruction technique to reanalyse the data and assess the PH assumption and ultimately perform alternative analyses. To my knowledge, this reconstruction technique has not been applied or recommended to be used to investigate the assumption of PH in the context of meta-analysis.

The aim of this chapter is to describe how the reconstruction technique is applied with results provided from digitising K-M curves from Chapter 4 to investigate non-PH in the meta-analysis setting. Later in this chapter, three examples are provided on digitising K-M curves to investigate non-PH in individual trials and within meta-analyses.

5.2 How to digitise K-M curves

The following steps are taken for extracting the data from published K-M curves:

1. Ensure that the survival data is in the correct format i.e. if presented as failure probabilities or failure percentages etc then these measures should be transformed arithmetically into survival probabilities.
2. The DigitizeIt (http://www.digitizeit.de/) software is used to extract data from graphical images of the published K-M curves which is achieved by clicking on individual points of the curve using a mouse, which are recorded as x and y values by the software. If possible the number of patients at risk for each arm should also be extracted, as the accuracy of the approximated TTE data can be improved by including this information[39].
3. The x and y values should then be saved as CSV files, one for each treatment group curve. The extracted data and the number at risk data which is normally included below the K-M plot should then be saved in a text file and imported into Stata. The text file should contain the data extracted from the x axis and y axis of each curve so time and survival as well as time and number at risk from the risk table.

4. Once the data is extracted, the Stata[69] package, ipdfc[80] is used for fitting the Cox PH model to the reconstructed data using the stcox command and then the PH assumption is tested using Schoenfeld residuals using the estat phtest command in Stata. The Stata package, ipdfc is based on the reconstruction algorithm which was written by Guyot et al[29] which is available in R[70]. Using the reconstruction approach, the PH assumption can be assessed even when only K-M curves are available.

5.3 Digitising K-M curves to investigate non-PH in individual trials: A worked example

For example 1 the following trial was used "A phase 3 trial of Bevacizumab in Ovarian Cancer" by Perren et al[7], published in 2011. This trial is the ICON7 trial which is a two arm RCT in advanced ovarian cancer. In total 1528 women were randomized to receive standard chemotherapy plus bevacizumab therapy or standard chemotherapy alone. The main outcomes of interest include PFS and OS. The total number of events for PFS included 464 for standard chemotherapy alone and 470 for standard chemotherapy plus bevacizumab therapy. Perren et al[7] concluded that standard chemotherapy plus bevacizumab therapy improves PFS compared to receiving standard chemotherapy alone. The authors found evidence of non-PH, so presented RMST estimates as an alternative analysis.

Figure 8: Published K-M plot for Example 1

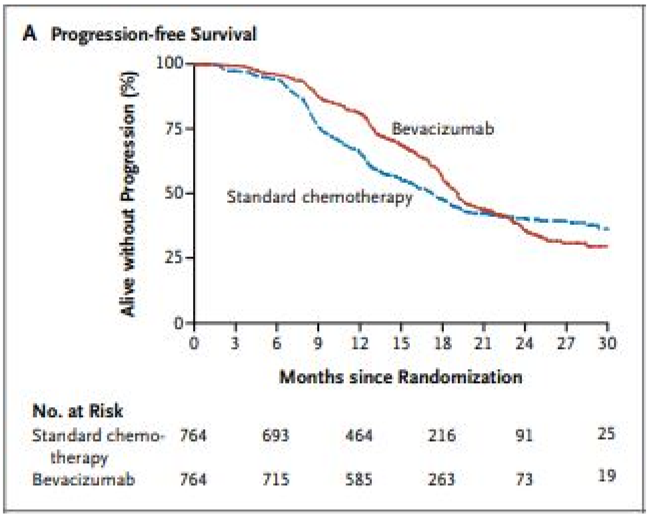
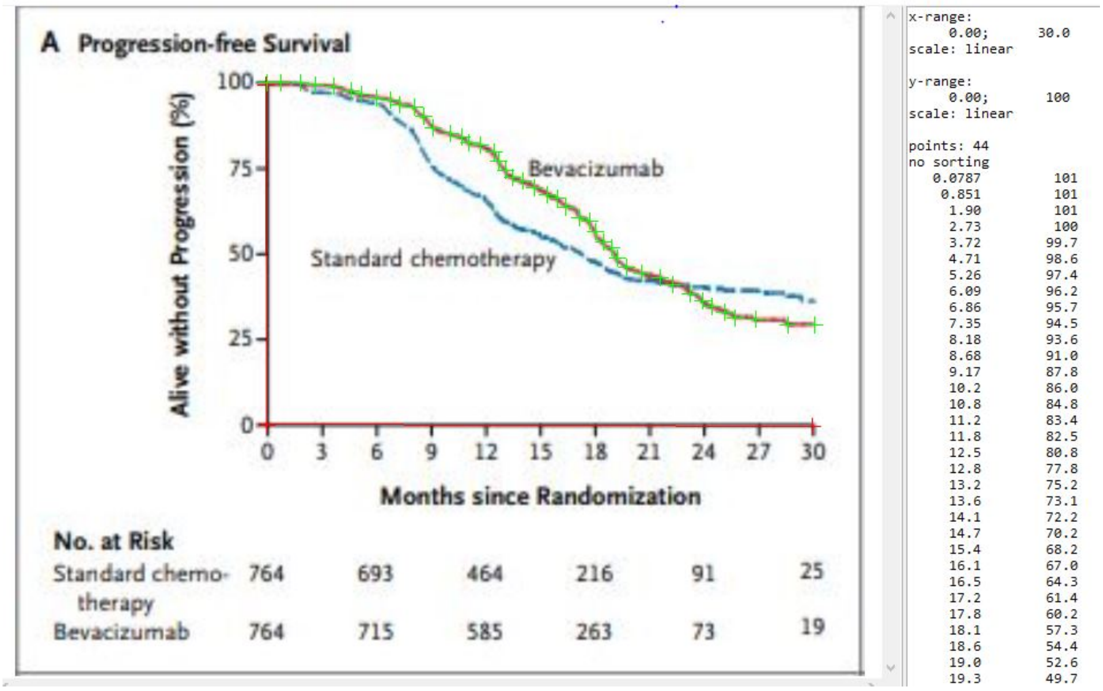


Figure 8 was uploaded into the DigitiseIt software and then the x and y values were extracted by clicking on individual points on the curve as seen in Figure 9. The right hand side of Figure 9 shows the x and y values that were extracted that are then saved as CSV files.

Figure 9: Screenshot of DigitiseIt software for Example 1


A Progression-free Survival
Alive without Progression (%)
100
75
50
25
0
0 3 6 9 12 15 18 21 24 27 30
Months since Randomization
Bevacizumab
Standard chemotherapy
No. at Risk
Standard chemo-therapy 764 693 464 216 91 25
Bevacizumab 764 715 585 263 73 19

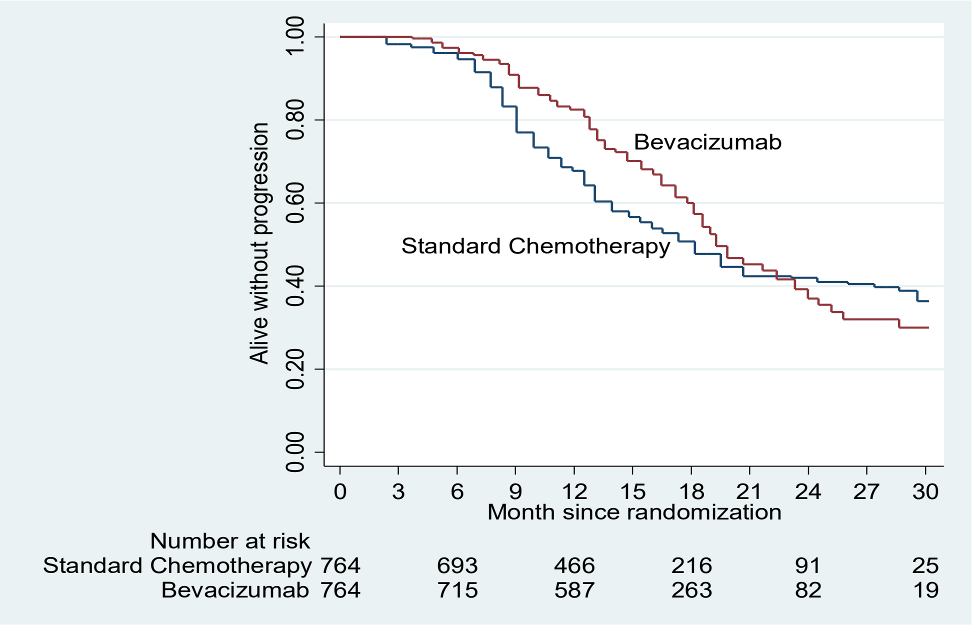
```
x-range:
    0.00;      30.0
scale: linear

y-range:
    0.00;       100
scale: linear

points: 44
no sorting
   0.0787        101
    0.851        101
     1.90        101
     2.73        100
     3.72       99.7
     4.71       98.6
     5.26       97.4
     6.09       96.2
     6.86       95.7
     7.35       94.5
     8.18       93.6
     8.68       91.0
     9.17       87.8
     10.2       86.0
     10.8       84.8
     11.2       83.4
     11.8       82.5
     12.5       80.8
     12.8       77.8
     13.2       75.2
     13.6       73.1
     14.1       72.2
     14.7       70.2
     15.4       68.2
     16.1       67.0
     16.5       64.3
     17.2       61.4
     17.8       60.2
     18.1       57.3
     18.6       54.4
     19.0       52.6
     19.3       49.7
```

The extracted data and the number at risk data which is included below the K-M plot were then saved in a text file and imported into Stata. The text file contained the data extracted from the x axis and y axis of each curve so time and survival as well as time and number at risk from the risk table. The reconstructed K-M plot can be seen in Figure 10.

Figure 10: Reconstructed K-M plot for Example 1



A cox PH model was fitted in Stata using the stcox command. The reconstructed HR (95% CI) is 0.82 (0.71 to 0.95), p=0.010 which was very similar to the results obtained from the original trial, 0.81 (0.70 to 0.94), p=0.004. There are slight differences present between the p-value from the reconstructed data compared to the original data, even though the differences are very small and the conclusions are unaltered based on the p-value. Potential reason why the p-value is more sensitive than the HR and CI is due to the assumptions around censoring. The reconstructed data is not the same as the original data but gives very similar KM-curves, however there are assumptions about where censoring occurs which could affect the p-value. From all the validation work that was carried out in the Guyot et al[29] paper, no validation of p-values was carried out.

The PH assumption was tested using the Schoenfeld residuals approach with p<0.001 suggesting that the PH assumption is invalid and in agreement with the trial authors results.

This example shows that when only aggregate data is available then how the K-M plot can be used to reconstruct pseudo-IPD and calculate the HR and associated 95% CI as well as how to investigate the PH assumption.

5.4 Digitising K-M curves to investigate non-PH in meta-analysis: Worked examples

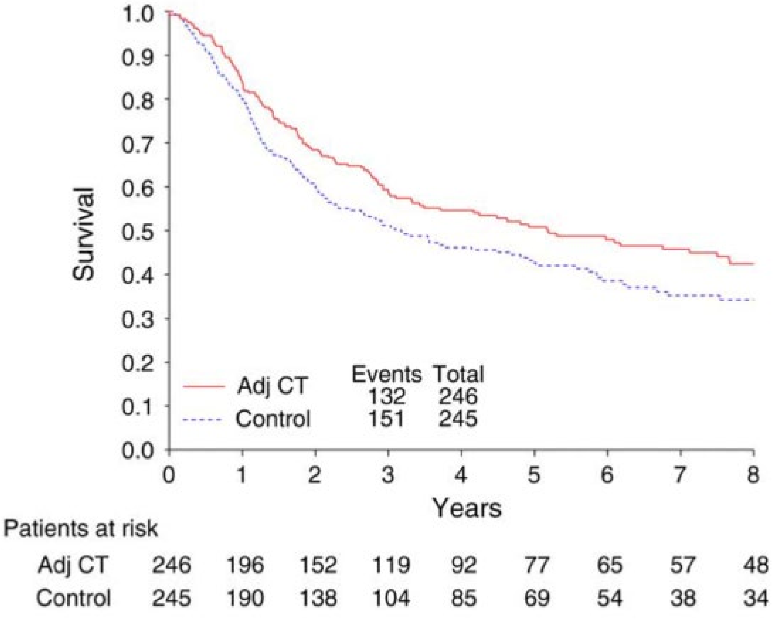
Currently the digitisation approach has been used mainly within individual trials to obtain reconstructed pseudo-IPD in order to perform survival analysis and more recently[55 56] as an approach for assessing the PH assumption. To check how robust the digitisation approach is within meta-analyses and to assess the PH assumption within the reviews, two worked examples are presented below. The main aim of performing this approach within meta-analyses is to estimate the prevalence of non-PH across the reviews but also as a method that consumers of reviews could use to assess the assumption when the HR is the treatment effect that is proposed.

Currently, methods available for assessing the PH assumption at the trial level can be used for assessing the PH assumption at the review level. It is important to note that the reviews did not present enough information to determine how the review level K-M curves were constructed so the K-M curves could be review level summary curves which may not necessarily be based on the most appropriate method.

5.4.1 Example 2

For example 2 the following review was used: “Adjuvant chemotherapy in invasive bladder cancer: a systematic review and meta-analysis of individual patient data” by Claire Vale[6], published in 2005. This IPD meta-analysis includes six RCTs with 491 patients comparing local treatment plus adjuvant chemotherapy to the same local treatment alone. The primary outcome was OS with survival analysis based on 283 events. The overall HR (95% CI) is 0.75 (0.60 to 0.96) representing a 25% relative decrease in the risk of death on local treatment plus adjuvant chemotherapy compared with that on control. The published K-M plot can be seen in Figure 11.

Figure 11: Published Kaplan-Meier plot for Example 2



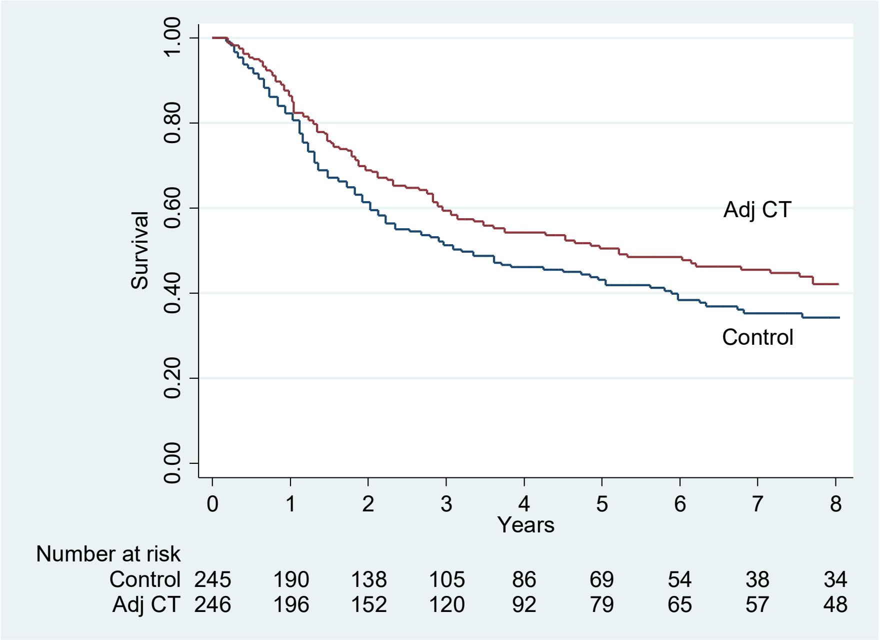
Patients at risk

Adj CT	246	196	152	119	92	77	65	57	48
Control	245	190	138	104	85	69	54	38	34

Fig. 2. Kaplan-Meier curve for survival (All trials).

Figure 11 was uploaded into the DigitiseIt software and then the x and y values were extracted by clicking on individual points on the curve like seen in Figure 9. The extracted data and the number at risk data which is included below the K-M plot were then saved in a text file and imported into Stata. The reconstructed K-M plot can be seen below in Figure 12.

Figure 12: Reconstructed K-M plot for Example 2



The survival analysis was carried out using stcox. The reconstructed HR (95% CI) is 0.78 (0.61 to 0.99) which was similar to the results obtained from the original review, 0.75 (0.60 to 0.96). These results demonstrated the accuracy of the approximated TTE data. The PH assumption was tested again using the Schoenfeld residuals approach. The p-value = 0.535 suggesting that the PH assumption was valid. The original review did not mention anything about testing the PH assumption within the review.

5.4.2 Example 3

In example 3 the following review was used: "Cisplatin- versus carboplatin-based chemotherapy in first-line treatment of advanced non-small-cell lung cancer: an individual patient data meta-analysis" by Ardizzoni et al[93], published in 2007. An IPD meta-analysis comparing carboplatin to cisplatin in first line treatment of advanced non-small cell lung cancer was performed. Nine RCTs were included with a total of 2968 patients. The primary outcome was OS with survival analysis based on 1298 events for cisplatin and 1316 events for carboplatin. Ardizzoni et al[94] concluded that the risk of death was higher with carboplatin compared to cisplatin, but the difference was not statistically significant. The authors did not mention or assess the PH assumption. The published K-M plot can be seen in Figure 13.

Figure 13: Published K-M plot for Example 3

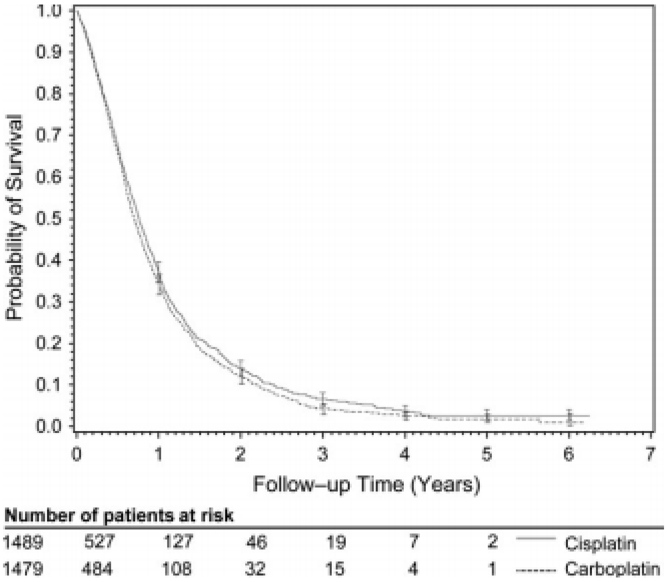
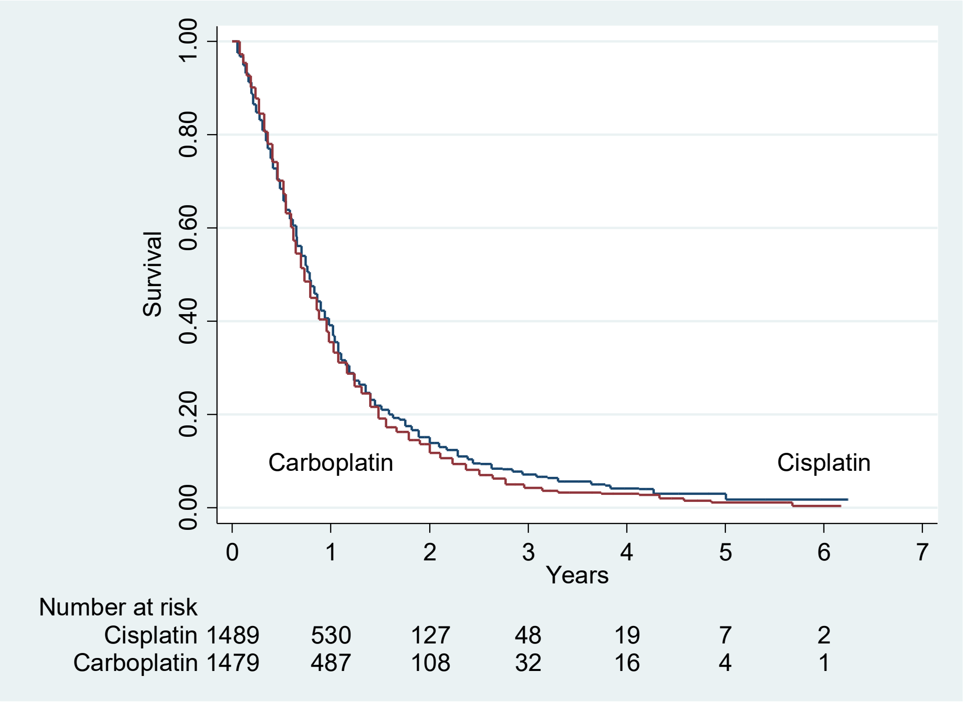


Figure 13 was uploaded into the DigitiseIt software and then the x and y values were extracted by clicking on individual points on the curve like seen in Figure 9. The extracted data and the number at risk data were saved in a text file and imported into Stata. Once the text file was imported into Stata it was important to state that the y values were probabilities and not percentages which is the default setting within the ipdfc[80] command. The reconstructed K-M plot can be seen below in Figure 14.

Figure 14: Reconstructed K-M plot for Example 3



Number at risk

	0	1	2	3	4	5	6
Cisplatin	1489	530	127	48	19	7	2
Carboplatin	1479	487	108	32	16	4	1

The survival analysis was carried out using stcox. The reconstructed HR (95% CI) is 1.06 (0.98 to 1.14) which was similar to the results obtained from the original review, 1.07 (0.99 to 1.15). These results demonstrated the accuracy of the approximated TTE data. The PH assumption was tested using the Schoenfeld residuals approach. The p-value=0.042 suggesting that the PH assumption was invalid. The original review did not mention testing the PH assumption within the review. The results suggest that the Cox PH model may not be the most appropriate method to use for analysing the survival data and perhaps a secondary analysis such as RMST could be performed which is not dependent on the PH assumption. This example shows why it is important to assess the PH assumption, although in this example a visual inspection of the K-M plot would be suffice to know that the survival curves cross so the PH assumption may be invalid.

5.5 Results from digitizing K-M curves

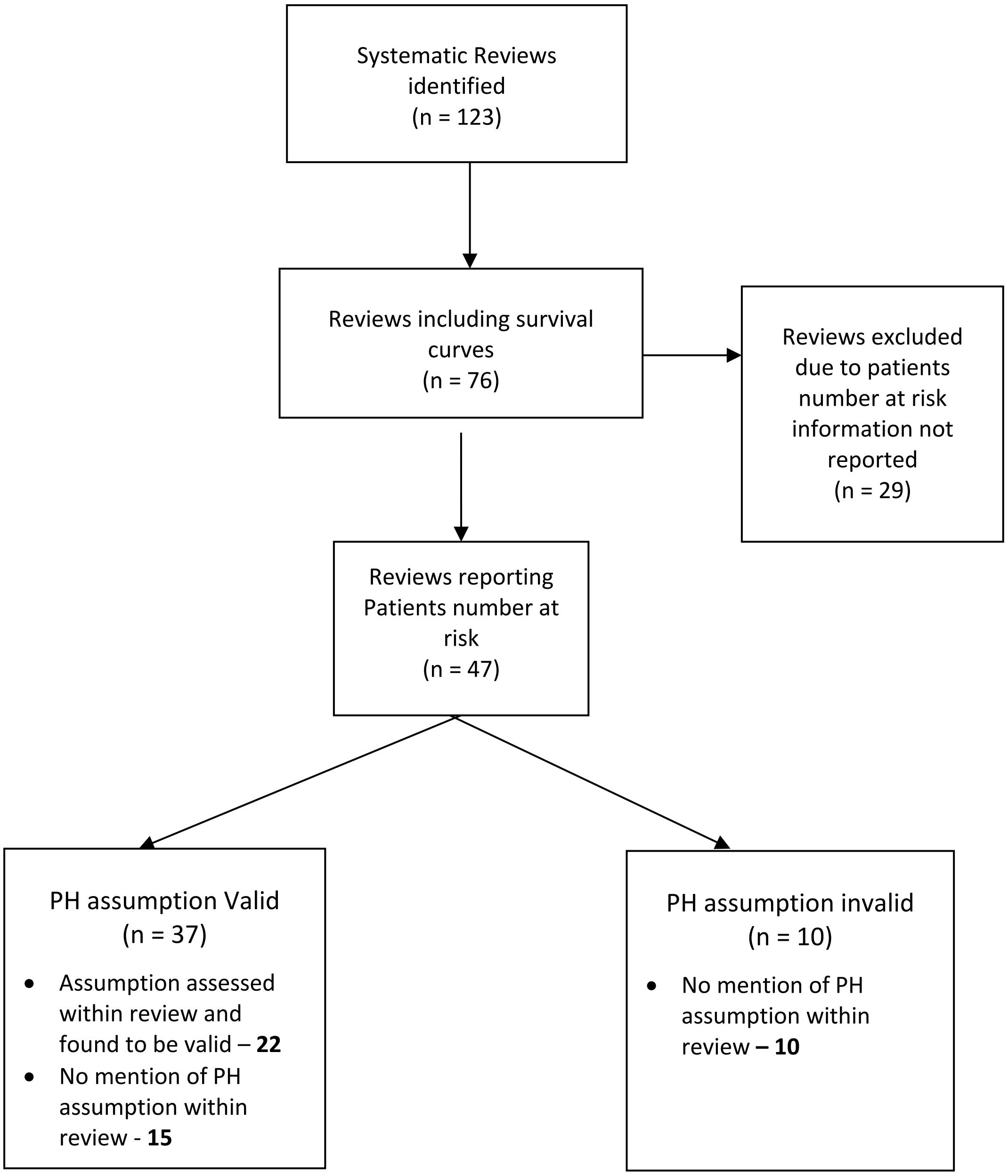
The K-M curves included in the systematic reviews from Chapter 4 were reconstructed in order to investigate the PH assumption. Only 76 out of 123 (62%) reviews reported survival curves in the systematic reviews. The method of survival curve calculation across all reviews was the K-M method, although this was not always clearly stated. The patient numbers at risk were reported in 47 (38%) of the reviews, hence it was only these 47 reviews which could be included in this digitisation work. The reviews did not include details on how the review level K-M curves were created. The majority (44 out of 47 (94%)) of the reviews included IPD with the remaining being aggregate data reviews. Two ways in which the IPD review level K-M curves could be created are as follows:

1. IPD would be lumped together from multiple trials and then the curves would be created. Using this approach could potentially dilute any trial specific issues about non-PH and the issues of heterogeneity in the patterns of effect as well as heterogeneity of the trial level curves will be lost. Additionally, using this approach goes against the spirit of meta-analysis as the data for the curve calculations will include patients from across trials rather than comparing patients within trials.

2. Trial level curves would be constructed and then pooled across the trials. This would be a sensible approach and has been described in the Parmar et al[14] paper and the Williamson et al[22] paper. They do describe this approach for aggregate data but the same approach could be used for IPD.

Using the digitising K-M data method presented by Guyot et al[29], the validity of the PH assumption was further investigated to understand whether the assumption was valid or not. The breakdown of the results are given in Figure 15. The results from this work identified that the PH assumption was valid for 37 (79%) reviews. From these 37 reviews, 22 review authors assessed the PH assumption within the review whilst a further 15 review authors did not mention the assumption within the review. From the 22 review authors who assessed the PH assumption, 20 authors had IPD whilst two review authors had access to IPD and aggregate data. All of the 15 review authors who did not mention the assumption had access to IPD. In total 10 reviews were identified where the PH assumption was invalid (21%). For all of these 10 reviews the review authors did not mention the assumption. All of the reviews had access to IPD apart from one review which only had aggregate data available.

Figure 15: PRISMA flow diagram for PH assumption checking


Systematic Reviews identified (n = 123)
Reviews including survival curves (n = 76)
Reviews excluded due to patients number at risk information not reported (n = 29)
Reviews reporting Patients number at risk (n = 47)
PH assumption Valid (n = 37)
Assumption assessed within review and found to be valid – 22
No mention of PH assumption within review - 15
PH assumption invalid (n = 10)
No mention of PH assumption within review – 10

5.6 Discussion

This chapter has demonstrated that the assumption of PH is violated in 21% of reviews examined. None of these reviews mentioned the PH assumption within the review, yet they presented a pooled HR, which is also a matter of concern.

The 76 reviews that presented published K-M curves did not present enough detail to understand how the review level K-M curves had been created. 47 reviews that presented numbers at risk were included in the reconstruction exercise with 44 (94%) of them being IPD reviews.

Using digitisation to reconstruct pseudo-IPD from published K-M curves to perform TTE analysis and assess the PH assumption is becoming a popular choice especially where IPD is not available as performed by Trinquart et al[56] and Liang et al[55] in recent years. Both Trinquart et al[56] and Liang et al[55] use the reconstruction technique to assess the PH assumption within the included RCTs. There are plenty of reviews where the reconstruction technique has been applied to individual trials. To my knowledge, however there are no reviews or papers where the reconstruction technique has been applied to meta-analyses in order to assess the PH assumption. As seen in the examples presented in Section 5.4, it is quite straightforward to apply the reconstruction technique to meta-analyses. The two examples included were IPD meta-analyses but there is no reason why the reconstruction technique can't be applied to aggregate data. Instead the example shown in Section 5.3 includes aggregate data, and shows how accurately the HR and 95% CI can be estimated when only the K-M plot and some summary measures are available.

Example 1 in this chapter also highlights that if performing a meta-analysis and only aggregate data from trials was available, then the DigitiseIt method followed by the ipdfc[80] command can be used to assess the PH assumption as part of the meta-analysis of TTE data.

When IPD is available it is not an issue as the reviewers can use the raw data to assess the assumption.

In situations where the PH assumption is invalid, the reconstructed pseudo-IPD can then be used to perform alternative analyses such as RMST that are not dependent on the PH assumption. However, it is important to note that these digitisation methods are limited by the quality of the curves that is available.

6 Review of proportional hazards within Single Technology Appraisals

6.1 Introduction

The NICE is an independent organization responsible for providing national guidance to the NHS in England on a range of clinical and public health issues, including the appraisal of new health technologies. The NICE STA process is specifically designed for the appraisal of a single health technology for a single indication, where most of the relevant evidence lies with one manufacturer or sponsor and typically covers new technologies shortly after UK market authorization is granted[27]. Within the STA process, the manufacturer or sponsor provides a written submission (alongside a decision-analytic model) that summarizes the estimate of the clinical effectiveness and cost-effectiveness of the technology. An external independent organisation (typically, an academic group) known as the ERG, provides a critique of the company's submission (the ERG report). Consultees, clinical specialists and patient representatives also provide additional information during the appraisal process.

Following a specification developed by NICE (the final scope), the NICE Appraisal Committee (AC) considers the company's submission, the ERG report and testimonies from experts and stakeholders in order to determine whether the technology represents a clinical- and cost-effective use of NHS resources. All stakeholders and the public have an opportunity to comment on the preliminary guidance issued by NICE in the form of an Appraisal Consultation Document (ACD), after which the AC meets again to produce the final guidance (Final Appraisal Determination [FAD]). The final guidance constitutes a legal obligation for National Health Service (NHS) providers in England and Wales to provide a technology that is approved within its licensed indication[27].

Oncology denotes a major disease area where TTE analysis is an important aspect of clinical management and motivates decision-making around treatment options. This is critical for STAs where majority of the outcomes within Oncology are OS and PFS. The most common method used to analyse TTE data is the Cox PH model as highlighted in Chapters 3 and 4. One issue that has repeatedly arisen within STAs is the incorrect use of the Cox PH model. In many instances, the PH assumption is not mentioned within STAs by the companies and often the ERG reports don't mention that the ERG have assessed the assumption themselves. When estimates of relative treatment effect are based on statistical models of TTE data, where the PH assumption is violated, both the results from the study and meta-analyses will be biased[35]. Where access to K-M curves and/or IPD is available from the individually included trials, it is possible for the ERG to assess the PH assumption for validity. Along with clinical-effectiveness analyses, cost-effectiveness analyses also rely on the TTE data to estimate the treatment effect, which is commonly assessed using PH models. Therefore, the implausibility of the PH assumption can impact upon decisions based upon cost-effectiveness analyses as well. Currently, it is still unknown what the impact of incorrectly assuming proportionality is.

6.1.1 NICE guidelines

The Decision Support Unit (DSU) which is funded by NICE have written a research and training resource to support their Technology Appraisal programme. The DSU have put together a TSD on survival analysis[83] which was published in 2011. In the document, they focus on survival analysis modelling methods with a strong focus on the PH assumption. The TSD mentions a few different approaches for assessing the PH assumption including visual inspection of Kaplan-Meier curves, using log-cumulative hazard plots as well as using AIC and Bayesian Information Criterion (BIC) statistics.

The TSD includes details of a review conducted on TTE methods used in NICE Technology Appraisals (TAs) which includes 21 TAs completed as of December 2009. The review summarises the modelling approaches that have been used and proposes a model selection process algorithm of how TTE analysis methods should be undertaken appropriately. The step-by-step process starts with using the log-cumulative hazard plots in order to assess which parametric model should be used as well as whether the PH assumption is valid. The next step explains what should be done if the log-cumulative hazard plot does or does not produce approximately straight lines.

An updated TSD was published in 2020 on "Flexible Methods for Survival Analysis"[84] which focuses on methods such as flexible parametric survival approaches, mixture models and landmark analysis which can be used on individual trials before focusing on extrapolating. The focus of this TSD is on methods not dependent on the PH assumption.

To my knowledge, this is the first review to assess the prevalence and issues surrounding the PH assumption within STAs. The main objective of this review is to assess the reporting and assessment of the PH assumption in the included STAs by both companies and the ERGs in terms of clinical- and cost-effectiveness.

6.2 Methods

6.2.1 Identification and selection of STAs for inclusion

In order to obtain a convenience sample of recently completed STAs, a list of Technology Appraisal guidance was obtained from the NICE website[27]. STAs that had been published on the NICE website between 1st April 2017 and 31st March 2018 were included. As oncology represents a major disease area where survival analysis typically drives decisions around treatment options, the review focused only on STAs in oncology. The company's submission of evidence, the ERG report (and any other documents relating to the company's initial submission of evidence and ERG's review of this initial submission) were all obtained from the NICE website.

Once the list of Technology Appraisal guidance had been screened for STAs in oncology and for those published between 1st April 2017 and 31st March 2018 the STAs were assessed for eligibility using the following criteria:

Inclusion criteria

- Included only if TTE outcomes present
- Included only if obtained the company's submission of evidence

Exclusion criteria

- No TTE outcome included in STA
- Multiple technology appraisals (MTAs)
- Company submission of evidence unavailable online
- If there is a reconsideration of the Cancer Drugs Fund as the company submission would then be unavailable.

6.2.2 Data extraction

Data extraction was performed using a database created in Microsoft Excel, as seen in Appendix 7. Data was extracted from the documents listed above. Any disagreements were resolved through discussion and by consulting CTS. Data was collected for each STA including: the company submitting evidence for appraisal, the ERG, the indication, the technology of interest, comparators, details regarding the use and assessment of the PH assumption within the STA for the outcomes of OS and PFS. OS and PFS were chosen as they both were the most commonly reported outcomes.

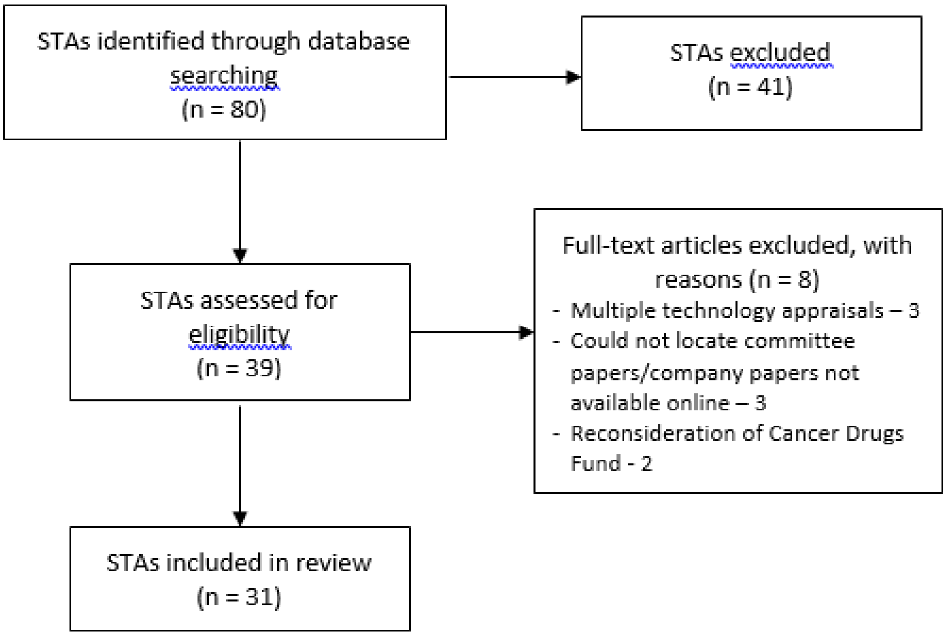
Data was extracted on clinical direct and indirect and cost-effectiveness evidence. Clinical direct evidence has been obtained directly from the pivotal trials. The indirect evidence is obtained using a mixture of direct and indirect evidence also known as NMA as described in Section 1.3.2. The cost-effectiveness evidence comes from the company submission who included published literature on the relevant patient population with the disease of interest.

6.3 Results

6.3.1 Review of STAs

The search for a list of Technology appraisals, where appraisals have been published on the NICE website between 1st April 2017 and 31st March 2018 yielded 80 Technology appraisals. After removing the Technology appraisals outside of oncology, the inclusion and exclusion criteria were applied to 39 Technology appraisals. A further eight Technology appraisals were excluded resulting in 31 STAs being eligible for inclusion in the review. The PRISMA diagram in Figure 16 describes the screening process, including reasons for exclusion at each screening stage. See Appendix 8 for a reference list of the 31 included reviews.

Figure 16: PRISMA flow diagram for STAs


STAs identified through database searching
(n = 80)
STAs excluded
(n = 41)
STAs assessed for eligibility
(n = 39)
Full-text articles excluded, with reasons (n = 8)
- Multiple technology appraisals – 3
- Could not locate committee papers/company papers not available online – 3
- Reconsideration of Cancer Drugs Fund - 2
STAs included in review
(n = 31)

6.3.2 STA results

In total, 27 out of 31 companies (87%) made an assumption of PH in their submission in either the clinical effectiveness, or cost-effectiveness sections or both. The results from the STA review of PH assumption checking are given below split by clinical effectiveness and cost-effectiveness. The clinical effectiveness section is further split by whether the results are from direct evidence or indirect evidence, i.e. network meta-analysis.

6.3.2.1 Clinical effectiveness – direct evidence

The results in Table 10 highlight that 23 out of 31 (74.2%) of STAs made an assumption of PH in the clinical effectiveness direct evidence section. Rather than choosing which outcome to present results for, it was agreed that both PFS and OS are both important survival outcomes so both should be reported in all the tables.

For all 23 STAs, the companies used methods that assumed PH for all PFS and OS outcomes. For both PFS and OS, 22 out of 23 STAs (96%) used the Cox PH model for the method of analysis. For PFS, from the 23 STAs assuming PH, only 15 (65%) STAs reported testing the PH assumption. Out of these 15 STAs, one company did not originally assess PH and only assessed after the ERG requested this at clarification stage. For another STA, there were two trials included in the direct evidence section and the company assessed PH for one trial but not the other. No additional details were given on why the company assessed PH for one trial and not the other. As the results in this review are by STA and not individual trial, this particular STA has been included as having tested the PH assumption but still reported separately in the table for clarity.

For PFS, out of the 15 companies who tested the assumption of PH, nine (60%) STAs reported that PH was violated, four (27%) STAs reported that PH was valid and two STAs did not report the results from the PH assumption testing. The most common methods used for testing the assumption included log cumulative hazard plots and Schoenfeld residuals. For OS, out of the 16 (70%) STAs that included an assessment of the PH assumption, 10 (63%) STAs reported that PH was violated, five (31%) STAs reported that PH was valid and for one STA the result was unknown. Out of the 10 STAs that reported PH was violated, for one STA, PH was found to be violated in one population but not in another population, so here it has been included in the PH violated category. Despite PH being violated, for all the STAs none of the companies made any changes to the analysis approach.

For those companies that did not assess PH, for the majority of the cases for both PFS and OS, the ERG also did not assess PH. Similarly, for cases where the company did assess, again the ERG did not assess PH for the majority of the cases for PFS (80%) and OS (69%).

Table 10: STA results from Clinical direct evidence

	STA n/N	Details (n/N (%))
Clinical direct		
How many STAs made an assumption of PH in the clinical effectiveness direct evidence section?	**23/31**	
What methods were used? I.e. cox PH or other or not reported?	**PFS (n=23)**	Cox PH model – 22/23 (96%) Log rank test, HRs presented but methods not stated – 1/23 (4%)
	OS (n=23)	Cox PH model – 21/23 (91%) Log rank test, HRs presented but methods not stated – 1/23 (4%) HRs presented but methods not stated – 1/23 (4%)
Of these, how many tested the assumption of PH?	**PFS (n=23)**	No – 8/23 (35%) For 1 trial yes, for the other trial no – 1/23 (4%) Yes – 14/23 (61%)
	OS (n=23)	No – 7/23 (30%) Yes – 16/23 (70%)
For those that tested, how many found that PH was violated?	**PFS (n=15)**	PH violated - 9/15 (60%) PH holds - 4/15 (27%) Result unknown - 2/15 (13%)
	OS (n=16)	PH violated - 10/16 (63%) PH holds - 5/16 (31%) Result unknown - 1/16 (6%)
For those who found that PH was violated, what did the company then do?	**PFS (n=9)**	No changes to analysis approach – 9/9 (100%)
	OS (n=10)	No changes to analysis approach – 10/10 (100%)
For those that did not test, did the ERG assess PH and if so, what were the results?	**PFS (n=8)**	Did not assess – 5/8 (63%) Yes (PH valid) – 2/8 (25%) Yes (PH violated) – 1/8 (12%)
	OS (n=7)	Did not assess – 5/7 (71%) Yes (PH violated) – 2/7 (29%)
For those that did test, did the ERG assess PH and if so, what were the results?	**PFS (n=15)**	Did not assess – 12/15 (80%) Yes – 1/15 (7%) (PH valid, in agreement with company) Yes – 2/15 (13%) (PH not valid, in agreement with company)
	OS (n=16)	Did not assess – 11/16 (69%) Yes – 2/16 (13%) (PH valid, in agreement with company) Yes – 3/16 (19%) (PH not valid, in agreement with company)

6.3.2.2 Clinical effectiveness – indirect evidence

Table 11: STA results from Clinical Indirect evidence

	n/N	n/N (%)
Clinical indirect		
How many STAs made an assumption of PH in the clinical effectiveness indirect evidence section?	**13/31**	-
Of these, how many tested the assumption of PH?	**PFS (n=13)**	No – 4/13 (31%) Yes – 9/13 (69%)
	OS (n=13)	No – 3/13 (23%) Yes – 10/13 (77%)
For those that tested, how many found that PH was violated?	**PFS (n=9)**	PH violated - 8/9 (89%) PH holds - 1/9 (11%)
	OS (n=10)	PH violated - 8/10 (80%) PH holds - 2/10 (20%)
For those who found that PH was violated, what did the company then do?	**PFS (n=8)**	Re-do NMA using fractional polynomials after ERG requested them to test PH at clarification – 2/8 (25%) Declared PH didn't hold, so ITC is unreliable – 2/8 (25%) Used RMST – 1/8 (12.5%) Used Bayesian fractional polynomial methods – 1/8 (12.5%) Used Bayesian parametric methods – 1/8 (12.5%) No change to analysis approach – 1/8 (12/5%)
	OS (n=8)	Re-do NMA using fractional polynomials after ERG requested them to test PH at clarification – 2/8 (25%) Declared PH didn't hold, so ITC is unreliable – 2/8 (25%) Used RMST – 1/8 (12.5%) Used Bayesian fractional polynomial methods – 1/8 (12.5%) Used Bayesian parametric methods – 1/8 (12.5%) No change to analysis approach – 1/8 (12.5%)
For those that did not test, did the ERG assess PH and if so, what were the results?	**PFS (n=4)**	Did not assess – 4/4 (100%)
	OS (n=3)	Did not assess – 3/3 (100%)
For those that did test, did the ERG assess PH and if so, what were the results?	**PFS (n=9)**	Did not assess – 7/9 (78%) Yes – ERG concludes PH is violated, in agreement with company – 2/9 (22%)
	OS (n=10)	Did not assess - 6/10 (60%) Yes – ERG concludes PH is violated whereas the company concluded PH is valid – 1/10 (10%) Yes – ERG concludes PH is violated, in agreement with company – 3/10 (30%)

The results in Table 11 highlight that only 13 out of 31 (42%) STAs made an assumption of PH in the clinical effectiveness indirect evidence section. For all 13 STAs, the companies used methods that assumed PH for PFS and OS outcomes.

For PFS, from the 13 STAs assuming PH, nine (69%) STAs reported testing the PH assumption. Out of the nine STAs, two companies did not originally assess PH but only assessed after the ERG requested at the clarification stage. For OS, 10 (77%) STAs reported testing the PH assumption and similarly to PFS, two companies only assessed the PH assumption after the ERG requested it during the clarification stage. After assessing the PH assumption, PH was found to be violated in the majority of the STAs for both PFS and OS.

Table 11 highlights the different approaches taken by the company due to the PH assumption being violated. Again, similarly to the results in Table 10, in majority of the cases the ERG did not assess the PH assumption in both situations where the company did assess PH and where the company did not assess PH.

6.3.2.3 Cost-effectiveness evidence

For the cost-effectiveness results, Table 12 suggests that 20 out of 31 (65%) STAs assume PH. However, only 18 STAs used methods that assumed PH for PFS and only 17 STAs assumed PH for OS. For PFS, from the 18 STAs assuming PH, 13 (72%) STAs reported testing the PH assumption. Out of the 13 STAs, two companies did not originally assess PH but only assessed after the ERG requested at the clarification stage. For OS, 15 out of 17 (88%) STAs reported testing the PH assumption and similarly to PFS, two companies only assessed the PH assumption after the ERG requested it during the clarification stage. For both PFS and OS, one company tested the PH assumption for the trials included in the direct evidence, however did not test the PH assumption for the trials included in the indirect treatment comparisons. For this review, as the company have tested the PH assumption for some trials, we have included the STA within the "tested PH" category.

After assessing the PH assumption, PH was found to be violated in around 70% of the STAs for both PFS and OS. Table 12 highlights the different approaches taken by the companies due to the PH assumption being violated. Again, similar to the results in Table 10 and Table 11, in majority of the cases the ERG did not assess the PH assumption in both situations where the company did assess PH and where the company did not assess PH.

Table 12: STA results from Cost-effectiveness data

	n/N	**n/N (%)**
Cost effectiveness		
How many companies made an assumption of PH in the cost effectiveness section (i.e. direct or indirect evidence)?	**20/31**	-
Of these, how many tested the assumption of PH?	**PFS (n=18)**	Did not assess– 4/18 (22%) After ERG requested at clarification – 2/18 (11%) Tested for HRs that came from trials, but not for HRs that came from the NMA – 1/18 (5%) Yes – 11/18 (61%)
	OS (n=17)	Did not assess - 1/17 (6%) After ERG requested at clarification – 2/17 (12%) Tested for HRs that came from trials, but not for HRs that came from the NMA – 1/17 (6%) Yes – 13/17 (76%)
For those that tested, how many found that PH was violated?	**PFS (n=14)**	PH violated - 10/14 (71%) PH holds - 4/14 (29%)
	OS (n=16)	PH violated - 11/16 (69%) PH holds - 5/16 (31%)
For those who found that PH was violated, what did the company then do?	**PFS (n=10)**	Redo NMA using fractional polynomials and modelling after ERG requested them to test PH at clarification – 2/10 (20%) Used HRs in model base-case – 1/10 (10%) Fit independent fitted curves – 4/10 (40%) Stratified Parametric models – 2/10 (20%) Fractional polynomials – 1/10 (10%)
	OS (n=11)	Redo NMA using fractional polynomials and modelling after ERG ask them to test PH at clarification – 2/11 (18%) Used a delayed exponential fit which was found to satisfy PH assumption – 1/11 (9%) Used HRs in model base-case – 2/11 (18%) Fit independent fitted curves – 3/11 (27%) Stratified Parametric models – 1/11 (9%) Fractional polynomials – 2/11 (18%)
For those that did not test, did the ERG assess PH and if so, what were the results?	**PFS (n=4)**	Did not assess – 4 STAs
	OS (n=1)	Did not assess – 1 STA
For those that did test, did the ERG assess PH and if so, what were the results?	**PFS (n=14)**	Did not assess – 13 STAs Yes – ERG concludes PH is violated, in agreement with company – 1 STA
	OS (n=16)	Did not assess - 13 STAs Yes - ERG concludes PH is violated whereas the company concluded PH is valid – 1 STA Yes – ERG concludes PH is violated, in agreement with company – 2 STAs

6.4 Discussion

6.4.1 Summary of main results

This review considers 31 STAs published between 1st April 2017 and 31st March 2018, in Oncology.

STA results from the clinical direct evidence suggest that 74% (23 out of 31) of the STAs use methods that assume PH. From these, 15 (65%) STAs tested PH for PFS and 16 (70%) tested for OS. Around 60% of companies found PH to be violated for both outcomes but no changes were made to the analysis approach. The ERG in majority of cases did not assess the PH assumption.

Similarly, for the clinical indirect evidence, 42% (13 out of 31) of the STAs used methods that assume PH. From these, nine (69%) STAs tested PH for PFS and 10 (77%) STAs tested PH for OS. Eight of the companies in both cases found PH to be violated. Alternative approaches were used for the analyses in all but one STA. The ERG in majority of cases did not assess the PH assumption.

For the cost-effectiveness evidence, 65% (20 out of 31) of the STAs used methods that assume PH. From these 14 out of 20 (70%) STAs tested PH for PFS and 16 out of 20 (80%) STAs tested PH for OS. For both outcomes, around 70% of the companies tested PH and found it to be invalid. For all these cases, the companies performed alternative analyses that did not assume PH. The ERG in majority of cases did not assess the PH assumption.

In summary, findings of this review demonstrate that around 70% of the companies are testing the PH assumption in STAs. However, in the majority of the cases, the ERGs are not assessing the PH assumption. It is vital for the ERG to be assessing to see if the assumption is valid and if they agree with the company's decision as there have been cases in the review where the ERG have not agreed with the company's conclusion. This is crucial for the clinical direct evidence where for all STAs where PH was found to be violated, the companies did not change the analysis approach. The companies chose to ignore the violation of the PH assumption and in majority of the cases the ERGs also chose to ignore the violation of the PH assumption and what the companies did. For clinical indirect evidence and cost-effectiveness evidence, in majority of the cases, the companies used alternative methods not dependent on the PH assumption to re-analyse the data.

This review highlighted that it is clear that the companies and ERGs are aware of methods for assessing the PH assumption but still choose to ignore the assumption in many cases. It is unclear whether this is happening as they do not think violation of the assumption could impact the results or whether some authors genuinely do not know how to check the assumption. Hence, further work is needed to understand the impact of the violation of the assumption and also understanding why companies and ERGs do not always assess the assumption.

6.4.2 Strengths and limitations

At the time of writing, this systematic review is believed to be the first review which systematically considers the reporting of TTE data and the PH assumption within STAs. A previous review has assessed new oncology drug approvals from ten HTA agencies, whilst we have only focused on NICE.

Due to time constraints, only STAs in Oncology were included and STAs published between 1st April 2017 and 31st March 2018. Arguably, it would have been informative to consider all disease areas and include data over several years rather than 12 months. However, it was considered that the disease area is unlikely to impact the reporting of the PH assumption and reviewing STAs published over 12 months should provide an adequate representation of current practice.

6.4.3 Comparison to previous work

Monnickendam et al[95] conducted a review on whether Health Technology Assessment (HTA) agencies routinely assessed PH, what methods were used when non-PH was detected in Oncology trials and how often the RMST was used as a measure of treatment benefit. They reviewed methodological guidelines from ten HTA agencies in eight countries and then to understand current practice reviewed new Oncology drugs published on the US Food and Drug Administration (FDA) website between 1st January 2014 to 31st December 2016. They included 52 Oncology drugs and 22 were from NICE. The PH assumption was assessed in 15 out of 22 (68%) NICE appraisals, which is in line with results found in our review. The authors state that formal statistical methods were used for assessing the assumption but that the methods varied in terms of type of test used. No further details were given in the paper on what methods were used or what the outcome of the assessment was.

The authors mention that RMST was used most often by NICE. They used it in a number of assessments for validating estimates of extrapolated mean survival in cost-utility analyses. The authors did not identify any cases where it was used directly for the primary cost-effectiveness analyses. It was used in 10 out of 22 HTAs (45%) whilst for our review, RMST was only included in one STA in an indirect treatment comparison to address non-PH. The authors say that for one STA, RMST was used as a treatment effect measure to address issues of non-PH in an indirect treatment comparison and in one case, RMST was used in the manufacturers submission to demonstrate the impact of non-PH and provide evidence of the treatment effect when median difference was not statistically significant.

Previous research that has been published has focused on the reporting of TTE outcomes including reporting of the PH assumption. Further details on Batson *et al*[35], Altman *et al*[30] and Abraira *et al*[31] have been given in Chapters 1, 3 and 4. All three reviews highlight the poor reporting of the PH assumption within RCTs. In comparison to those reviews, the PH assumption is not being assessed in all cases of direct evidence by the companies as the direct evidence is coming from other RCTs so they are only presenting the results. However, for the indirect evidence, where the PH assumption is violated, 77.5% (7 out of 8 STAs) of companies did perform alternative analyses or acknowledged that as the PH assumption doesn't hold, the indirect treatment comparison results are unreliable. The companies changed the methods as they are performing the NMA so have access to the data.

6.4.4 Concluding remarks

In conclusion, in line with previous work conducted in this area, the current review has shown concerns regarding the routine testing of the PH assumption. For cost-effectiveness analyses, it is increasingly important that appropriate methods and techniques are used to assess the PH assumption and analyse the data using appropriate methods to ensure treatments are valued appropriately.

Although, around 70% of companies are assessing PH but still ignoring the assumption where it is violated , there are still 30% of companies who are not checking the assumption. Results of this review are surprising given that the NICE guidelines recommend testing the PH assumption and have published the TSD[83] with details on how to assess the PH assumption. Hence, why it is then unclear on why so many companies and ERGs are not following the recommendations in the guidelines. An updated TSD[84] has been published in 2020 that focuses on methods not dependent on the PH assumption which could be an alternative to methods dependent on the PH assumption. However, it is important to note that the results from this review on how often the PH assumption is assessed are higher than what has been found in Chapters 3 and 4 and in the general medical literature as discussed in Chapter 1, Section 1.4.2. Although, it is important to note that there is no direct comparison between

STA evidence syntheses where ITCs are being conducted and IPD is available and traditional systematic reviews where IPD is not always available and meta-analysis may be conducted.

Findings of this review further highlight the greater need for an understanding of the impact of non-PH, on overall results and conclusions. Currently, there is advice and recommendations on testing the PH assumption and alternative methods to use if the assumption is invalid. However, there is no advice on how the assumption being invalid impacts the results. Therefore, there is an urgent need for the development of reporting standards for TTE analysis and in particular on the impact of non-PH on results and conclusions.

7 Simulation Study to assess the impact of non-PH in meta-analysis

7.1 Introduction

It is clear from preceding chapters and previous research that the analysis of TTE outcomes using methods that assume PH continue to be conducted in practice, despite a lack of testing of the assumption, or in some cases ignoring the assumption. Little is known about the impact of incorporating trial results that have inappropriately assumed PH, into meta-analysis. In this chapter, a novel simulation study is presented to explore the performance of different methods for meta-analysis of TTE data under various scenarios, including whether PH holds or not. To begin with, a brief recap is given below of the methods available for analysing TTE data when PH does not hold. The literature is also reviewed, comparing two of the summary measures used to pool results for survival data, the HR, and the RMST difference or ratio as well as summarising previously conducted simulation studies that have specifically looked at the performance of methods for meta-analysis of TTE data. This chapter then goes on to describe the methods and report the results of the novel simulation study that was conducted.

7.1.1 Summary treatment effects: the Restricted Mean Survival Time

A summary treatment effect measure that is increasingly being used as an alternative to the HR or as a secondary analysis measure is the RMST as described in Chapter 2, Section 2.1.9. The RMST is a good summary treatment effect measure which is not reliant on the PH assumption, and can therefore be used as a metric to compare performance of different methods.

In 2011, Royston and Parmar[60] described three approaches for estimating the RMST: using pseudo-values, flexible parametric survival models, and integrating difference of survival functions. They then used three RCTs all from advanced cancer, to demonstrate how simple it is to estimate the RMST, and then compared the RMST results across the three approaches. The authors recommend using either flexible parametric models or pseudo-values when estimating the RMST as they give appropriate estimates of treatment effects under PH or non-PH and estimates of RMST can easily be obtained from them. However, they also state that the choice of t* (fixed specified time point) is important and hence why it should be pre-specified where possible. The HRs are also mentioned but there are no conclusions drawn on whether the HRs and RMST results are similar or not.

Trinquart et al[56] conducted a review in 2016 comparing treatment effects measured by the HR and the ratio of RMST in oncology RCTs. They included studies published between July and December 2014 in three general medical journals and two oncology journals. They found 54 trials that were eligible to be included in the analysis. In five (9%) trials, the RCT authors reported that the PH assumption was checked and evidence of non-PH was found by Trinquart et al[56] in 13 (24%) other trials. For each trial, the authors reconstructed IPD for each treatment group from published K-M curves, an approach suggested by Guyot et al[29]. Using the reconstructed IPD, the HR and associated variance were estimated by fitting a Cox PH model. The RMST was calculated using the integration of the survival function approach. The authors state that the HR and ratio of RMST are two different ways of quantifying the difference between two survival curves and do not have the same meaning. The authors aimed to assess if one was systematically further from the null than the other. Trinquart et al[56] explain that there was agreement on the direction of the treatment effect between the difference of RMST and HR in all but four trials. The authors mention that in 41 (76%) of the

54 trials, the HR demonstrated a larger treatment effect for the experimental arm compared to the ratio of RMST and the difference was statistically significant in 20 (37%) trials. To conclude, Trinquart et al[56] say that their findings suggest that the treatment effects are "systematically more beneficial when measured with HRs rather than RMST, whether the PH assumption holds or not." Therefore, they recommend that RMST-based measures of treatment effects are reported in any trial with TTE outcomes. It is important to note that although the HR and ratio of RMST are being compared they are not identical effect measures, with the HR measuring the risk of having an event in one group compared to another whilst the RMST difference measures whether patients in one group have X number of years gained or lost in life expectancy from 0 to t^*compared to patients in the other group.

Liang et al[55] conducted a systematic review in 2018 where they included all phase II and III RCTs with TTE outcomes, which included parallel group RCTs of immune checkpoint inhibitors. 25 trials were included in total that were all published between 2010 and 2018. Liang et al[55] report that evidence of non-PH was found in seven (28%) trials, although all the trials reported a HR. The authors used the K-M curves to obtain the time-dependent probability of overall survival. The K-M data was reconstructed for each arm using the numbers at risk and total number of events, where available. The authors estimated the ratio of RMST and difference in RMST between the treatment groups. Liang et al[55] state that "the ratio of RMST was transformed so that a ratio of RMST less than 1 indicated superiority of the experimental treatment, as would be the case for an estimate of HR in a trial where proportional hazards were satisfied." The authors assessed the PH assumption using a treatment-time interaction term in a time-dependent Cox model. The results illustrate that there was agreement on the direction of the treatment effect between the ratio of RMST and the HR for all trials. In 23 (92%) of the trials, the HR and ratio of RMST were in favour of the experimental arm. However, as seen in the research papers by Rulli et al[36] and Trinquart et al[56], the HR systematically overestimated the treatment effect compared to the ratio of RMST in all of these 23 trials. These 23 trials included trials where PH was valid and some where it was violated. The authors conclude by saying that failure to illustrate that the PH assumption is violated may actually showcase the lack of power to detect a violation rather than confirming that the hazards are valid. Liang et al[55] state that for immune checkpoint inhibitors very little or no difference is seen in the survival curves for the initial period and then a late separation occurs. In such a scenario, the PH assumption is not valid, hence the HR is difficult to interpret and in such cases the ratio of RMST or difference in RMST are valid methods for summarising the treatment effect.

In Rulli et al[36], the authors used nine of the superiority trials where the PH assumption was violated and compared the HR and RMST (ratio and difference). The results expressed that for all of these studies, there was agreement on the direction of the treatment effect between the RMST difference and HR for all the trials. Rulli et al[36] explained that similar to Trinquart et al[56], they also found that the HR systematically overestimates the treatment effect compared to the RMST ratio.

7.1.2 Previous simulation studies for meta-analysis of TTE outcomes

Simulation studies[96-98] are datasets that are created by pseudo-random sampling from a known probability distribution. Simulation studies are used to obtain empirical results about the performance of statistical methods in various scenarios. This is important as it is not always possible to obtain data that can be used to evaluate methods especially as the data can be messy or could be based on various assumptions and settings. Additionally, with real data the model parameters and functional form are unknown, whereas as simulated data is artificial the user is in control of the model parameters and functional form so nothing is left unknown. Therefore, simulation studies can be used to assess situations where commonly used methods may be based on incorrect assumptions, or used to evaluate new methods or competing methods[98].

In 2015, Wei et al[66] also conducted a simulation study where they compared the three methods used in the Royston and Parmar[60] paper in 2011, namely pseudo-values, flexible parametric model and integrated difference of survival functions. The authors start by applying the three methods to two IPD meta-analyses which were originally performed by the Medical Research Council Clinical Trials Unit (MRC CTU) on behalf of collaborative groups. The results from the two meta-analyses highlight that the conclusions from the RMST analysis are similar to the meta-analysis of HRs. The authors also state that although most of the trials included in the two meta-analyses provided no evidence of non-PH, 26% and 17% of the weight in the meta-analyses came from trials with non-PH. It is important to note that if the meta-analyses were dominated by trials with non-PH, then it is uncertain how reliable and informative such meta-analyses based on HRs could be. The authors state that they consider "the difference in RMST as a safer measure because it is free of the PH assumption."

Wei et al[66] explain that to their knowledge no comparison of the three estimation methods has taken place before and therefore they will conduct a simulation study to compare the three methods in terms of bias, mean square error and coverage of the difference in RMST. The Weibull distribution is used to simulate both the time to the event of interest and time to censoring. The study authors simulate 1000 survival datasets for each scenario with 16 scenarios in total. Wei et al[66] vary censoring (low and high), sample size (250 and 500 observations) and t^*(3, 5 and 10 years of follow-up). The results for the two non-parametric methods (pseudo-values and integrated difference of survival functions) produce similar results in terms of bias and mean square errors. This could be as both methods use a K–M estimate for the survival function. This leads to similar but not identical coverage probabilities. For all three methods, the coverage probabilities are close to their nominal value of 95%. There is no clear indication of one method being better than the other in terms of the coverage. However, in the flexible parametric survival model, mean square errors are smaller than the other two methods. Wei et al[66] explain that this could be as the flexible parametric model is able to correctly specify the survival function when the survival time follows a Weibull distribution. The authors explain that the non-parametric methods do not assume any parametric distribution, so the mean square errors are inflated. The study authors are well aware that the survival time does not always follow a Weibull distribution and that the mean square errors of the difference in RMST from the flexible parametric method are not always smaller compared with that from the other two estimation methods. Wei et al[66] also mention that a further advantage of estimating the RMST by a flexible parametric model is that the RMST can be predicted beyond the actual follow-up time, which allows all of the trial data in a meta-analysis to be included even when some trials have follow-up less than t^*. This is appealing in a meta-analysis context since trials typically have different lengths of follow-up. The study authors conclude by saying that the difference in RMST is a useful effect measure in a meta-analysis since it avoids the PH assumption. Wei et al[66] highlight that the measure is interpretable and helpful in situations where the treatment effects may change with time. The authors state that recent developments in data reconstruction techniques[29] including the Wei and Royston[80] Stata journal in 2017 on reconstructing TTE data from published K-M curves enables the extension of RMST meta-analysis for aggregate data.

7.1.3 Motivation of simulation study

The summary of treatment effects compares HRs to the ratio of RMST or difference in RMST and suggests that although the HR and ratio of RMST are two different measures, both are comparable to one another. In cases where the PH assumption is invalid, the HR is then dependent on time and thus a single HR summary is inappropriate. However, the summary of treatment effects in Section 7.1.1 found that the HR systematically overestimates the treatment effect, but it is important to note that in those reviews the PH was violated in many studies or poorly reported. However, the authors suggest that the HR tends to overestimate treatment effects when PH is violated, hence making the RMST a more conservative treatment effect measure.

The summary of previous simulation studies for meta-analysis of TTE outcomes highlights that further work is required to investigate the impact of the violation of the PH assumption on meta-analysis as currently only one research paper[66] has been identified which focuses on TTE outcomes within a meta-analytic setting where the focus is on the PH assumption. Wei et al[66] compared three approaches for estimating the RMST using a simulation study. This is the only simulation study conducted so far on estimating RMST, with all other results coming from real-life data[36 55 56 60]. The summary of previous simulation studies highlights that although there has been research conducted comparing the HR obtained from fitting a Cox model to the RMST, there hasn't been any research carried out comparing the results from the Cox, Weibull PH, AFT and flexible parametric models to one another using real or simulated data.

A simulation study is needed to evaluate different methods in cases where PH is valid, and in cases where PH is invalid, to compare different methods of analysing TTE data under different scenarios within a meta-analytic setting. Such a simulation study is described during this chapter.

7.2 Methods

A simulation study was undertaken to compare the performance of different modelling approaches for meta-analysis of TTE data under conditions, where the validity of the assumption of PH varies.

7.2.1 Data Simulation Methods and Data-generating mechanisms

Data were generated by assuming a Weibull distribution for survival times, and thus a hazard function for the *ith* individual is as follows:

$$\mathrm{h_i(t)} = \lambda_e \gamma \mathrm{t}^{\gamma-1} \exp(\beta_0 \mathrm{treat}), \tag{13}$$

where $\lambda_e > 0$ is the scale parameter, $\gamma > 0$ is the shape parameter, and β_0 is the log hazard ratio for treatment when time t=1 (i.e. when log(t)=0). Here treat is a binary variable denoting treatment group (with 0 representing a control, and 1 an experimental treatment)

Twelve data generating mechanisms (DGMs) were considered, where DGM is a unique scenario each with different characteristics in which the different modelling approaches will be compared[98]. As the main focus for this work is within evidence synthesis, for all DGMs, data are simulated for five studies, each containing $\mathrm{n_{obs}}$= 500 patients, designed to represent an IPD meta-dataset containing five trials with TTE outcome. The number of trials and patients were selected to be similar to the characteristics of our review described in Chapter 3, for which the median number of studies included was five and median number of patients was 2579 for each review. For the length of follow-up, the median length of follow-up in the studies in the review was 4.5 years so for the simulations a five year length of follow-up was used. In total, 1000 meta-datasets were generated for each DGM.

The covariate $treat$ was binary and generated from a Bernoulli(0.5) distribution – representing simple randomisation with an equal allocation ratio. Simulated survival times were generated using the Simsurv package version 0.2.3[99] in the R software package, version

3.4.0[70] using the method proposed by Bender et al[100] and extended by Crowther and Lambert[97].

Before presenting the DGMs, a description of the parameters is included below including details on how the values for the parameters were selected:

Hazards - to set the λ_e and γ values initial tuning simulations were undertaken to identify censoring levels of 25% and 65%. Weibull parameter values were chosen that gave median event times of approximately 3.5 years when the censoring level was 25% and median event times of approximately five years when the censoring level was 65%.

Censoring Level/Rate: to set the censoring level as discussed in Table 13, a random sample was drawn from the exponential distribution using a sample size of 500 for each study and a rate value, λ_c of 0.03 for 25% censoring and 0.001 for 65% censoring. Initial tuning simulations were undertaken to identify λ_c values that gave average mean censoring levels of 25% and 65%. Appendix 9 presents the mean and range of all censoring levels used for each DGM.

Time dependent treatment effect: for scenarios where PH is valid, the survival times were simulated from a baseline Weibull distribution as in equation (1) using λ_e and γ values as defined in Table 13 with β_0 set to zero, with equivalent hazard ratio value of one, indicating no treatment effect.

In order to introduce non-proportional hazards into the DGM, model (13) is extended to include an interaction between treatment covariate and function of time, as in model (14):

$$\mathrm{h_i(t)} = \lambda\gamma \mathrm{t}^{\gamma-1} \exp\{(\beta_0 \mathrm{treat} + \beta_1 \mathrm{treat\,x\,log(t)})\}, \qquad (14)$$

where β_1 is the amount by which the log HR for treatment changes for each one unit increase in log(t). The true value of β_1 was set to zero (no time dependent log HR so PH is valid), 0.1 (small time varying effect) and 0.5 (large time varying effect).

Meta-analysis heterogeneity: in order to introduce heterogeneity into the simulated datasets, values for β_0 were drawn for each study from a random effects distribution with between study variance of 0.11. The standard deviation value was selected after undertaking some initial simulations to identify a standard deviation value that gave an average I^2value of 20-50%.

The characteristics of the twelve DGMs are as follows:

Table 13 : List of DGMs

DGM	Set	Parameters					
		Lambda & Gamma values	Censoring Level	Censoring Rate: λ_c	β_0	β_1	Meta-analysis -Heterogeneity (Between studies variance)
1	1	$\lambda = 0.1;\ \gamma = 1.5$	25%	0.03	0	-	Homogeneous (0)
2		$\lambda = 0.1;\ \gamma = 1.5$	25%	0.03	0	0.1	Homogeneous (0)
3		$\lambda = 0.1;\ \gamma = 1.5$	25%	0.03	0	0.5	Homogeneous (0)
4		$\lambda = 0.075;\ \gamma = 1.0$	65%	0.001	0	-	Homogeneous (0)
5		$\lambda = 0.075;\ \gamma = 1.0$	65%	0.001	0	0.1	Homogeneous (0)
6		$\lambda = 0.075;\ \gamma = 1.0$	65%	0.001	0	0.5	Homogeneous (0)
7	2	$\lambda = 0.1;\ \gamma = 1.5$	25%	0.03	0	-	Heterogeneous: (0.11)
8		$\lambda = 0.1;\ \gamma = 1.5$	25%	0.03	0	0.1	Heterogeneous: (0.11)
9		$\lambda = 0.1;\ \gamma = 1.5$	25%	0.03	0	0.5	Heterogeneous: (0.11)
10		$\lambda = 0.075;\ \gamma = 1.0$	65%	0.001	0	-	Heterogeneous: (0.11)
11		$\lambda = 0.075;\ \gamma = 1.0$	65%	0.001	0	0.1	Heterogeneous: (0.11)
12		$\lambda = 0.075;\ \gamma = 1.0$	65%	0.001	0	0.5	Heterogeneous: (0.11)

The RMST was chosen as a measure that can be used for comparison across the different methods of analysis. The interpretation of the difference in RMST between two treatment groups is quite simple, e.g. the difference in 5-year RMST is 0.75 years which can be interpreted as patients in the treatment group have a prolongation of 9 months in life expectancy during the first 5 years compared to patients in the control group.

7.2.2 Restricted Mean Survival Time

Using the definition from Royston and Parmar[44], the RMST of a random variable T is the mean of the survival time $\mathrm{X} = \min(\mathrm{T}, \mathrm{t}^*)$ limited to a specific time point $\mathrm{t}^* > 0$, so the RMST equals the area under the survival curve $\mathrm{S}(\mathrm{t})$ from $\mathrm{t} = 0$ to $\mathrm{t} = \mathrm{t}^*$.

$$\mathrm{RMST} = \mu = \mathrm{E}[\mathrm{X}] = \mathrm{E}[\min(\mathrm{T}, \mathrm{t}^*)] = \int_0^{\mathrm{t}^*} \mathrm{S}(\mathrm{t})\,\mathrm{dt} \qquad (15)$$

where S(t) is the survival function for a given distribution, and t* is the specific time point of interest.

For this simulation study, a time point of 4.5 years was used. A value of 4.5 years was chosen for t* in order to calculate the RMST as close to the last observed event time as possible as it was more likely to have patients included in the simulation at 4.5 years rather than at the maximum follow-up time of five years.

The “true” RMST was also calculated as seen in section 7.2.1, in order to allow direct comparison of the results of the four different models under the different scenarios. The true RMST was calculated by integrating the formula in equation (13), where the Weibull distribution is used. As the integral does not have an easy closed form, it was approximated using Gauss-Legendre quadrature[97] to calculate S(t) by firstly reparametrizing the integral in the following way:

$$\int_{0}^{t^*} \mathrm{S(u)\ du} = \frac{t^*}{2}\int_{-1}^{1} \mathrm{S}\left(\frac{t^*}{2}\mathrm{z} + \frac{t^*}{2}\right)\ \mathrm{dz}$$

Then using Gauss-Legendre quadrature[97]:

$$\approx \frac{t^*}{2}\sum_{i=1}^{m} \mathrm{w_i S}\left(\frac{t^*}{2}\mathrm{z_i} + \frac{t^*}{2}\right)$$

where $\mathrm{w_i}$ and $\mathrm{z_i}$ are sets of weights and node locations respectively, where the total number of nodes is denoted m. These nodes and weights can be generated using statistical software, where the nodes are vectors of values at which to evaluate the function and weights are vectors of weights to give the function values. The R package Statmod[101], version 1.4.32 was used to generate the values. As the data was simulated under the Weibull distribution, the "true" RMST value was calculated as:

$$\mathrm{RMST} = \int_{0}^{t*} \exp(-\lambda t^{\gamma} \exp(\beta_0 \mathrm{treat}))\ \mathrm{dt}$$

$$\approx \frac{t^*}{2}\sum_{i=1}^{m} \mathrm{w_i} \exp\left(-\lambda\left[\left(\frac{t^*}{2}\mathrm{z_i} + \frac{t^*}{2}\right)^{\gamma}\right]\exp(\beta_0 \mathrm{treat})\right)$$

The parameter values given above in the DGMs were used for calculating the true RMST per scenario. For example, for scenario 1 the following true parameter values λ=0.1, γ=1.50, β_0=0, t* = 4.5 and m=100 were used. For scenarios where the treatment effect is time dependent, the formula for "true" RMST was:

$$RMST = \int_{0}^{t*} \exp(-\lambda t^{\gamma} \exp(\beta_0 \mathrm{treat} + \ \beta_1 treat\ x\ \log(t)))\ \mathrm{dt}$$

$$\approx \ \frac{t^*}{2}\sum_{i=1}^{m} w_i\, exp\left\{-\left(\lambda\gamma \exp(\beta_0 treat)\ \frac{1}{\gamma +\ (\beta_1 treat)}\ \left[\left(\frac{t^*}{2}z_i +\ \frac{t^*}{2}\right)^{(\gamma + \,\beta_1 treat)}\right]\right)\right\}$$

7.2.3 Analysis of simulated data

7.2.3.1 Models examined

Each simulated dataset in each DGM was analysed in four ways, using the following:

1. A Cox proportional hazards model;

$$h_i(t) = h_0(t)\exp(\beta_0 \text{treat})$$

   where $h_0(t)$ is the unspecified baseline hazard, where treat is the binary treatment indicator for individual i, and β_0 is the log HR for treatment. The stcox package, version 7.5.2 in the Stata software package, version 14.1[69] was used to fit the Cox model. The Cox model[5] was chosen, as it is the most commonly used method for analysing TTE data in RCTs, however ignores any time varying coefficients.

2. A Weibull proportional hazards model;

$$h_i(t) = \lambda\gamma t^{\gamma-1}\exp(\beta_0 \text{treat})$$

   where treat is the binary treatment indicator for individual i, λ and γ are the scale and shape parameters for the Weibull baseline hazard and β_0 is the log hazard ratio for treatment. The Stata package streg, version 6.4.3 was used to fit the Weibull model. The Weibull model was chosen as it is the most commonly chosen parametric model[3] and as the data has been generated under the Weibull distribution, it is expected that it will perform well.

3. An Accelerated failure time model;

$$h_i(t) = e^{-\eta_i} h_0(t/_{e^{\eta_i}})$$

where $\eta_i = \alpha_1 x_{1i}$ is the linear component of the model, in which x_{1i} is the value of the explanatory variable, for the ith individual, i=1,2,…,500. As in the proportional hazards model, the baseline hazard function, $h_0(t)$, is the hazard of death at time t for an individual for whom the values of the p explanatory variables are all equal to zero.[3] The Stata package streg, version 6.4.3 was used to fit the accelerated failure time (AFT) model. The AFT model was chosen, as it is an alternative approach to use when PH is invalid, where using the Cox model is not deemed appropriate. Although they are not used often in RCTs, they are easy to use and the effect measure, time ratios are easier to interpret than a hazard ratio[8]. For example, a time ratio of 1.20 suggests that the treated group dies at a 20% faster rate compared to the control group.

4. A flexible parametric model with time-dependent coefficients

A flexible parametric survival model as proposed by Royston and Parmar, which incorporate time-dependent effects of covariates x, is expressed as:

$$\log H_i\,(t) = \gamma_0 + \gamma_1 y + \gamma_2 \upsilon_1(y) + \gamma_3 \upsilon_3(y) + \beta_0 \text{treat} + \beta_1 \text{treat x} \log(t)$$

where $y = \log t$, β_0 is the log HR for treatment, β_1 is the amount by which the log HR for treatment changes for each one unit increase in log(t) and for the *j*th knot at $k_j, j = 1,2,3,$

$$\upsilon_j(y) = \left(y - k_j\right)_+^3 - \lambda_j(y - k_{min})_+^3 - \left(1 - \lambda_j\right)(y - k_{max})_+^3,$$

and $$\lambda_j = \frac{k_{max} - k_j}{k_{max} - k_{min}},$$

For the simulation study, the following model was fitted with 2 knots for the restricted cubic spline function for the baseline hazard function and 1 degree of freedom for the time-dependent effect, with 1 degree of freedom stating that a linear effect of log time is being fitted. The Stata package stpm2[102], version 1.4.4 was used to fit the flexible parametric models. It was decided that two knots would be sufficient for this model as the recommendations are between one and five knots[102] with two knots being the default number in Stata. As only one time-dependent effect is included in our model, 1 degree of freedom for each time-dependent effect seemed sufficient.

The flexible parametric model was chosen, as it is sometimes difficult to determine which probability distribution should be used to model the survival times, as the given distribution may not fit the data perfectly. Instead, the Royston and Parmar method can be applied, which models the underlying baseline hazard parametrically but allows the model to have greater flexibility than is possible with fully parametric models. The benefits of using flexible parametric models are becoming more recognised in applied research[103 104], such as that it is possible to allow covariates to have time-dependent effects by fitting interactions between the covariate and time using a second spline function and that flexible parametric models offer a good alternative to standard parametric models as they are able to summarize simple and complex effects which standard parametric models sometimes struggle to highlight.

7.2.3.2 Software

The R software package was used for data generation as running simulations on single computers can be very time consuming so during the simulation studies, the University of Liverpool's HTCondor system was used[105 106]. The model fitting was performed in Stata as my previous programming experience was mainly in Stata. For estimating RMST, a mix of Stata and the R package were used due to the chance to have multiple datasets open at the same time whilst Stata was the better performing software for fitting meta-analysis. However, all of this work can be performed in both the R package and Stata.

7.2.3.3 Estimating RMST

To calculate the estimated RMST for each proposed modelling approach, the four analysis methods stated in Section 7.2.3.1 were applied to the simulated data. The definitions of the RMST for all the models is given below:

Cox model:

$$\mathrm{RMST} = \int_{0}^{t*} \exp(-h_0(t)\exp(\beta_0 \mathrm{treat}))\,dt \tag{16}$$

Weibull model:

$$\mathrm{RMST} = \int_{0}^{t*} \exp(-\lambda t^{\gamma}\exp(\beta_0 \mathrm{treat}))\,dt \tag{17}$$

AFT model:

$$\mathrm{RMST} = \int_{0}^{t*} \exp(-\lambda t^{\gamma}\exp(-\beta_0 treat)^{\gamma})\,dt \tag{18}$$

Flexible Parametric model:

$$\mathrm{RMST} = \int_{0}^{t*} \exp\{-\exp[\gamma_0 + \gamma_1 y + \gamma_2 \upsilon_1(y) + \gamma_3 \upsilon_3(y)] + \beta_0 \mathrm{treat} + \beta_1 \mathrm{treat}\ x \log(t)\}\ dt \qquad (19)$$

These integrals were difficult to evaluate directly or by using numerical approximation procedures, so values were sampled from the distributions of the parameters from the fitted models and using these RMST values were calculated for each of these sampled parameter sets. The following procedure was followed for estimating the RMST:

1. The estimated coefficients, parameter values and their (joint) uncertainty were obtained directly from the fitted models, e.g. shape, scale, treatment estimate and variance-covariance matrix for a Weibull model.
2. A total of 1000 values for the treatment coefficient were generated using a normal distribution for the Cox, Weibull and AFT models, where the mean was the treatment coefficient and the variance term was the uncertainty term from the fitted models so variance or variance-covariance matrix. For the flexible parametric model, a multivariate normal distribution was used as it includes more than one treatment parameter so the vector of means is made up of the treatment and treatment*time coefficients and the variance-covariance matrix is used to account for the uncertainty around the parameters.
3. For each iteration, the RMST was calculated using the formulae in (16), (17), (18) and (19) where the parameter values extracted from the fitted models in step 1 and the treatment coefficients generated in step 2 were used. The Gauss-Legendre quadrature[97] method as described in section 7.2.2 to calculate the "true" RMST was used as the integrals did not have an easy closed form. A t* value of 4.5 years was

used and the total number of nodes was 100. The R package Statmod[101], version 1.4.32 was used to generate the values. The RMST for each of the two treatment groups was estimated.

4. This then gave a set of simulated RMST values, from which the mean and standard error of the RMST for each fitted model could be extracted.

7.2.3.4 Meta-analysis

Once the data had been simulated, meta-analysis was conducted using a two-stage approach where the models discussed in section 7.2.3.1 were fitted separately to each study within each meta-dataset within each DGM. A two-stage approach was chosen as they are considered to be less complicated to fit compared to one-stage methods[71 107 108]. For the Cox, Weibull and AFT models, the Stata package ipdmetan[109], version 1.06 was used to conduct the two-stage approach. The initial plan was to fit both random and fixed effects models and to compare the performance. However, due to the amount of time taken to fit all the models to the datasets, it was too time consuming to fit both types of models. Therefore, a decision was made to fit random-effects models for all the models, as there was some heterogeneity present between the studies even for scenarios where the meta-analysis was homogeneous. It is important to note that for cases where there was little between study heterogeneity, it is expected that fixed and random effects models would give similar results. The Cox model for the event time of the i^{th} participant, i = 1,…,500, in study j, j = 1,…,5 at time t can be written as:

$$h_{ij}(t) = h_{0j}(t)\exp\left(\beta_{0j} treat_{ij}\right) \tag{20}$$

where $treat_{ij}$ is the treatment group indicator for the i^{th} participant in the j^{th} study, β_{0j} represents the log HR between the treatment groups for the j^{th} study and h_{ij}, h_{0j} represent the subject specific and baseline hazard functions, respectively. The estimated HRs for the treatment effect from each study may then be combined in a meta-analysis, using the DerSimonian-Laird random effects model[40].

All of the overall effect estimates across the five studies within each meta-dataset (β_0 for the Cox, Weibull and AFT models) were pooled in order to obtain the average pooled hazard ratio or time ratio as well as standard deviations to account for the variability across the 1000 simulated datasets.

For the flexible parametric model, a flexible parametric model using the Royston and Parmar method was fitted to each individual study and then the studies were combined together before performing a multivariate meta-analysis. A multivariate meta-analysis was used as multiple parameters from the flexible parametric model were extracted from the single model. The Stata package mvmeta[110], version 3.2.0 was used to conduct a random effects multivariate meta-analysis. The delta method[111] was used in order to obtain the overall estimate of the standard error. The delta method is a procedure for approximating expected values of functions of random variables where direct estimation of the expected values is not feasible[111]. Due to the ease of coding the delta method and the chance to have multiple datasets open at the same time, the R package was used to estimate the overall effect estimate and standard error and then obtain the effect estimates and 95% CIs at various timepoints as seen in Table 14 to Table 17. The R package msm, version 1.6.7 was used to perform the delta method in order to obtain the overall standard error.

Similarly, once the estimated RMST values had been calculated, the Stata package admetan[109], version 1.06 was used to perform a two-stage random effects aggregate data meta-analysis. The difference in RMST between the two treatment groups along with the standard deviation surrounding the difference in RMST was meta-analysed. The average pooled difference in RMST and standard deviation was then calculated across all 1000 datasets.

The between study heterogeneity for the Cox, Weibull and AFT models was assessed using I^2 and τ^2 values as well as the standard deviations. For the flexible parametric model, the multivariate I^2 statistic was calculated in order to assess the between study heterogeneity. For the estimated RMST values, the between study heterogeneity was assessed using the I^2 statistic and then presented using the mean pooled I^2 value as well as the standard deviation.

7.2.4 Performance measures

The following performance measures were calculated for each analysis method within each DGM to assess the performance of the four different survival models at estimating the RMST:

Bias: the bias is the deviation in an estimate from the true value so here it is the difference between the average estimate and the true value.[96] $E[\hat{\theta}] - \theta$, where $\hat{\theta}$ is the point estimate and θ is the true value. It is desirable to have a bias close to zero.

Empirical SE: the empirical SE is the standard deviation of the point estimates. $\sqrt{Var(\hat{\theta})}$

Model SE: the model SE is the square root of the average standard error: $\sqrt{E\left[\widehat{se}(\hat{\theta})^2\right]}$

Relative % error in Model SE: the definition of the relative % error in model SE is: $100 * \left(\frac{ModSE}{EmpSE} - 1\right)$

This measure explains whether the model-based SE is being overestimated or not[98]. The relative % error in model SE will be provided in the results tables as it is a more useful measure rather than the individual empirical SE and model SE.

Mean squared error: the mean squared error (MSE) is the average squared difference between the true value and the point estimate. It is desirable to have the mean squared error (MSE) close to zero: $E\left[(\hat{\theta} - \theta)^2\right]$

Coverage: the coverage of a confidence interval is the proportion of times the 95% confidence interval of the point estimate contains the true parameter value. If the method is performing well, it is expected coverage to be 95%. Coverage higher than 95% indicates an inefficient estimator and coverage less than 95% indicates an inaccurate estimator: $P\left(\hat{\theta}_{low} \leq \theta \leq \hat{\theta}_{upp}\right)$

The coverage has been presented graphically in the form of a "zip plot" which helps with understanding coverage by seeing the confidence intervals directly. For all of the DGMs and methods, the confidence intervals are fractional-centile ranked according to $|z_i|$, where $z_i = \left(\widehat{\theta_\iota} - \theta\right)/ModSE$, where $\hat{\theta}$ is the point estimate, θ is the true value and ModSE is the model SE as mentioned in this section. This ranking is then used for the vertical axis and is plotted against the confidence intervals. The intervals which cover θ are blue (as seen in the bottom end of intervals) and those that do not cover θ are in purple (as seen towards the top of the intervals). If a method has 95% coverage, the colour of the interval changes at 95 on the vertical axis. The blue horizontal lines that are the full width of the intervals are Monte Carlo 95% confidence intervals for per cent coverage.

The Monte Carlo SEs (MCSEs) were also presented in order to quantify simulation uncertainty, which is: $\sqrt{\frac{1}{n_{sims}}\widehat{Var}(\hat{\theta})} = \frac{EmpSE}{\sqrt{n_{sims}}}$

The performance measures were calculated using the Stata package simsum[112], version 0.17.1.

7.3 Results

The results for the simulations have been split into four sets of simulation results. The parameters being varied include censoring level and whether the treatment effect across studies is homogeneous or heterogeneous. Within the sets of simulations, the level of the time-dependent log HR has been varied. All of these details are included in Table 13 including details on the values of the parameters being used. In this section, the results will be referred to as on the exponential scale so as HRs rather than log HRs.

7.3.1 Simulation set 1: Varying censoring, Meta-analysis: Homogeneous

For the first set of simulations (Table 14 and Table 15 and Figure 17 and Figure 18) all the models converged except for DGMs 5 and 6 where the flexible parametric model failed to converge for two datasets.

The results for DGM 1 where PH is valid highlight that this DGM performs well with low bias, coverage around 95% and low MSE as seen in Table 14. However, as soon as a small time-dependent effect is introduced all models other than the flexible parametric model perform poorly. The mean pooled HRs for the Cox and Weibull models for DGMs 2, 3, 5 and 6 are close to the results at year 5 obtained from the flexible parametric model, highlighting when it is inappropriate to use a Cox and Weibull model. It is important to note that as the data has been generated under the Weibull distribution it would be expected that the Weibull model for DGMs 1 and 4 would be the gold standard.

The mean pooled HR in Table 14 for the Cox model under DGM 3 suggests that there may be a 42% greater risk of dying in the treatment group compared to the control group at any time during follow-up. However, 95% confidence intervals would be needed to be certain of evidence of an effect. Similarly, for the AFT model under DGM 3, the mean pooled TR suggests that the patients in the treatment group have a 19% shorter life expectancy compared to patients in the control group. A t* value of 4.5 years was used for the RMST analysis. Using the Cox model, the mean pooled difference in RMST was -0.387 years, suggesting a loss in RMST of around 4.5 months.

The results in Figure 17 for the bias around the difference in RMST illustrate that for DGM 1 all models have bias close to zero, suggesting well performing methods. However, compared to DGMs 2 and 3 the Cox, Weibull and AFT models are not performing well especially for DGM 3.

In Figure 18, the "zip plot" for DGM 1 shows that all survival models are performing well. However, for DGM 2, the coverage is lower at around 92% for all models except the flexible parametric model. There are more intervals to the left of θ than to the right, especially intervals not covering θ which suggest that the model SEs must underestimate the empirical SE, although the coverage is also slightly lower than the nominal 95% level. For DGM 3, only the flexible parametric model is performing well as expected under a DGM where the PH is invalid. The "zip plots" are slightly "hairy" as compared to other examples[98], however as the RMST values are so small there are some results included that are close to the "true" value but have slightly wider confidence intervals and some results that are close to the true value but have narrower confidence intervals.

In comparison, by varying the censoring, the results for DGMs 4 to 6 were similar for the pooled effect estimates and for majority of the performance measures. The main differences were found when comparing the coverage, where the coverage was slightly higher for DGM

3 for the Cox, Weibull and AFT models compared to DGM 6. For the flexible parametric model, coverage was below 95% for DGM 3 whereas coverage was above 95% for DGM 6 highlighting under and over coverage.

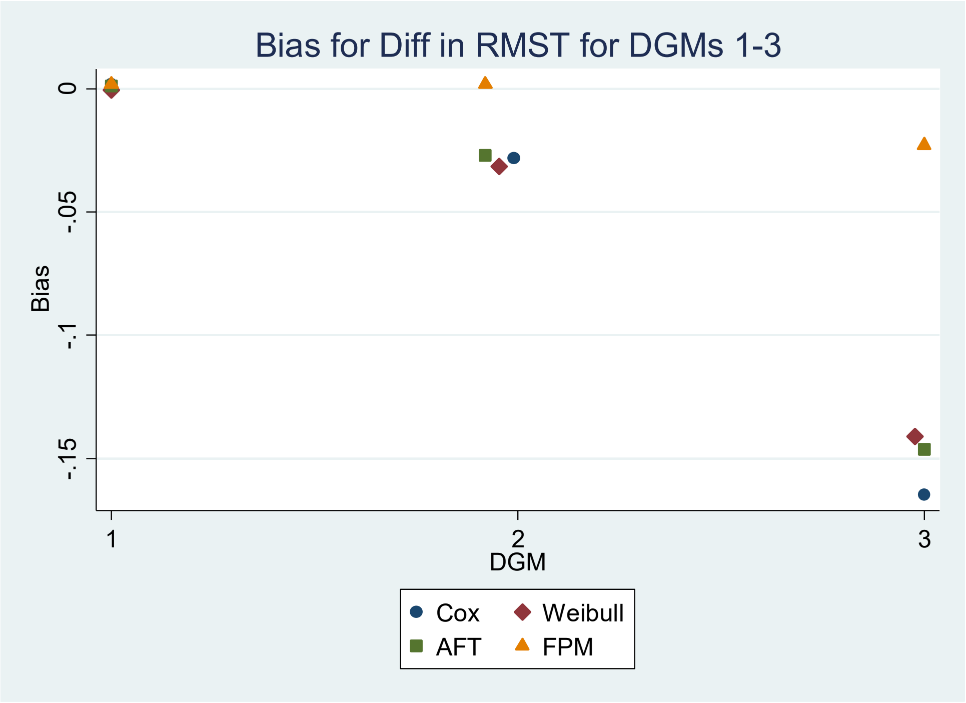
Table 14 : Simulation results for DGMs 1 to 3 and by model being performed.

DGM	Parameters	Models	Mean pooled HR (SD)	Mean pooled TR (SD)	Mean pooled difference in RMST (SD)	Bias (MCSE)	Coverage (MCSE)	MSE (MCSE)	Relative error in Model SE (MCSE)
1	*Censoring: 25%* *Homogeneity: Yes* *Time-dependent log HR: None* *PH Valid: Yes*	Cox	1.00 (0.05)	-	0.001 (0.05)	0.001 (0.002)	95.5% (0.66)	0.003 (<0.001)	5.15% (2.45)
		Weibull	1.00 (0.05)	-	0.001 (0.05)	0.001 (0.002)	95.3% (0.67)	0.003 (<0.001)	5.47% (2.45)
		AFT	-	1.00 (0.04)	0.001 (0.05)	0.001 (0.002)	95.6% (0.65)	0.003 (<0.001)	5.47% (2.45)
		FPM	Yr 1: 1.00 (0.09) Yr 3: 1.00 (0.05) Yr 5: 1.00 (0.05)	-	0.002 (0.06)	0.002 (0.002)	95.0% (0.69)	0.004 (<0.001)	6.42% (2.48)
2	*Censoring: 25%* *Homogeneity: Yes* *Time-dependent log HR: 0.1* *PH Valid: No*	Cox	1.07 (0.05)	-	-0.069 (0.05)	-0.033 (0.002)	92.3% (0.84)	0.004 (<0.001)	10.62% (2.59)
		Weibull	1.07 (0.05)	-	-0.068 (0.05)	-0.032 (0.002)	92.2% (0.85)	0.004 (<0.001)	11.43% (2.60)
		AFT	-	0.96 (0.03)	-0.068 (0.05)	-0.032 (0.002)	92.3% (0.84)	0.004 (<0.001)	11.33% (2.60)
		FPM	Yr 1: 0.96 (0.09) Yr 3: 1.04 (0.05) Yr 5: 1.09 (0.05)	-	-0.038 (0.06)	-0.002 (0.002)	96.2% (0.60)	0.003 (<0.001)	12.84% (2.65)
3	*Censoring: 25%* *Homogeneity: Yes* *Time-dependent log HR: 0.5* *PH Valid: No*	Cox	1.42 (0.07)	-	-0.387 (0.05)	-0.165 (0.002)	16.5% (1.17)	0.030 (<0.001)	9.31% (2.53)
		Weibull	1.42 (0.07)	-	-0.367 (0.05)	-0.145 (0.002)	20.2% (1.27)	0.024 (<0.001)	12.50% (2.58)
		AFT	-	0.81 (0.02)	-0.367 (0.05)	-0.145 (0.002)	20.5% (1.28)	0.024 (<0.001)	12.43% (2.58)
		FPM	Yr 1: 0.86 (0.10) Yr 3: 1.21 (0.12) Yr 5: 1.43 (0.24)	-	-0.244 (0.06)	-0.022 (0.002)	94.6% (0.71)	0.004 (<0.001)	10.35% (2.59)

Table 15 : Simulations results for DGMs 4 to 6 and by models being performed

DGM	Parameters	Models	Mean pooled HR (SD)	Mean pooled TR (SD)	Mean pooled difference in RMST (SD)	Bias (MCSE)	Coverage (MCSE)	MSE (MCSE)	Relative error in Model SE (MCSE)
4	*Censoring: 65%* *Homogeneity: Yes* *Time-dependent log HR: None* *PH Valid: Yes*	Cox	1.00 (0.07)	-	0.003 (0.04)	0.003 (0.001)	95.5 (0.66)	0.002 (<0.001)	8.91% (2.53)
		Weibull	1.00 (0.07)	-	0.002 (0.04)	0.002 (0.001)	95.6 (0.65)	0.002 (<0.001)	9.77% (2.55)
		AFT	-	1.00 (0.07)	0.002 (0.04)	0.002 (0.001)	96.0 (0.62)	0.002 (<0.001)	10.57% (2.56)
		FPM	Yr 1: 1.00 (0.12) Yr 3: 1.00 (0.08) Yr 5: 1.00 (0.07)	-	0.002 (0.05)	0.002 (0.001)	96.2 (0.60)	0.002 (<0.001)	14.51% (2.70)
5	*Censoring: 65%* *Homogeneity: Yes* *Time-dependent log HR: 0.1* *PH Valid: No*	Cox	1.06 (0.08)	-	-0.034 (0.05)	-0.033 (0.001)	92.4 (0.84)	0.003 (<0.001)	10.41% (2.70)
		Weibull	1.06 (0.08)	-	-0.035 (0.04)	-0.034 (0.001)	91.3 (0.90)	0.003 (<0.001)	12.57% (2.62)
		AFT	-	0.94 (0.06)	-0.035 (0.04)	-0.034 (0.001)	91.5 (0.88)	0.003 (<0.001)	12.93% (2.63)
		FPM	Yr 1: 0.92 (0.12) Yr 3: 1.02 (0.08) Yr 5: 1.07 (0.08)	-	-0.004 (0.05)	-0.003 (0.002)	96.6 (0.57)	0.002 (<0.001)	10.23% (2.59)
6	*Censoring: 65%* *Homogeneity: Yes* *Time-dependent log HR: 0.5* *PH Valid: No*	Cox	1.42 (0.09)	-	-0.244 (0.05)	-0.180 (0.002)	9.8 (0.94)	0.035 (<0.001)	10.03% (2.57)
		Weibull	1.42 (0.09)	-	-0.217 (0.04)	-0.153 (0.001)	10.2 (0.96)	0.025 (<0.001)	18.09% (2.71)
		AFT	-	0.76 (0.04)	-0.216 (0.04)	-0.152 (0.001)	10.6 (0.97)	0.025 (<0.001)	18.78% (2.72)
		FPM	Yr 1: 0.67 (0.09) Yr 3: 1.15 (0.09) Yr 5: 1.49 (0.10)	-	-0.054 (0.05)	0.010 (0.002)	95.9 (0.63)	0.003 (<0.001)	12.84% (2.67)

Figure 17: Plots of Bias for the difference in RMST for DGMs 1 to 3 and then for DGMs 4 to 6



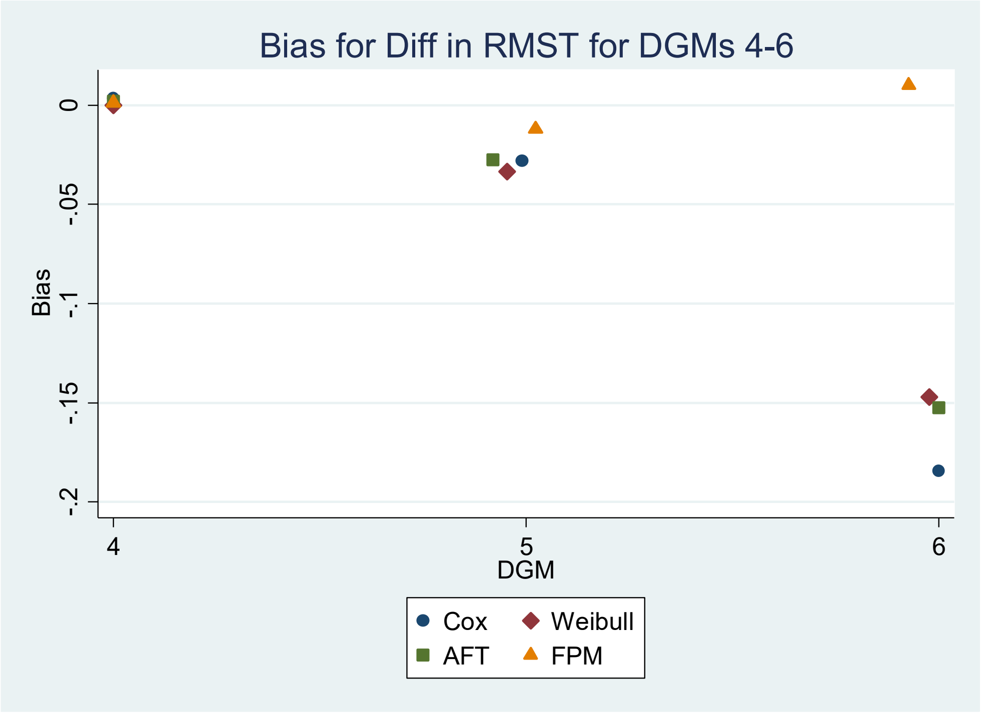
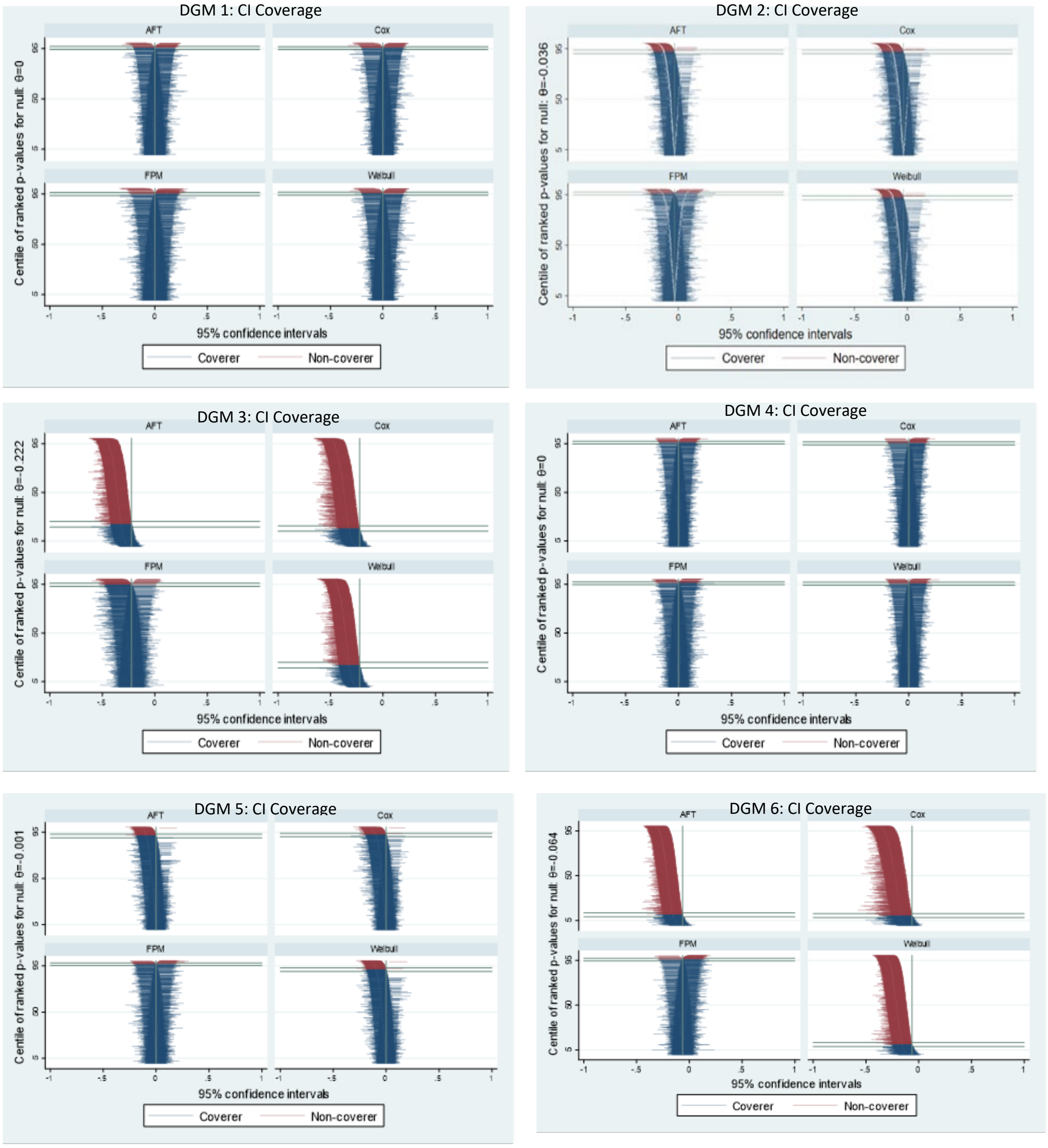


Figure 18: “Zip plot” of the 1000 confidence intervals for DGM 1 to 6 by analysis method



7.3.2 Simulation set 2: Varying censoring, Meta-analysis: Heterogeneous

For the second set of simulations (Table 16 and Table 17, and Figure 19 and Figure 20) all the models converged except for DGMs 8, 10 and 11 where up to four datasets failed to converge.

The estimates from DGM 7 where no time-dependent effect was included, highlights that all models have low bias and low MSE. The coverage is around 98% suggesting some over-coverage. However, as soon as a small time-dependent effect is included, the estimated mean pooled HR increases compared to the “true” HR the data is simulated under. Similar patterns are seen when a larger time dependent effect is included where the mean pooled HR is 1.45 compared to the “true” HR of 1.0.

The mean pooled HR in Table 16 for the Cox model under DGM 8 suggests that there may be a 7% greater risk of dying in the treatment group compared to the control group at any time during follow-up. However, 95% confidence intervals would be needed to be certain of evidence of an effect. Similarly, for the AFT model under DGM 8, the mean pooled TR suggests that the patients in the treatment group have a 4% shorter life compared to patients in the control group. A t* value of 4.5 years was used for the RMST analysis. Using the Cox model, the mean pooled difference in RMST was -0.068 years, suggesting a loss in RMST of less than 1 month.

Figure 19 illustrates the bias for the DGMs for all four models. Compared to DGM 7, the flexible parametric model has lower bias compared to the other three models for both DGM 8 and 9 suggesting that the estimates from the flexible parametric model are closer to the “true” difference in RMST.

In Figure 20, the “zip plot” for DGM 7 shows that the plot is “hairier” than seen previously in Figure 18 suggesting wider confidence intervals throughout for this DGM. The coverage is higher at around 98% due to which the purple intervals not covering θ are not as evident. However, for DGM 8, all models are performing well with coverage around 93%, however the flexible parametric model is performing slightly better with coverage around the nominal level of 95%. There are more intervals to the left of θ than to the right, suggesting that the model SEs are underestimating the empirical SE. For DGM 9, the coverage is poor for the Cox, Weibull and AFT models ranging between 11% and 14%. The coverage increases to around 93% for the flexible parametric model suggesting a well performing model.

In comparison, by increasing the censoring level from around 25% to around 65%, DGM 10 where no time varying effect is present, all of the survival models are performing well with low bias, low MSE and high coverage around 95%. The "zip plot" however suggests that as there are more intervals to the left of θ for the Cox, Weibull and AFT models, the model SEs may be underestimating the empirical SE.

Even once a small time-dependent effect is included, DGM 11 is performing well for all models with again low bias, low MSE and coverage of around 97% for the Cox, Weibull and AFT models suggesting slight over-coverage, however for the flexible parametric model the coverage was around the nominal level of 95%. The "zip plot" for the flexible parametric model shows that there are more intervals to the right of θ, suggesting that the model SEs are overestimating the empirical SE. However, as soon as a large time-dependent effect is introduced for DGM 12 all models except the flexible parametric model are struggling.

In comparison, when comparing simulated datasets that are homogeneous as seen in DGMs 1 to 3 to simulated datasets that are heterogeneous as seen in DGMs 7 to 9, there is evidence of over-coverage and considerably higher relative error in model SE present in results where the simulated datasets are heterogeneous. Similarly, for DGMs where censoring is high and the simulated datasets are homogeneous and heterogeneous, similar patterns are seen with again only the flexible parametric model performing well.

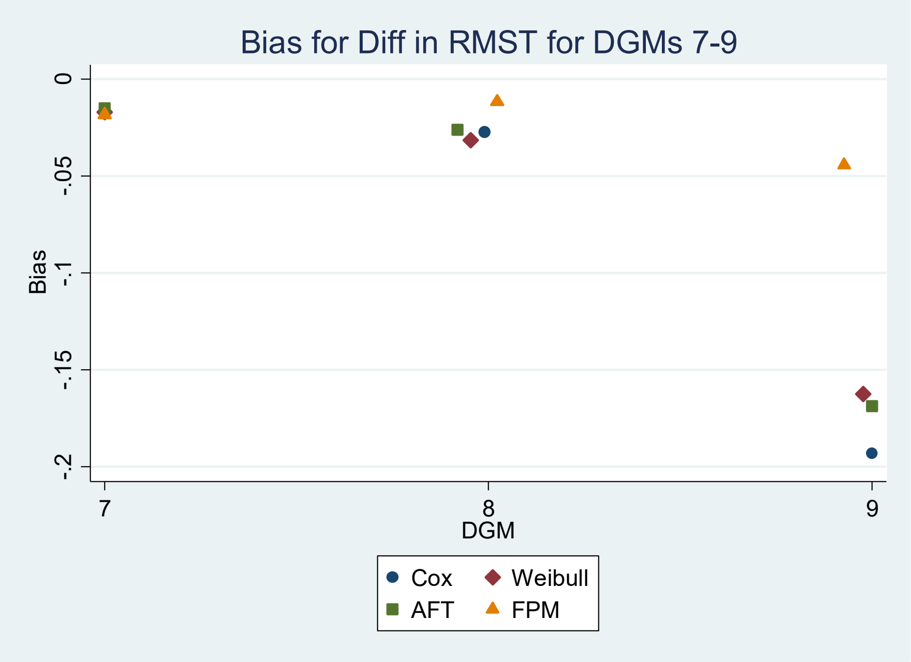
Table 16 : Simulation results for DGMs 7 to 9 and by model being performed

DGM	Parameters	Models	Mean pooled HR (SD)	Mean pooled TR (SD)	Mean pooled difference in RMST (SD)	Bias (MCSE)	Coverage (MCSE)	MSE (MCSE)	Relative error in Model SE (MCSE)
7	*Censoring: 25%* *Homogeneity: No* *Time-dependent log HR: None* *PH Valid: Yes*	Cox	1.02 (0.05)	-	-0.016 (0.05)	-0.016 (0.002)	98.5 (0.38)	0.003 (<0.001)	52.35% (3.64)
		Weibull	1.02 (0.05)	-	-0.015 (0.05)	-0.015 (0.002)	98.3 (0.41)	0.003 (<0.001)	53.43% (3.67)
		AFT	-	0.98 (0.03)	-0.015 (0.05)	-0.015 (0.002)	98.5 (0.38)	0.003 (<0.001)	53.77% (3.68)
		FPM	Yr 1: 1.02 (0.09) Yr 3: 1.02 (0.05) Yr 5: 1.02 (0.05)	-	-0.018 (0.06)	-0.018 (0.002)	97.7 (0.47)	0.004 (<0.001)	42.49% (3.43)
8	*Censoring: 25%* *Homogeneity: No* *Time-dependent log HR: 0.1* *PH Valid: No*	Cox	1.07 (0.05)	-	-0.068 (0.05)	-0.032 (0.002)	92.9 (0.81)	0.004 (<0.001)	14.30% (2.68)
		Weibull	1.07 (0.05)	-	-0.068 (0.05)	-0.032 (0.002)	93.0 (0.81)	0.004 (<0.001)	15.32% (2.70)
		AFT	-	0.96 (0.03)	-0.068 (0.05)	-0.032 (0.002)	93.2 (0.80)	0.004 (<0.001)	15.41% (2.70)
		FPM	Yr 1: 0.96 (0.09) Yr 3: 1.05 (0.05) Yr 5: 1.09 (0.05)	-	-0.039 (0.06)	-0.003 (0.002)	95.7 (0.64)	0.003 (<0.001)	12.41% (2.63)
9	*Censoring: 25%* *Homogeneity: No* *Time-dependent log HR: 0.5* *PH Valid: No*	Cox	1.45 (0.06)	-	-0.411 (0.05)	-0.189 (0.002)	11.0 (0.99)	0.038 (<0.001)	19.00% (2.79)
		Weibull	1.45 (0.07)	-	-0.390 (0.05)	-0.168 (0.001)	13.3 (1.07)	0.030 (<0.001)	22.06% (2.84)
		AFT	-	0.80 (0.02)	-0.390 (0.05)	-0.168 (0.001)	14.0 (1.10)	0.030 (<0.001)	22.03% (2.84)
		FPM	Yr 1: 0.83 (0.08) Yr 3: 1.31 (0.06) Yr 5: 1.62 (0.08)	-	-0.268 (0.06)	-0.046 (0.002)	92.9 (0.81)	0.006 (<0.001)	16.33% (2.74)

Table 17 : Simulation results for DGMs 10 to 12 and by model being performed

DGM	Parameters	Models	Mean pooled HR (SD)	Mean pooled TR (SD)	Mean pooled difference in RMST (SD)	Bias (MCSE)	Coverage (MCSE)	MSE (MCSE)	Relative error in Model SE (MCSE)
10	*Censoring: 65%* *Homogeneity: No* *Time-dependent log HR: None* *PH Valid: Yes*	Cox	1.03 (0.08)	-	-0.014 (0.05)	-0.014 (0.001)	95.3% (0.67)	0.002 (<0.001)	14.63% (2.71)
		Weibull	1.03 (0.08)	-	-0.015 (0.04)	-0.015 (0.001)	95.6% (0.65)	0.002 (<0.001)	17.02% (2.76)
		AFT	-	0.98 (0.07)	-0.015 (0.04)	-0.015 (0.001)	95.8% (0.63)	0.002 (<0.001)	17.66% (2.77)
		FPM	Yr 1: 1.04 (0.14) Yr 3: 1.04 (0.08) Yr 5: 1.04 (0.08)	-	-0.015 (0.05)	-0.015 (0.001)	96.4% (0.59)	0.003 (<0.001)	14.34% (2.71)
11	*Censoring: 65%* *Homogeneity: No* *Time-dependent log HR: 0.1* *PH Valid: No*	Cox	1.03 (0.07)	-	-0.014 (0.04)	-0.013 (0.001)	97.1% (0.53)	0.002 (<0.001)	24.99% (2.97)
		Weibull	1.03 (0.07)	-	-0.014 (0.04)	-0.013 (0.001)	97.2% (0.52)	0.002 (<0.001)	27.89% (3.04)
		AFT	-	0.98 (0.06)	-0.014 (0.04)	-0.013 (0.001)	97.3% (0.51)	0.002 (<0.001)	28.52% (3.05)
		FPM	Yr 1: 0.89 (0.11) Yr 3: 0.99 (0.08) Yr 5: 1.04 (0.07)	-	0.018 (0.05)	0.019 (0.001)	95.8% (0.64)	0.003 (<0.001)	25.44% (3.01)
12	*Censoring: 65%* *Homogeneity: No* *Time-dependent log HR: 0.5* *PH Valid: No*	Cox	1.35 (0.09)	-	-0.201 (0.05)	-0.137 (0.002)	43.9% (1.57)	0.021 (<0.001)	33.26% (3.18)
		Weibull	1.35 (0.09)	-	-0.180 (0.04)	-0.116 (0.001)	42.6% (1.56)	0.015 (<0.001)	39.79% (3.30)
		AFT	-	0.79 (0.04)	-0.179 (0.04)	-0.115 (0.001)	44.3% (1.57)	0.015 (<0.001)	40.55% (3.32)
		FPM	Yr 1: 0.64 (0.09) Yr 3: 1.10 (0.09) Yr 5: 1.42 (0.10)	-	-0.022 (0.05)	0.042 (0.002)	91.3% (0.89)	0.004 (<0.001)	30.71% (3.13)

Figure 19: Plots of Bias for the difference in RMST for DGMs 7 to 12



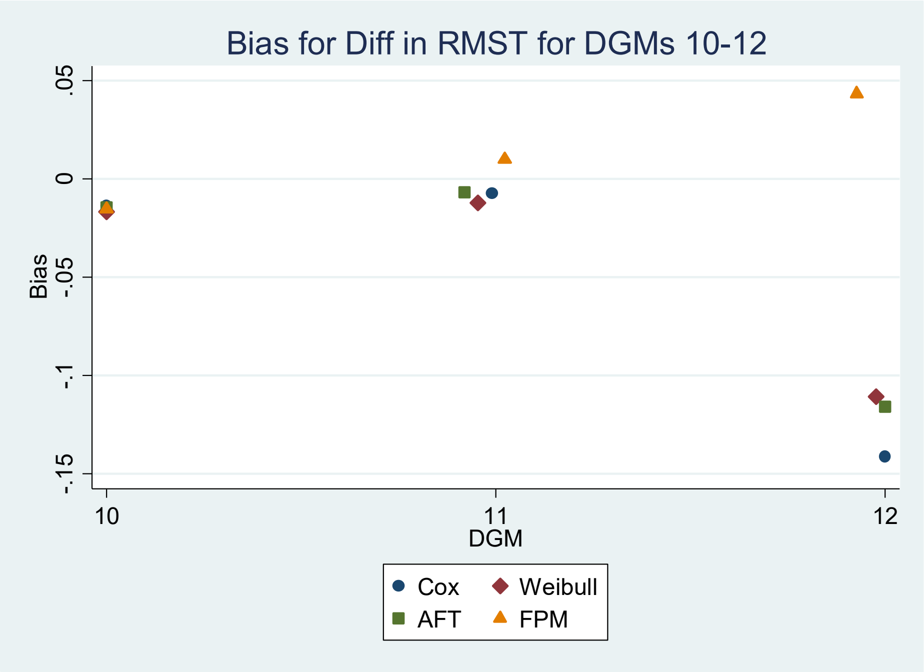
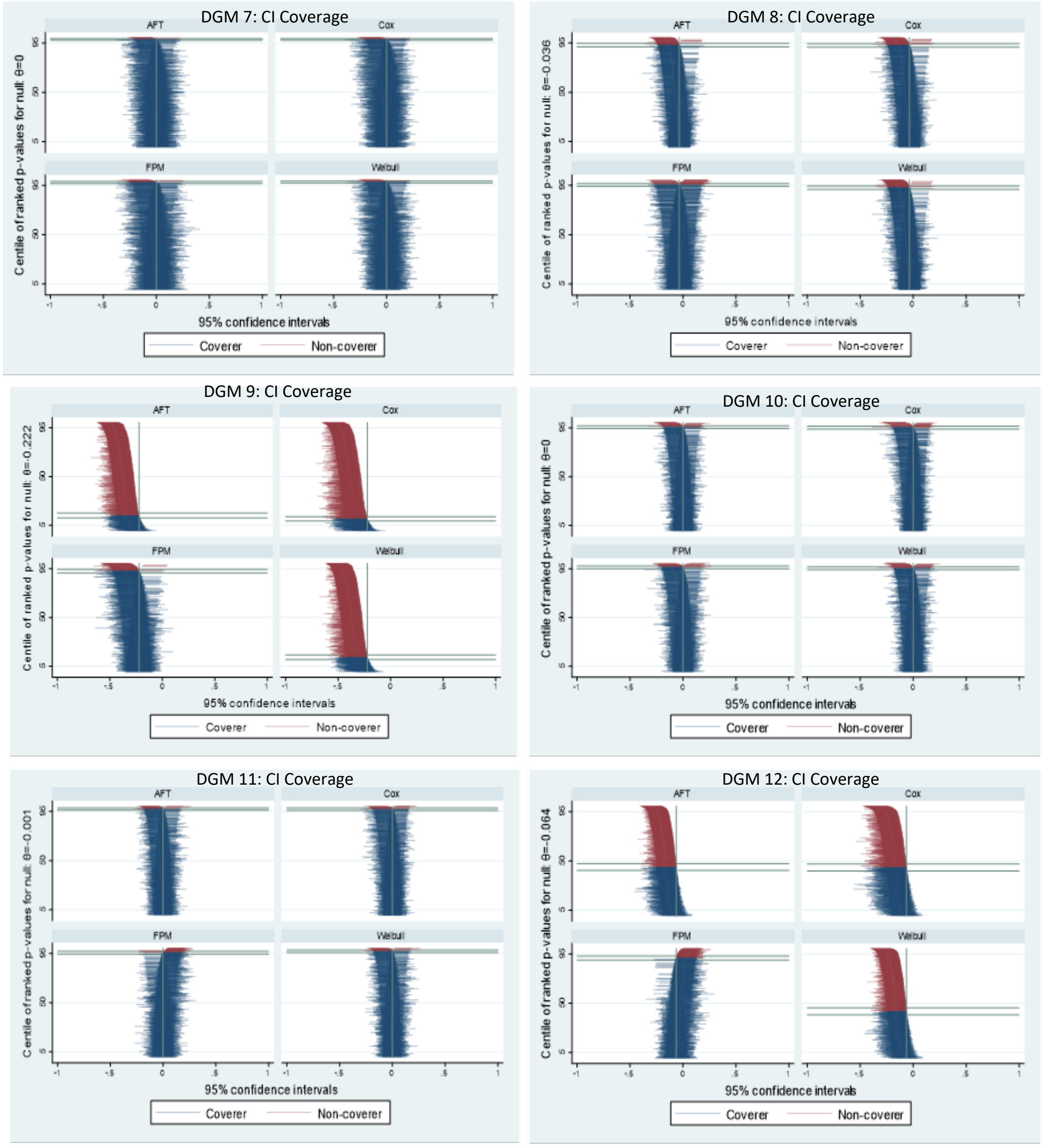


Figure 20: “Zip plot” of the 1000 confidence intervals for DGM 7 to 12 by analysis method



7.4 Discussion

7.4.1 Summary of key points and implications

A simulation study was undertaken to evaluate the performance of various different survival analysis models for meta-analysis of study data generated either assuming PH holds or not. Meta-analysis of TTE outcomes normally use the HR as a measure of the treatment effect. However, the PH assumption may not be valid for all included studies. In cases where different assumptions may have been made in the analysis of each study in the meta-analysis, the RMST is an appealing effect measure, as it does not depend on the PH assumption. This method allowed the direct comparison of the four different survival models which otherwise would not have been possible with comparing HRs, TRs and HRs varying over time.

The results presented indicate that under a valid assumption of PH, all models performed well with low bias, low MSE and high coverage. As soon as, a small-time dependent effect was introduced the Cox, Weibull and AFT models started to struggle with the mean pooled HR/TR moving away from the "true" estimate. The performance measures also demonstrated that the bias was higher and coverage was lower in comparison to the flexible parametric model which continued to do well. Similarly, once a larger time-dependent effect was introduced all models except the flexible parametric model struggled to perform well.

It is important to note that as the data was simulated under a Weibull distribution, the AFT model used is a Weibull AFT model. Hence why under the Weibull distribution, the AFT and PH models can be shown to be the same[8]. The only difference being that the AFT models differ in terms of their interpretation of effect sizes with TRs instead of HRs. Therefore, in cases where PH is invalid, there is no surprise that the RMST results based on the AFT models are behaving in a similar way to the Weibull model and not performing as well.

For simulation set 1, where censoring was varied and the simulated datasets were homogeneous, there was not much difference in the estimates. For DGMs 1 and 4 where PH is valid, there was little difference between the results for the Weibull, AFT and flexible parametric models. However, as soon as a time-dependent effect was introduced only the flexible parametric model performed well.

For simulation set 2, where the censoring levels were varied and the simulated datasets were heterogeneous, the estimates suggest that for DGMs where the censoring was increased from around 25% to around 65%, they performed marginally better with lower bias, lower MSE and coverage closer to the nominal 95% level for some DGMs and for others like DGM 12 where a large time-dependent effect is present the coverage was slightly higher. The "zip plots" presented in Figure 20 also highlight that for DGMs with higher censoring had narrower intervals compared to the DGMs with lower coverage.

The results from the simulation study illustrate that in situations where the PH assumption holds, all the survival models performed well. However, when a small-time dependent effect was introduced, the Cox, Weibull and AFT models had slightly higher bias and lower coverage compared to the results obtained under the PH valid cases. In situations where a large time-dependent effect was included the Cox, Weibull and AFT models could not cope and this was reflected in the poor effect estimates which were further from the "true" HR and the performance measures which performed poorly. The best performing model in all cases and DGMs was the flexible parametric model. The different simulation sets demonstrated which parameters are more likely to impact the results with simulations with high censoring and where the simulated datasets were homogeneous performing better.

It is important to note that both the DGM and model fitting has assumed that the model structure is the same in each study in the meta-analysis, e.g. PH valid for all studies or PH invalid for all studies. In future work, including some studies where PH is valid and some where it is invalid could be explored.

7.4.2 Limitations and future work

The data were generated under a Weibull distribution as it was identified as being the most commonly used distribution for survival data[3]. Although, this is rather restrictive, due to time constraints it was not possible to re-run the simulations and generate the data using an alternative distribution such as a Gompertz distribution which is another popular choice for survival data but this could be considered during future work.

Similarly, a Weibull model was also selected to be fitted to the simulated data as it is a model that is commonly chosen[3] and as the data was generated under the Weibull distribution so it was expected to perform well. However, this means it was not possible to assess the performance of fitting other models which may be inappropriate in the presence of non-PH. This would allow one to understand whether incorrect model functions are likely to be chosen if PH is assumed. It would be useful to consider this work in the future to understand the likelihood of using alternative models which are inappropriate.

Due to time constraints, it was not possible to vary the “true” HR to less than 1.0 to reflect a beneficial treatment effect in favour of experimental arm compared to control which could be a limitation of the simulation study so only a no treatment effect scenario has been presented. This could be considered for future work to see whether varying the “true” HR has an impact on the performance of the four models.

A random-effects approach was used for conducting the meta-analysis which could be a further limitation. The initial plan was to fit fixed and random-effects models and compare the performance, however due to time constraints and some level of heterogeneity being present in all DGMs, the decision was made to only use a random-effects model. A comparison of the performance of the two models could be considered during future work.

Additionally, an extreme censoring level of around 65% was selected to highlight how the simulation study would perform in such a situation, although this would be very unlikely to happen in reality. In future possibly two lower censoring levels could be used to reflect current practice.

In order to create non-PHs, a time-dependent effect was included in the models at two different levels to create crossing curves. No alternative approach was considered to introduce non-PH which could be a further limitation. An alternative approach could be considered in future.

7.4.3 Concluding remarks

The results from the simulation study have confirmed how well all four models performed when the PH assumption was valid. However, also highlighted how poorly the Cox, Weibull and AFT models perform when a time-dependent effect is present and hence the PH assumption does not hold. The best performing model in all cases was the flexible parametric model. Overall, these simulation studies have given a chance to conduct in depth investigations of the behaviour of all survival models for meta-analysis of study data generated either assuming PH holds or not.

In the future, I recommend that as a first step the PH assumption is assessed so better decisions can be made on appropriate methods to use. Chapters 2 and 4 include details on how to assess the PH assumption when raw data is not available but a K-M plot and summary statistics are reported. If the K-M plot and summary statistics are not available then every effort should be made to contact the study authors to obtain the raw data or at least the K-M plot.

If the PH assumption is not valid but raw data is available then the flexible parametric model is a conservative approach to take and the best model to use to analyse TTE data. The choice of parameter (baseline hazard, level of censoring, time-dependent effect present or not, whether simulated datasets were homogeneous or heterogeneous, follow-up period etc) should be made based on the requirements of each individual investigation as one particular parameters value could prove to be more appropriate than another.

If it is known that the PH assumption is not valid but only the HR is reported then it is worth contacting the study authors for the K-M plots as the plots can be used to estimate the RMST which doesn't depend on the validity of the PH assumption.

8 Discussion and Conclusions

8.1 Summary of main findings of the thesis

A range of methods for individual studies and for the meta-analytic synthesis of TTE data have been proposed over several decades and applied to a wide range of clinical and methodological scenarios; Chapter 2 of this thesis presents a literature review of this methodology according to whether the Cox PH assumption is valid or not.

It is well documented within RCTs that the Cox PH model is the most commonly used method for analysing TTE data[30 31 35]. The first known review to investigate the reporting of TTE outcomes within individual studies was carried out by Altman *et al*[30] in the 1990s. Since then there has been two further reviews[31 35] conducted to assess the reporting of TTE outcomes in individual studies. Although there has been some improvement in the reporting of TTE analyses, further work is still required to ensure authors are not only assessing the PH assumption and reporting it, but are also explaining how they assessed the assumption and what appropriate action was taken if the PH assumption is invalid. However, the recent published research has highlighted that the PH assumption has only been assessed in around 4% - 26% of individual studies. Chapter 3 of this thesis presents a systematic review of the reporting of TTE outcomes within RCTs with no restriction on the disease area and includes results from a survey of current practice targeted at the UKCRC network of registered clinical trials units.

This is believed to be the first systematic review carried out since 2007[31] outside of the field of oncology, reflecting reporting standards over the past 13 years. In line with previous work[30 31 35], although the reporting of TTE outcomes has greatly improved with 93% of RCTs reporting what analysis method was used, the reporting of the PH assumption is still an issue.

Results from the UKCRC CTU survey in Chapter 3 reflected that over 90% of CTUs use the Cox PH regression method to analyse TTE data and although most of the CTUs said they would use alternative methods if the PH assumption was invalid, the decision to perform an alternative analysis is dependent on other factors such as survival curves cross and if there is a total absence of valid interpretation, despite whether the results are biased and based on incorrect assumptions. All of the CTUs mentioned that they do assess the PH assumption when using a method that assumes PH. It is unclear whether the CTUs mentioned this as they are telling me what they think I want to hear about what the correct approach is whether this is what they are doing in practice. Unfortunately, as the studies included in this review are only until 2013, and more recent publications are not included it is difficult to know what CTUs are doing in practice.

The necessary published information required to perform aggregate data meta-analysis of TTE data is often not reported or is reported inconsistently[30 31 35 113-116]. A range of accessible and user-friendly methods have been developed with the plan to make use of more commonly reported summary statistics and published survival curves to indirectly estimate HRs and associated variances[14 39]. Nevertheless, whether these methods can be used in practice has been questioned since many alternative summary statistics are also not reported or published graphical figures are of inadequate quality[14 114 117]. Chapter 4 of this thesis presents a novel systematic review of the reporting of TTE outcomes within meta-analyses.

This is believed to be the first systematic review to be conducted on the reporting of TTE outcomes within meta-analyses. All of the previous work has been conducted within individual trials with no clear understanding of how well TTE outcomes are reported in meta-analyses or how often the PH assumption is tested. All of the included systematic reviews reported on the analysis method of choice which was the Cox PH regression method. IPD meta-analyses were conducted in 66% of reviews with 28% of reviews using aggregate data and the remaining 6% of reviews including a mix of IPD and aggregate data. Although, the reporting of the TTE outcomes in the review was excellent, the reporting of the PH assumption was rather poor with only 27% of reviews assessing the PH assumption. These findings suggest that the results for the PH assumption assessment are poor across individual trials and within meta-analyses.

In 2012 Guyot et al[29] published a paper describing how to use published survival curves to obtain pseudo IPD. This approach was considered to investigate the PH assumption within the meta-analyses in Chapter 4. Chapter 5 of this thesis presents results from assessing the PH assumption using pseudo IPD to compare the results obtained to what has been carried out within the reviews.

In the 123 included trials only 76 (62%) reviews included K-M curves and only 47 of those reviews included numbers at risk enabling only those reviews to be included in the digitisation work. The reconstructed data demonstrated estimates close to the HRs reported in the systematic reviews. Evidence of non-PH was found in 10 (21%) of the reviews.

Chapter 6 highlights the results from assessing the reporting and assessment of the PH assumption within STAs by both companies and the ERGs in terms of clinical and cost effectiveness. Chapter 6 shows that around 70% of the companies are testing the PH assumption in STAs but in majority of the cases the ERGs are failing to independently assess the assumption. This became an issue in the clinical direct evidence for STAs where PH was found to be violated, the companies did not change the analysis approach and the ERG didn't perform the necessary checks.

Chapter 7 presents results from a novel simulation study that was undertaken to evaluate the performance of various different survival analysis models for meta-analysis of study data generated either assuming PH holds or not. Using RMST allowed the direct comparison of the four different survival models which otherwise would not have been possible with comparing HRs, TRs and HRs varying over time. The results indicated that where PH is valid, all models performed well with low bias, low MSE and high coverage. As soon as, a small time-dependent effect was introduced the Cox, Weibull and AFT models started to struggle with the mean pooled HR/TR moving away from the "true" estimate. The performance measures also demonstrated that the bias was higher and coverage was lower in comparison to the flexible parametric model which continued to do well. Similarly, once a larger time-dependent effect was introduced all models except the flexible parametric model struggled to perform well.

8.2 Implications for practice and research

The work of this thesis adds to the evidence base on the poor quality of the PH assumption reporting. The following recommendations can be used by trialists, reviewers and 'consumers' of reviews on how to approach the PH assumption based on the results from the thesis:

- For a trialist, if a HR is the treatment effect estimate of interest then a simple Cox PH model can be fitted to the data. Once the HR and 95% CI has been obtained it is important to assess the PH assumption which can be done visually by examining a K-M plot or even better by using a plot of the log cumulative hazard where the logarithm of time is plotted against the log cumulative hazard. If the curves for the two treatment groups are approximately parallel, the PH assumption is said to be valid. Although, this graphical approach is useful for visualising clear departures from PH, they can be subjective so a formal test may be preferred. To formally assess the PH assumption the simplest and most popular method as seen in Chapter 3 is the Schoenfeld residuals approach. Using this method if the p-value is statistically significant then the PH assumption is invalid.

  - If the PH assumption is valid then the HR and 95% CI can be presented along with a clear explanation to say how the PH assumption was tested and that the assumption is valid.
  - If the PH assumption is invalid then an alternative analysis should be presented. Alternative analyses include using a time-dependent treatment effect in the Cox model and presenting the results or using an approach not dependent on the validity of the PH assumption such as RMST, fitting an AFT model or using flexible parametric models, all methods seen and interpreted in Chapter 7. It would still be beneficial to present the HR and 95% CI but with additional details such as how the HR changes over time as well as clinical reasons behind why the HR is changing over time.
- For a reviewer performing a systematic review, if IPD is available then the methods mentioned above can be used for assessing the PH assumption. However, if only aggregate data is available then the PH assumption can be assessed using the reconstruction technique mentioned in Chapter 5. In order to perform the reconstruction technique the K-M plot and information on the number at risk would be needed in order to digitise the K-M data in order to obtain pseudo-IPD, which can be performed in the DigitiseIt software. The pseudo-IPD which includes time and survival for each curve will be merged with the data for time and number at risk which can be uploaded into Stata or R. Then depending on which software will be used, the Cox PH model can be fitted and then the PH assumption can be assessed using methods suggested above.

Currently in terms of reporting guidelines there is no mention of the PH assumption in the CONSORT guidelines[34] for RCTs or the PRISMA guidelines[118] for systematic reviews and meta-analyses. There is also no mention of the reporting of the PH assumption in the Cochrane Handbook[68] or in the ICH E9[119] guidelines. The only guidelines that are available are at the individual trial level in the form of a set of "minimum requirements" as suggested by Altman et al in 1995[30] and updated in 2013 by Abraira et al[31] and then the two NICE DSU TSDs[83 84] which are written to be used within STAs which are generally conducted by Pharmaceutical companies. Altman et al[30] suggested that "When Cox regression analyses are performed, describe the criteria used to select the variables in the initial model, the procedure to specify the final model and describe any methods used to assess the model assumptions." Abraira et al[31] updated the set of "minimum requirements" and suggested that "When using regression models, report the method used and results of model assumptions checking (e.g., the proportional hazards assumption in Cox models or distributional form in parametric models)." The first NICE DSU TSD published in 2011[83] states methods that can be used for assessing the PH assumption but the main focus is on the extrapolation of patient level data. The more recent TSD published in 2020[84] focuses more on methods that can be used on individual studies and to extrapolate that are not dependent on the PH assumption.

8.3 Limitations and future work

The first major limitation is that rapid review approaches have been used for reviewing the systematic reviews included in Chapter 4 as they were published between 2005 and 2015 and then the focused sample of RCTs included in Chapter 3 were published between 1985 and 2013. These reviews highlighted the poor reporting of the PH assumption in individual studies and within meta-analyses, however the set of "minimum requirements" updated by Abraira et al[31] weren't published until 2013. It is unclear whether this set of guidelines could have impacted the results obtained if the sample of studies included in Chapter 3 was based on individual studies published between 1995 and 2015 when both sets of "minimum requirements" were available. However, it is worth noting that the Batson et al[35] review was published in 2016 and includes RCTs published between April and July 2015 and they reported that only 7% of the publications reported the assessment of the PH assumption. Hence, it is unclear whether changing the inclusion criteria on publication years included would have altered the results on the reporting of the PH assumption.

Due to the number of reviews and individual studies included in the systematic reviews presented in this thesis, it was out of the scope of the work to contact original investigators individually to request additional or unpublished information but this could be worthwhile to expand our understanding of the reasons behind particular choices of methods of analysis, why certain practices are taken, and what would be needed to change practice. Such insights could help with updating the CONSORT guidelines[34] to include assumption checking in that more trials may adhere to as well as writing a guidance document which could be shared with experts in the field on how to assess the PH assumption and what to do if the assumption is invalid. Hence, improving the consistency of assumption checking and reporting across trials and facilitating synthesis of trials.

The second main limitation of the work in this thesis was in relation to the simulation study which was slightly restrictive as the data was simulated under a Weibull distribution followed by using a Weibull model for analysing the data as well as only using a random-effects approach. Using a Weibull distribution for simulating the data followed by using a Weibull model for analysing the data has meant that it was not possible to assess the performance of fitting other models which may be inappropriate in the presence of non-PH. This is important as this would allow one to understand whether incorrect model functions are likely to be chosen if PH is assumed.

It would be useful to consider using alternative scenarios in the future including simulating the data under an alternative distribution such as Gompertz, using a different sample of patients within each simulated study, varying the “true” HR so choosing a value where there is a beneficial treatment effect in favour of experimental arm compared to control, varying the number of studies included in the meta-analysis, using an alternative analysis method to the Weibull model to understand the likelihood of using alternative models which are inappropriate and possibly comparing the results from fitting a fixed effects and random effects model.

Lastly, another recommendation for future work would be to use the methods used within the simulation study and applying them to real data to explore the results obtained from the different analysis methods and especially when PH is invalid.

8.4 Concluding Remarks

The work of this thesis has provided a detailed insight into the assessment and reporting of the PH assumption within individual studies, meta-analyses and STAs and highlighted many inadequacies in the application and reporting of the PH assumption across a wide range of clinical disciplines. It is essential for clinical research of all sources, whether an original trial or synthesis, that the PH assumption is reported transparently. Whilst the HR remains the most commonly used effect measure for TTE outcomes, it is important that advantages of using alternative methods to the HR are reported so they can be used instead of the HR or in addition to the HR. Currently there are no guidelines published on the PH assumption reporting or mention of assumption checking in the CONSORT/PRISMA guidelines[34 118], the ICH E9 guidelines[119] or the Cochrane handbook[68].

References

1. Kane RL, Wang J, Garrard J. Reporting in randomized clinical trials improved after adoption of the CONSORT statement. *J Clin Epidemiol* 2007;60(3):241-9.

2. Keech A, Gebski V, Pike R. Interpreting and Reporting Clinical Trials. A Guide to the Consort Statement and Principles of Randomised Controlled Trials. *Australasian Medical Publishing Company* 2007

3. Collett D. Modelling Survival Data in Medical Research. *A Chapman & Hall Book* 2015;Third Edition

4. Clark TG, Bradburn MJ, Love SB, et al. Survival analysis part I: basic concepts and first analyses. *Br J Cancer* 2003;89(2):232-8.

5. Cox DR. Regression Models and Life-Tables. *Journal of the Royal Statistical Society Series B (Methodological)* 1972;34(2):187-220.

6. Advanced Bladder Cancer Meta-analysis C. Adjuvant chemotherapy in invasive bladder cancer: a systematic review and meta-analysis of individual patient data Advanced Bladder Cancer (ABC) Meta-analysis Collaboration. *European urology* 2005;48(2):189-99; discussion 99-201.

7. Perren TJ, Swart AM, Pfisterer J, et al. A phase 3 trial of bevacizumab in ovarian cancer. *The New England journal of medicine* 2011;365(26):2484-96.

8. Bradburn MJ, Clark TG, Love SB, et al. Survival analysis part II: multivariate data analysis--an introduction to concepts and methods. *Br J Cancer* 2003;89(3):431-6.

9. Pignon JP, Auperin A, Hill C. [Meta-analyses on individual patient data and treatment evaluation in oncology]. *Bulletin du cancer* 2007;94(11):957-64.

10. Gotzsche PC. Why we need a broad perspective on meta-analysis. It may be crucially important for patients. *BMJ* 2000;321(7261):585-6.

11. Salanti G. Indirect and mixed-treatment comparison, network, or multiple-treatments meta-analysis: many names, many benefits, many concerns for the next generation evidence synthesis tool. *Res Synth Methods* 2012;3(2):80-97.

12. Group. EBCTC. Statistical methods: Treatment of early breast cancer. *Oxford: Oxford University Press* 1990

13. Peto R. Why do we need systematic overviews of randomized trials? *Stat Med* 1987;6(3):233-44.

14. Parmar MKB, Torri V, Stewart L. Extracting summary statistics to perform meta-analyses of the published literature for survival endpoints. *Statistics in Medicine* 1998;17(24):2815-34.

15. Whitehead A, Whitehead J. A general parametric approach to the meta-analysis of randomized clinical trials. *Stat Med* 1991;10(11):1665-77.

16. OCEBM Levels of Evidence Working Group. The Oxford 2011 Levels of Evidence. *https://wwwcebmnet/indexaspx?o=5653 (accessed 09/01/2017)* 2011

17. Whitehead A. Meta-analysis of controlled clinical trials. Chichester ;: John Wiley & Sons 2002.

18. Yusuf S, Peto R, Lewis J, et al. Beta blockade during and after myocardial infarction: An overview of the randomized trials. *Progress in Cardiovascular Diseases* 1985;27(5):335-71.

19. Lin E, Tong T, Chen Y, et al. Fixed-effects model: the most convincing model for meta-analysis with few studies 2020.

20. Sutton AJ. Methods for meta-analysis in medical research: John Wiley 2000.

21. DerSimonian R, Laird N. Meta-analysis in clinical trials. *Control Clin Trials* 1986;7(3):177-88.

22. Williamson PR, Smith CT, Hutton JL, et al. Aggregate data meta-analysis with time-to-event outcomes. *Stat Med* 2002;21(22):3337-51.

23. 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.

24. Stewart LA, Tierney JF. To IPD or not to IPD? Advantages and disadvantages of systematic reviews using individual patient data. *Evaluation & the health professions* 2002;25(1):76-97.

25. Hua H, Burke DL, Crowther MJ, et al. One-stage individual participant data meta-analysis models: estimation of treatment-covariate interactions must avoid ecological bias by separating out within-trial and across-trial information. *Stat Med* 2017;36(5):772-89.

26. Caldwell DM, Ades AE, Higgins JP. Simultaneous comparison of multiple treatments: combining direct and indirect evidence. *Bmj* 2005;331(7521):897-900.

27. National Institute for H, Care E. NICE Process and Methods Guides. Guide to the Methods of Technology Appraisal 2013. London: National Institute for Health and Care Excellence (NICE) unless otherwise stated. All rights reserved. 2013.

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

29. Guyot P, Ades AE, Ouwens MJ, et al. Enhanced secondary analysis of survival data: reconstructing the data from published Kaplan-Meier survival curves. *BMC Med Res Methodol* 2012;12:9.

30. Altman DG, De Stavola BL, Love SB, et al. Review of survival analyses published in cancer journals. *British journal of cancer* 1995;72(2):511-18.

31. Abraira V, Muriel A, Emparanza JI, et al. Reporting quality of survival analyses in medical journals still needs improvement. A minimal requirements proposal. *Journal of Clinical Epidemiology* 2013;66(12):1340-46.

32. Begg C CM, Eastwood S, Horton R, Moher D, Olkin I, et al. Improving the quality of reporting of randomized controlled trials. The CONSORT statement. *JAMA* 1996;276(8):637-9.

33. Altman DG SK, Moher D, Egger M, Davidoff F, Elbourne D, et al. The Revised CONSORT Statement for Reporting Randomized Trials: Explanation and Elaboration. *Ann Intern Med* 2001;134(8):663-94.

34. Moher D, Hopewell S, Schulz KF, et al. CONSORT 2010 Explanation and Elaboration: updated guidelines for reporting parallel group randomised trials. *BMJ* 2010;340:c869.

35. Batson S, Greenall G, Hudson P. Review of the Reporting of Survival Analyses within Randomised Controlled Trials and the Implications for Meta-Analysis. *PLoS One* 2016;11(5):e0154870.

36. Rulli E, Ghilotti F, Biagioli E, et al. Assessment of proportional hazard assumption in aggregate data: a systematic review on statistical methodology in clinical trials using time-to-event endpoint. *Br J Cancer* 2018;119(12):1456-63.

37. Royston P, Parmar MKB. An approach to trial design and analysis in the era of non-proportional hazards of the treatment effect. *Trials* 2014;15(1):314. doi: 10.1186/1745-6215-15-314

38. Antoniou GA, Antoniou SA, Smith CT. A guide on meta-analysis of time-to-event outcomes using aggregate data in vascular and endovascular surgery. *Journal of Vascular Surgery* 2020;71(3):1002-05.

39. Tierney JF, Stewart LA, Ghersi D, et al. Practical methods for incorporating summary time-to-event data into meta-analysis. *Trials* 2007;8(1):16.

40. Bowden J, Tierney JF, Simmonds M, et al. 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. doi: 10.1002/jrsm.45

41. 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.

42. de Jong VMT, Moons KGM, Riley RD, et al. Individual participant data meta-analysis of intervention studies with time-to-event outcomes: A review of the methodology and an applied example. *Research synthesis methods* 2020;11(2):148-68.

43. Simmonds MC, Tierney J, Bowden J, et al. Meta-analysis of time-to-event data: a comparison of two-stage methods. *Research Synthesis Methods* 2011;2(3):139-49.

44. Royston P, Parmar MKB. Restricted mean survival time: an alternative to the hazard ratio for the design and analysis of randomized trials with a time-to-event outcome. *BMC Medical Research Methodology* 2013;13(1):152.

45. Siannis F, Barrett JK, Farewell VT, et al. One-stage parametric meta-analysis of time-to-event outcomes. *Statistics in Medicine* 2010;29(29):3030-45.

46. Barrett JK, Farewell VT, Siannis F, et al. Two-stage meta-analysis of survival data from individual participants using percentile ratios. *Statistics in Medicine* 2012;31(30):4296-308.

47. Bland JM, Altman DG. Survival probabilities (the Kaplan-Meier method). *BMJ* 1998;317(7172):1572.

48. Bland JM, Altman DG. The logrank test. *BMJ* 2004;328(7447):1073.

49. Bewick V, Cheek L, Ball J. Statistics review 12: survival analysis. *Critical care* 2004;8(5):389-94.

50. Schoenfeld D. Partial residuals for the proportional hazards regression model. *Biometrika* 1982;69(1):239-41.

51. Grambsch PM, Therneau TM. Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika* 1994;81(3):515-26.

52. Ng'andu NH. An empirical comparison of statistical tests for assessing the proportional hazards assumption of Cox's model. *Stat Med* 1997;16(6):611-26.

53. Harrell F. The PHGLM procedure, SAS supplemental Library User's Guide,. *SAS Institute, Cary, NC* 1986;Version 5 Edition

54. Lee L, Pirie R. A graphical method for comparing trends in series of events. *Communications in Statistics - Theory and Methods* 1981;10(9):827-48.

55. Liang F, Zhang S, Wang Q, et al. Treatment effects measured by restricted mean survival time in trials of immune checkpoint inhibitors for cancer. *Annals of oncology : official journal of the European Society for Medical Oncology* 2018;29(5):1320-24.

56. Trinquart L, Jacot J, Conner SC, et al. Comparison of Treatment Effects Measured by the Hazard Ratio and by the Ratio of Restricted Mean Survival Times in Oncology Randomized Controlled Trials. *Journal of Clinical Oncology* 2016;34(15):1813-19.

57. 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. *Stat Med* 2002;21(15):2175-97.

58. Schemper M, Wakounig S, Heinze G. The estimation of average hazard ratios by weighted Cox regression. *Stat Med* 2009;28(19):2473-89.

59. Kalbfleisch JD, Prentice RL. Estimation of the average hazard ratio. *Biometrika* 1981;68(1):105-12.

60. Royston P, Parmar MKB. The use of restricted mean survival time to estimate the treatment effect in randomized clinical trials when the proportional hazards assumption is in doubt. *Statistics in Medicine* 2011;30(19):2409-21.

61. Irwin JO. The standard error of an estimate of expectation of life, with special reference to expectation of tumourless life in experiments with mice. *J Hyg (Lond)* 1949;47(2):188-88.

62. Andersen PK, Hansen MG, Klein JP. Regression Analysis of Restricted Mean Survival Time Based on Pseudo-Observations. *Lifetime Data Analysis* 2004;10(4):335-50.

63. Andersen PK, Perme MP. Pseudo-observations in survival analysis. *Stat Methods Med Res* 2010;19(1):71-99.

64. Overgaard M, Andersen PK, Parner ET. Regression Analysis of Censored Data Using Pseudo-observations: An Update. *The Stata Journal* 2015;15(3):809-21.

65. Parner ET, Andersen PK. Regression Analysis of Censored Data Using Pseudo-observations. *The Stata Journal* 2010;10(3):408-22.

66. Wei Y, Royston P, Tierney JF, et al. Meta-analysis of time-to-event outcomes from randomized trials using restricted mean survival time: application to individual participant data. *Stat Med* 2015;34(21):2881-98.

67. Hutton JL, Williamson PR. Bias in meta-analysis due to outcome variable selection within studies. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 2000;49(3):359-70.

68. Higgins JPT, Thomas J, Chandler J, et al. Cochrane Handbook for Systematic Reviews of Interventions version 6.1 (updated September 2020). *Cochrane,* 2020;Available from www.training.cochrane.org/handbook.

69. StataCorp. Stata Statistical Software: Release 14. College Station, TX: StataCorp LP. 2015

70. Team RC. R: A language and environment for statistical computing. . *R Foundation for Statistical Computing, Vienna, Austria* 2017;Version 3.4.0:https://www.R-project.org/.

71. Simmonds MC, Higgins JP, Stewart LA, et al. Meta-analysis of individual patient data from randomized trials: a review of methods used in practice. *Clin Trials* 2005;2(3):209-17.

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

73. Higgins JP, Thompson SG. Quantifying heterogeneity in a meta-analysis. *Stat Med* 2002;21(11):1539-58.

74. Moore MJ, Goldstein D, Hamm J, et al. Erlotinib plus gemcitabine compared with gemcitabine alone in patients with advanced pancreatic cancer: a phase III trial of the National Cancer Institute of Canada Clinical Trials Group. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* 2007;25(15):1960-6.

75. Ogawa H, Nakayama M, Morimoto T, et al. Low-dose aspirin for primary prevention of atherosclerotic events in patients with type 2 diabetes: a randomized controlled trial. *Jama* 2008;300(18):2134-41.

76. Raitt MH, Connor WE, Morris C, et al. Fish oil supplementation and risk of ventricular tachycardia and ventricular fibrillation in patients with implantable defibrillators: a randomized controlled trial. *Jama* 2005;293(23):2884-91.

77. Solomon SD, Rice MM, K AJ, et al. Renal function and effectiveness of angiotensin-converting enzyme inhibitor therapy in patients with chronic stable coronary disease in the Prevention of Events with ACE inhibition (PEACE) trial. *Circulation* 2006;114(1):26-31.

78. The Allhat Officers Coordinators for the Allhat Collaborative Research Group. Major Outcomes in High-Risk Hypertensive Patients Randomized to Angiotensin-Converting Enzyme Inhibitor or Calcium Channel Blocker vs DiureticThe Antihypertensive and Lipid-Lowering Treatment to Prevent Heart Attack Trial (ALLHAT). *JAMA* 2002;288(23):2981-97.

79. Wing LM, Reid CM, Ryan P, et al. A comparison of outcomes with angiotensin-converting--enzyme inhibitors and diuretics for hypertension in the elderly. *The New England journal of medicine* 2003;348(7):583-92.

80. Wei Y, Royston P. Reconstructing time-to-event data from published Kaplan-Meier curves. *The Stata journal* 2017;17(4):786-802.

81. Lyman GH, Kuderer NM. The strengths and limitations of meta-analyses based on aggregate data. *BMC Med Res Methodol* 2005;5:14.

82. https://www.cochranelibrary.com/.

83. Latimer N. NICE DSU Technical Support Document 14: Undertaking survival analysis for economic evaluations alongside clinical trials - extrapolation with patient-level data. 2011 doi: Available from http://www.nicedsu.org.uk

84. Rutherford M, Lambert P, Sweeting M, et al. NICE DSU Technical Support Document 21. Flexible Methods for Survival Analysis. 2020 doi: Available from http://www.nicedsu.org.uk

85. Simmonds M, Stewart G, Stewart L. A decade of individual participant data meta-analyses: A review of current practice. *Contemp Clin Trials* 2015;45(Pt A):76-83.

86. Brouwer IAR, M. H.;Dullemeijer, C.;Kraemer, D. F.;Zock, P. L.;Morris, C.;Katan, M. B.;Connor, W. E.;Camm, J. A.;Schouten, E. G.;McAnulty, J. Effect of fish oil on ventricular tachyarrhythmia in three studies in patients with implantable cardioverter defibrillators *European Heart Journal* 2009; 30(7). http://onlinelibrary.wiley.com/o/cochrane/cldare/articles/DARE-12009104937/frame.html.

87. Ronellenfitsch US, Matthias;Hofheinz, Ralf;Kienle, Peter;Kieser, Meinhard;Slanger Tracy, E.;Jensen, Katrin. Perioperative chemo(radio)therapy versus primary surgery for resectable adenocarcinoma of the stomach, gastroesophageal junction, and lower esophagus. *Cochrane Database of Systematic Reviews* 2013; (5). http://onlinelibrary.wiley.com/doi/10.1002/14651858.CD008107.pub2/abstract.

88. Unverzagt SB, Michael;de Waha, Antoinette;Haerting, Johannes;Pietzner, Diana;Seyfarth, Melchior;Thiele, Holger;Werdan, Karl;Zeymer, Uwe;Prondzinsky, Roland. Intra-aortic balloon pump counterpulsation (IABP) for myocardial infarction complicated by cardiogenic shock. *Cochrane Database of Systematic Reviews* 2015; (3). http://onlinelibrary.wiley.com/doi/10.1002/14651858.CD007398.pub3/abstract.

89. Huang Y, Mao C, Yuan J, et al. Distribution and Epidemiological Characteristics of Published Individual Patient Data Meta-Analyses. *PLOS ONE* 2014;9(6):e100151.

90. Riley RD, Lambert PC, Abo-Zaid G. Meta-analysis of individual participant data: rationale, conduct, and reporting. *BMJ* 2010;340:c221.

91. Bradburn MJ, Clark TG, Love SB, et al. Survival Analysis Part III: Multivariate data analysis – choosing a model and assessing its adequacy and fit. *British Journal of Cancer* 2003;89(4):605-11.

92. Gamble C, Krishan A, Stocken D, et al. Guidelines for the Content of Statistical Analysis Plans in Clinical Trials. *JAMA* 2017;318(23):2337-43.

93. Ardizzoni AB, L.;Tiseo, M.;Fossella, F. V.;Schiller, J. H.;Paesmans, M.;Radosavljevic, D.;Paccagnella, A.;Zatloukal, P.;Mazzanti, P.;Bisset, D.;Rosell, R. Cisplatin- versus carboplatin-based chemotherapy in first-line treatment of advanced non-small-cell lung cancer: an individual patient data meta-analysis (Structured abstract). *J Natl Cancer Inst* 2007; 99(11). http://onlinelibrary.wiley.com/o/cochrane/cldare/articles/DARE-12007005734/frame.html.

94. Ardizzoni A, Boni L, Tiseo M, et al. Cisplatin- versus carboplatin-based chemotherapy in first-line treatment of advanced non-small-cell lung cancer: an individual patient data meta-analysis. *J Natl Cancer Inst* 2007;99(11):847-57.

95. Monnickendam G, Zhu M, McKendrick J, et al. Measuring Survival Benefit in Health Technology Assessment in the Presence of Nonproportional Hazards. *Value in Health* 2019;22(4):431-38.

96. Burton A, Altman DG, Royston P, et al. The design of simulation studies in medical statistics. *Statistics in Medicine* 2006;25(24):4279-92. doi: 10.1002/sim.2673

97. Crowther MJ, Lambert PC. Simulating biologically plausible complex survival data. *Statistics in Medicine* 2013;32(23):4118-34.

98. Morris TP, White IR, Crowther MJ. Using simulation studies to evaluate statistical methods. *Statistics in Medicine* 2019;38(11):2074-102.

99. Brilleman S, Gasparini A. Simsurv: Simulating complex survival data. . *R package* 2019;version 0.2.3

100. Bender R, Augustin T, Blettner M. Generating survival times to simulate Cox proportional hazards models. *Statistics in Medicine* 2005;24(11):1713-23.

101. Smyth G HY, Dunn P, Phipson B, Yunshun C. Statistical Modelling. *R package* 2019;version 1.4.32

102. Lambert PC, Royston P. Further Development of Flexible Parametric Models for Survival Analysis. *The Stata Journal* 2009;9(2):265-90.

103. Colzani E LA, Johansson ALV, Adolfsson J, Hellborg H, Hall PFL, Czene K. Prognosis of patients with breast cancer: causes of death and effects of time since diagnosis, age, and tumor characteristics. *Journal of Clinical Oncology* 2011;29:4014-21.

104. McMinn DJW, Snell KIE, Daniel J, et al. Mortality and implant revision rates of hip arthroplasty in patients with osteoarthritis: registry based cohort study. *BMJ* 2012;344:e3319.

105. https://research.cs.wisc.edu/htcondor/.

106. http://condor.liv.ac.uk/.

107. Bowden J, Tierney JF, Simmonds M, et al. Individual patient data meta-analysis of time-to-event outcomes: one-stage versus two-stage approaches for estimating the hazard ratio under a random effects model. *Res Synth Methods* 2011;2(3):150-62.

108. Debray TP, Moons KG, van Valkenhoef G, et al. Get real in individual participant data (IPD) meta-analysis: a review of the methodology. *Res Synth Methods* 2015;6(4):293-309.

109. Fisher DJ. Two-stage individual participant data meta-analysis and generalized forest plots. *The Stata Journal* 2015;15:369-96.

110. White IR. Multivariate random-effects meta-regression: Updates to mvmeta. *The Stata Journal* 2011;11:255-70.

111. Oehlert GW. A Note on the Delta Method. *The American Statistician* 1992;46(1):27-29.

112. White IR. simsum: Analyses of simulation studies including Monte Carlo error. *Stata Journal* 2010;10(3):369-85.

113. Arkenau HT, Nordman I, Dobbins T, et al. Reporting time-to-event endpoints and response rates in 4 decades of randomized controlled trials in advanced colorectal cancer. *Cancer* 2011;117(4):832-40.

114. 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.

115. Mathoulin-Pelissier S, Gourgou-Bourgade S, Bonnetain F, et al. Survival end point reporting in randomized cancer clinical trials: a review of major journals. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* 2008;26(22):3721-6.

116. Michiels S, Piedbois P, Burdett S, et al. 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.

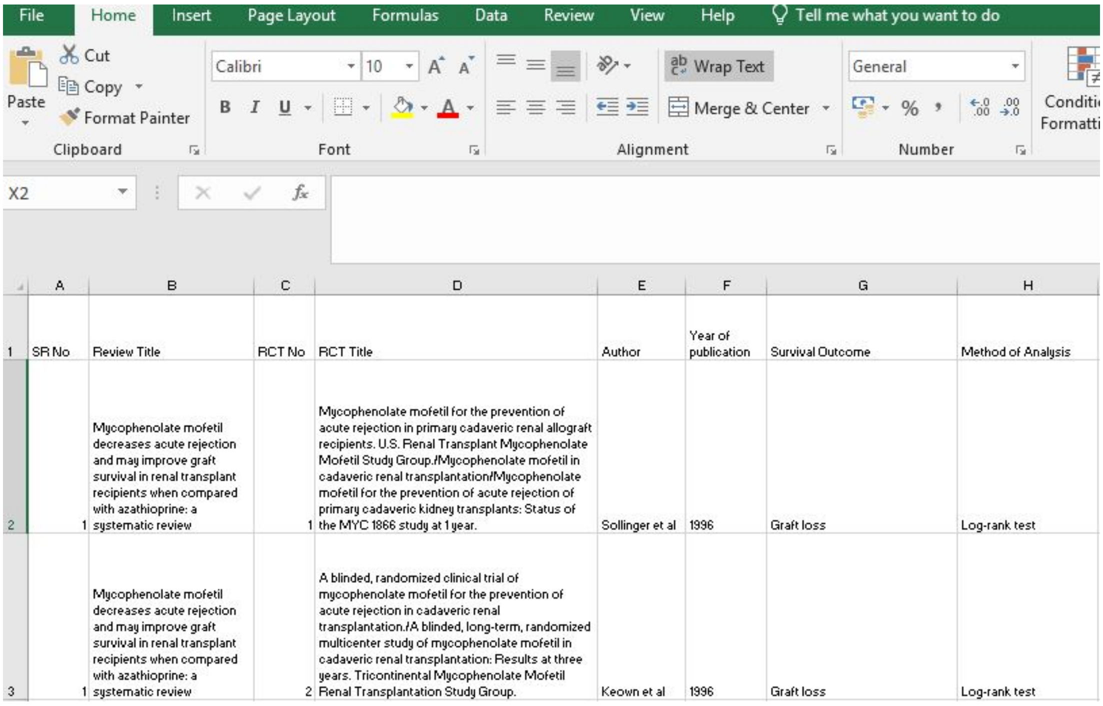
117. Tudur C, Williamson PR, Khan S, et al. The value of the aggregate data approach in meta-analysis with time-to-event outcomes. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 2001;164(2):357-70.

118. Moher D, Liberati A, Tetzlaff J, et al. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *PLoS medicine* 2009;6(7):e1000097.

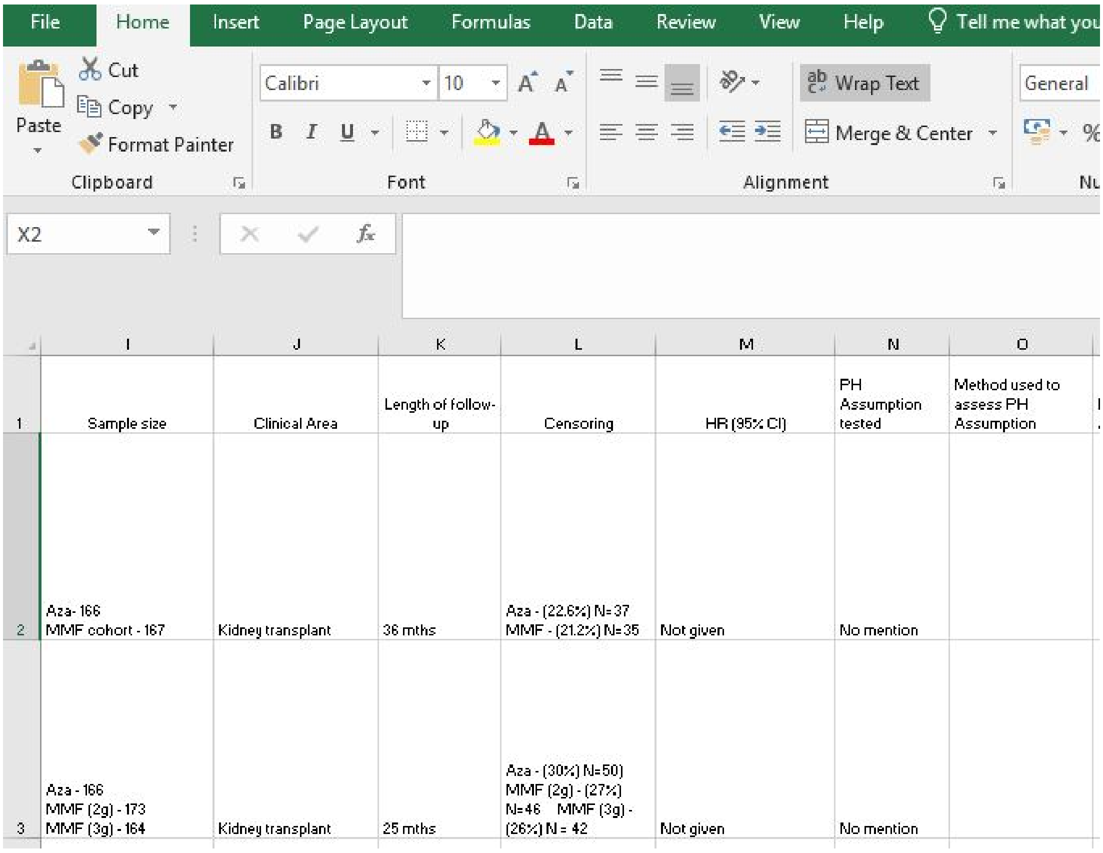
119. International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use. *ICH Harmonised Tripartite Guideline: Statistical Principles for Clinical Trials E9*;London, England: European Medicines Agency; 1998

Appendices

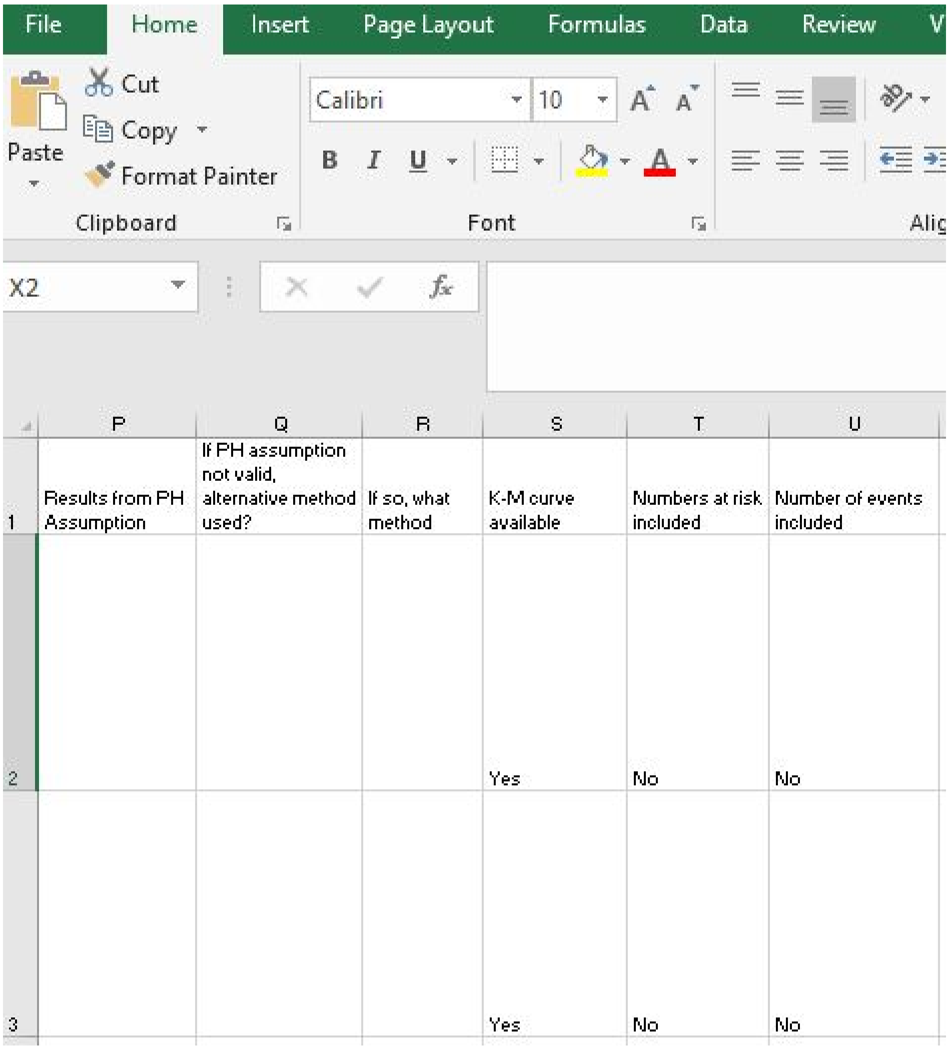
Appendix 1: Screenshot of Microsoft Excel Database used in Chapter 3



	A	B	C	D	E	F	G	H
1	SR No	Review Title	RCT No	RCT Title	Author	Year of publication	Survival Outcome	Method of Analysis
2	1	Mycophenolate mofetil decreases acute rejection and may improve graft survival in renal transplant recipients when compared with azathioprine: a systematic review	1	Mycophenolate mofetil for the prevention of acute rejection in primary cadaveric renal allograft recipients. U.S. Renal Transplant Mycophenolate Mofetil Study Group./Mycophenolate mofetil in cadaveric renal transplantation/Mycophenolate mofetil for the prevention of acute rejection of primary cadaveric kidney transplants: Status of the MYC 1866 study at 1 year.	Sollinger et al	1996	Graft loss	Log-rank test
3	1	Mycophenolate mofetil decreases acute rejection and may improve graft survival in renal transplant recipients when compared with azathioprine: a systematic review	2	A blinded, randomized clinical trial of mycophenolate mofetil for the prevention of acute rejection in cadaveric renal transplantation./A blinded, long-term, randomized multicenter study of mycophenolate mofetil in cadaveric renal transplantation: Results at three years. Tricontinental Mycophenolate Mofetil Renal Transplantation Study Group.	Keown et al	1996	Graft loss	Log-rank test



	I	J	K	L	M	N	O
1	Sample size	Clinical Area	Length of follow-up	Censoring	HR (95% CI)	PH Assumption tested	Method used to assess PH Assumption
2	Aza- 166 MMF cohort - 167	Kidney transplant	36 mths	Aza - (22.6%) N=37 MMF - (21.2%) N=35	Not given	No mention	
3	Aza - 166 MMF (2g) - 173 MMF (3g) - 164	Kidney transplant	25 mths	Aza - (30%) N=50) MMF (2g) - (27%) N=46 MMF (3g) - (26%) N = 42	Not given	No mention	



	P	Q	R	S	T	U
1	Results from PH Assumption	If PH assumption not valid, alternative method used?	If so, what method	K-M curve available	Numbers at risk included	Number of events included
2				Yes	No	No
3				Yes	No	No

Appendix 2: Copy of the survey: Analysis of time-to-event data in clinical trials: A survey of current practice

The following questions are about methods for analysing time-to-event outcomes in Randomised Controlled Trials (RCTs) and current practice within your Clinical Trials Unit (CTU). In particular we are interested in knowing how you deal with the Proportional Hazards (PH) assumption. We would also be grateful if your responses could be made on behalf of your CTU, so it may help to discuss the survey with your colleagues before returning it. The survey should take 5-10 minutes to complete.

1. What method do you use for the analysis of time-to-event (TTE) outcomes? *Please tick all that apply. (See below for a glossary of the different methods)*

Cox Proportional Hazards (PH) regression

Never ☐ Sometimes ☐ Frequently ☐

Kaplan-Meier Method

Never ☐ Sometimes ☐ Frequently ☐

Log-rank test

Never ☐ Sometimes ☐ Frequently ☐

Parametric PH model

Never ☐ Sometimes ☐ Frequently ☐

Accelerated Failure Time model

Never ☐ Sometimes ☐ Frequently ☐

Flexible parametric model

Never ☐ Sometimes ☐ Frequently ☐

Other (*Please specify*) ________________________________

2. If using a method that assumes PH which methods are used most commonly to assess the PH assumption? *Please tick all that apply. (See below for a glossary of the different methods)*

Assumption not assessed ☐

Kaplan-Meier plots ☐

Log-cumulative plot ☐

Schoenfeld residuals ☐

Lee and Pirie method (also known as H-H plots) ☐

Time-dependent covariates ☐

Other (*Please specify*) ____

3. If the PH assumption is invalid, what approach would you take for the analysis?

Ignore the assumption and still use the PH model ☐

Use an alternative method that does not assume PH ☐

- Please specify ____

Other strategy for analysis (*Please specify*)________________________________

Comments____

____

Thank you so much for taking the time to complete this survey. Your views are important, and we are grateful for your support. If you have any questions on any aspect of this survey then please do not hesitate to contact me.

Thank You,

Ashma Krishan

<u>Glossary of terms</u>

Term	Definition
Proportional Hazards	*Assuming that the hazard rates for the intervention groups are proportional over time.*
Cox PH regression	*An approach to explore the relationship between the survival experience of a patient and explanatory variable, which is dependent on the PH assumption being valid.*
Kaplan-Meier (K-M) method	*The K-M method estimates the survival probability nonparametrically from observed survival times which are both censored and uncensored.*
Log-rank test	*The log-rank test is used to test the null hypothesis that there is no difference between the intervention groups in the probability of an event (such as death or relapse) occurring at any time point. The observed number of events in each intervention group along with the expected number*

	of events are calculated under the null hypothesis of no difference between the intervention groups.
Parametric PH model	*Parametric PH models are similar in concept and interpretation to the Cox PH regression model, except parametric models follow a specific statistical distribution.*
Accelerated Failure Time (AFT) model	*The AFT model is a model for the analysis of survival data where the covariates measured for an individual are expected to act multiplicatively on the time-scale, so in other words say for the covariate treatment, the length of survival is either increasing or decreasing in the new treatment group compared to the standard treatment group. This method is not dependent on the PH assumption.*
Flexible Parametric model	*An approach which models the underlying baseline hazard, but allows the function to have greater flexibility than that allowed by the fully parametric models.*
Schoenfeld Residuals	*The Schoenfeld residuals is a graphical assessment of the PH assumption, but a graphical summary that tests the covariates for time-dependence. The Schoenfeld residuals take a set of values, so one set for each covariate included in the fitted Cox regression model.*
Time-dependent covariates	*A method for assessing departures from PH by introducing a time-dependent covariate to the Cox regression model. The time-dependent covariate is added to the model by adding an interaction term that involves time, e.g. age x log(t) and testing for significance*

Appendix 3: References of 106 RCTs included in Chapter 3

1. Sollinger HW. Mycophenolate mofetil for the prevention of acute rejection in primary cadaveric renal allograft recipients. U.S. Renal Transplant Mycophenolate Mofetil Study Group. Transplantation 1995;60(3):225-32.

2. The Tricontinental Mycophenolate Mofetil Renal Transplantation Study Group. A blinded, randomized clinical trial of mycophenolate mofetil for the prevention of acute rejection in cadaveric renal transplantation. Transplantation 1996;61(7):1029-37.

3. Ahsan N, Johnson C, Gonwa T, et al. Randomized trial of tacrolimus plus mycophenolate mofetil or azathioprine versus cyclosporine oral solution (modified) plus mycophenolate mofetil after cadaveric kidney transplantation: results at 2 years. Transplantation 2001;72(2):245-50.

4. Joh JW, Lee HH, Lee DS, et al. The influence of mycophenolate mofetil and azathioprine on the same cadaveric donor renal transplantation. J Korean Med Sci 2005;20(1):79-81.

5. Sadek S, Medina J, Arias M, et al. Short-term combination of mycophenolate mofetil with cyclosporine as a therapeutic option for renal transplant recipients: A prospective, multicenter, randomized study. Transplantation 2002;74(4):511-7.

6. Tuncer M, Gürkan A, Erdoğan O, et al. Mycophenolate mofetil in renal transplantation: five years experience. Transplantation proceedings 2002;34(6):2087-8.

7. Boku N, Yamamoto S, Fukuda H, et al. Fluorouracil versus combination of irinotecan plus cisplatin versus S-1 in metastatic gastric cancer: a randomised phase 3 study. The Lancet Oncology 2009;10(11):1063-9.

8. Komatsu Y, Takahashi Y, Kimura Y, et al. Randomized phase II trial of first-line treatment with tailored irinotecan and S-1 therapy versus S-1 monotherapy for advanced or recurrent gastric carcinoma (JFMC31-0301). Anti-cancer drugs 2011;22(6):576-83.

9. Bouché O, Raoul JL, Bonnetain F, et al. Randomized multicenter phase II trial of a biweekly regimen of fluorouracil and leucovorin (LV5FU2), LV5FU2 plus cisplatin, or LV5FU2 plus irinotecan in patients with previously untreated metastatic gastric cancer: a Federation Francophone de Cancerologie Digestive Group Study--FFCD 9803. Journal of clinical oncology : official journal of the American Society of Clinical Oncology 2004;22(21):4319-28.

10. Moehler M, Eimermacher A, Siebler J, et al. Randomised phase II evaluation of irinotecan plus high-dose 5-fluorouracil and leucovorin (ILF) vs 5-fluorouracil, leucovorin, and etoposide (ELF) in untreated metastatic gastric cancer. British Journal of Cancer 2005;92(12):2122-28.

11. Mutch DG, Orlando M, Goss T, et al. Randomized phase III trial of gemcitabine compared with pegylated liposomal doxorubicin in patients with platinum-resistant ovarian cancer. Journal of clinical oncology : official journal of the American Society of Clinical Oncology 2007;25(19):2811-8.

12. Pignata S, Scambia G, Ferrandina G, et al. Carboplatin plus paclitaxel versus carboplatin plus pegylated liposomal doxorubicin as first-line treatment for patients with ovarian cancer: the MITO-2 randomized phase III trial. Journal of clinical oncology : official journal of the American Society of Clinical Oncology 2011;29(27):3628-35.

13. Bookman MA, Brady MF, McGuire WP, et al. Evaluation of new platinum-based treatment regimens in advanced-stage ovarian cancer: a Phase III Trial of the Gynecologic Cancer Intergroup. Journal of clinical oncology : official journal of the American Society of Clinical Oncology 2009;27(9):1419-25.

14. Kaye SB, Lubinski J, Matulonis U, et al. Phase II, open-label, randomized, multicenter study comparing the efficacy and safety of olaparib, a poly (ADP-ribose) polymerase inhibitor, and pegylated liposomal doxorubicin in patients with BRCA1 or BRCA2 mutations and recurrent ovarian cancer. Journal of clinical oncology : official journal of the American Society of Clinical Oncology 2012;30(4):372-9.

15. Alberts DS, Liu PY, Wilczynski SP, et al. Randomized trial of pegylated liposomal doxorubicin (PLD) plus carboplatin versus carboplatin in platinum-sensitive (PS) patients with recurrent epithelial ovarian or peritoneal carcinoma after failure of initial platinum-based chemotherapy (Southwest Oncology Group Protocol S0200). Gynecologic oncology 2008;108(1):90-4.

16. Colombo N, Kutarska E, Dimopoulos M, et al. Randomized, open-label, phase III study comparing patupilone (EPO906) with pegylated liposomal doxorubicin in platinum-refractory or -resistant patients with recurrent epithelial ovarian, primary fallopian tube, or primary peritoneal cancer. Journal of clinical oncology : official journal of the American Society of Clinical Oncology 2012;30(31):3841-7.

17. Burger RA, Brady MF, Bookman MA, et al. Incorporation of Bevacizumab in the Primary Treatment of Ovarian Cancer. New England Journal of Medicine 2011;365(26):2473-83.

18. Perren TJ, Swart AM, Pfisterer J, et al. A phase 3 trial of bevacizumab in ovarian cancer. The New England journal of medicine 2011;365(26):2484-96.

19. Aghajanian C, Blank SV, Goff BA, et al. OCEANS: a randomized, double-blind, placebo-controlled phase III trial of chemotherapy with or without bevacizumab in patients with platinum-sensitive recurrent epithelial ovarian, primary peritoneal, or fallopian tube cancer. Journal of clinical oncology : official journal of the American Society of Clinical Oncology 2012;30(17):2039-45.

20. Pujade-Lauraine E, Hilpert F, Weber B, et al. AURELIA: A randomized phase III trial evaluating bevacizumab (BEV) plus chemotherapy (CT) for platinum (PT)-resistant recurrent ovarian cancer (OC). Journal of Clinical Oncology 2012;30(18_suppl):LBA5002-LBA02.

21. Delaney CP, Weese JL, Hyman NH, et al. Phase III Trial of Alvimopan, a Novel, Peripherally Acting, Mu Opioid Antagonist, for Postoperative Ileus After Major Abdominal Surgery. Diseases of the Colon & Rectum 2005;48(6):1114-29.

22. Herzog TJ, Coleman RL, Guerrieri JP, Jr., et al. A double-blind, randomized, placebo-controlled phase III study of the safety of alvimopan in patients who undergo simple total abdominal hysterectomy. American journal of obstetrics and gynecology 2006;195(2):445-53.

23. Viscusi ER, Goldstein S, Witkowski T, et al. Alvimopan, a peripherally acting mu-opioid receptor antagonist, compared with placebo in postoperative ileus after major abdominal surgery: results of a randomized, double-blind, controlled study. Surgical endoscopy 2006;20(1):64-70.

24. Wolff BG, Michelassi F, Gerkin TM, et al. Alvimopan, a novel, peripherally acting mu opioid antagonist: results of a multicenter, randomized, double-blind, placebo-controlled, phase III trial of major abdominal surgery and postoperative ileus. Annals of surgery 2004;240(4):728-34.

25. Stöckle M, Meyenburg W, Wellek S, et al. Adjuvant polychemotherapy of nonorgan-confined bladder cancer after radical cystectomy revisited: long-term results of a controlled prospective study and further clinical experience. The Journal of urology 1995;153(1):47-52.

26. Freiha F, Reese J, Torti FM. A Randomized Trial of Radical Cystectomy Versus Radical Cystectomy Plus Cisplatin, Vinblastine and Methotrexate Chemotherapy for Muscle Invasive Bladder Cancer. The Journal of urology 1996;155(2):495-500.

27. The EPIC Investigators. Use of a Monoclonal Antibody Directed against the Platelet Glycoprotein IIb/IIIa Receptor in High-Risk Coronary Angioplasty. New England Journal of Medicine 1994;330(14):956-61.

28. EPILOG Investigators. Platelet glycoprotein IIb/IIIa receptor blockade and low-dose heparin during percutaneous coronary revascularization. The New England journal of medicine 1997;336(24):1689-96.

29. EPISTENT Investigators. Randomised placebo-controlled and balloon-angioplasty-controlled trial to assess safety of coronary stenting with use of platelet glycoprotein-IIb/IIIa blockade. Lancet 1998;352(9122):87-92.

30. CAPTURE Investigators. Randomised placebo-controlled trial of abciximab before and during coronary intervention in refractory unstable angina: the CAPTURE Study. Lancet 1997;349(9063):1429-35.

31. Brener Sorin J, Barr Lawrence A, Burchenal JEB, et al. Randomized, Placebo-Controlled Trial of Platelet Glycoprotein IIb/IIIa Blockade With Primary Angioplasty for Acute Myocardial Infarction. Circulation 1998;98(8):734-41.

32. IMPACT-II Investigators. Randomised placebo-controlled trial of effect of eptifibatide on complications of percutaneous coronary intervention: IMPACT-II. Integrilin to Minimise Platelet Aggregation and Coronary Thrombosis-II. Lancet 1997;349(9063):1422-8.

33. Topol EJ, Moliterno DJ, Herrmann HC, et al. Comparison of two platelet glycoprotein IIb/IIIa inhibitors, tirofiban and abciximab, for the prevention of ischemic events with percutaneous coronary revascularization. The New England journal of medicine 2001;344(25):1888-94.

34. Steinhubl SR, Berger PB, Mann JT, 3rd, et al. Early and sustained dual oral antiplatelet therapy following percutaneous coronary intervention: a randomized controlled trial. Jama 2002;288(19):2411-20.

35. Marson AG, Al-Kharusi AM, Alwaidh M, et al. The SANAD study of effectiveness of carbamazepine, gabapentin, lamotrigine, oxcarbazepine, or topiramate for treatment of partial epilepsy: an unblinded randomised controlled trial. Lancet 2007;369(9566):1000-15.

36. Brodie MJ, Overstall PW, Giorgi L. Multicentre, double-blind, randomised comparison between lamotrigine and carbamazepine in elderly patients with newly diagnosed epilepsy. The UK Lamotrigine Elderly Study Group. Epilepsy Res 1999;37(1):81-7.

37. Herrmann R, Bodoky G, Ruhstaller T, et al. Gemcitabine plus capecitabine compared with gemcitabine alone in advanced pancreatic cancer: a randomized, multicenter, phase III trial of the Swiss Group for Clinical Cancer Research and the Central European Cooperative Oncology Group. Journal of clinical oncology : official journal of the American Society of Clinical Oncology 2007;25(16):2212-7.

38. Cunningham D, Chau I, Stocken DD, et al. Phase III randomized comparison of gemcitabine versus gemcitabine plus capecitabine in patients with advanced pancreatic cancer. Journal of clinical oncology : official journal of the American Society of Clinical Oncology 2009;27(33):5513-8.

39. Moore MJ, Goldstein D, Hamm J, et al. Erlotinib plus gemcitabine compared with gemcitabine alone in patients with advanced pancreatic cancer: a phase III trial of the

National Cancer Institute of Canada Clinical Trials Group. Journal of clinical oncology : official journal of the American Society of Clinical Oncology 2007;25(15):1960-6.

40. Boeck S, Hoehler T, Seipelt G, et al. Capecitabine plus oxaliplatin (CapOx) versus capecitabine plus gemcitabine (CapGem) versus gemcitabine plus oxaliplatin (mGemOx): final results of a multicenter randomized phase II trial in advanced pancreatic cancer. Annals of oncology : official journal of the European Society for Medical Oncology 2008;19(2):340-7.

41. Heinemann V, Quietzsch D, Gieseler F, et al. Randomized phase III trial of gemcitabine plus cisplatin compared with gemcitabine alone in advanced pancreatic cancer. Journal of clinical oncology : official journal of the American Society of Clinical Oncology 2006;24(24):3946-52.

42. Conroy T, Desseigne F, Ychou M, et al. FOLFIRINOX versus gemcitabine for metastatic pancreatic cancer. The New England journal of medicine 2011;364(19):1817-25.

43. Colucci G, Labianca R, Di Costanzo F, et al. Randomized phase III trial of gemcitabine plus cisplatin compared with single-agent gemcitabine as first-line treatment of patients with advanced pancreatic cancer: the GIP-1 study. Journal of clinical oncology : official journal of the American Society of Clinical Oncology 2010;28(10):1645-51.

44. Louvet C, Labianca R, Hammel P, et al. Gemcitabine in combination with oxaliplatin compared with gemcitabine alone in locally advanced or metastatic pancreatic cancer: results of a GERCOR and GISCAD phase III trial. Journal of clinical oncology : official journal of the American Society of Clinical Oncology 2005;23(15):3509-16.

45. Nakai Y, Isayama H, Sasaki T, et al. A multicentre randomised phase II trial of gemcitabine alone vs gemcitabine and S-1 combination therapy in advanced pancreatic cancer: GEMSAP study. Br J Cancer 2012;106(12):1934-9.

46. Ozaka M, Matsumura Y, Ishii H, et al. Randomized phase II study of gemcitabine and S-1 combination versus gemcitabine alone in the treatment of unresectable advanced pancreatic cancer (Japan Clinical Cancer Research Organization PC-01 study). Cancer chemotherapy and pharmacology 2012;69(5):1197-204.

47. Poplin E, Feng Y, Berlin J, et al. Phase III, randomized study of gemcitabine and oxaliplatin versus gemcitabine (fixed-dose rate infusion) compared with gemcitabine (30-minute infusion) in patients with pancreatic carcinoma E6201: a trial of the Eastern Cooperative Oncology Group. Journal of clinical oncology : official journal of the American Society of Clinical Oncology 2009;27(23):3778-85.

48. Ueno H, Ioka T, Ikeda M, et al. Randomized phase III study of gemcitabine plus S-1, S-1 alone, or gemcitabine alone in patients with locally advanced and metastatic pancreatic cancer in Japan and Taiwan: GEST study. Journal of clinical oncology : official journal of the American Society of Clinical Oncology 2013;31(13):1640-8.

49. Von Hoff DD, Ervin T, Arena FP, et al. Increased Survival in Pancreatic Cancer with nab-Paclitaxel plus Gemcitabine. New England Journal of Medicine 2013;369(18):1691-703.

50. Mann JF, Gerstein HC, Pogue J, et al. Renal insufficiency as a predictor of cardiovascular outcomes and the impact of ramipril: the HOPE randomized trial. Annals of internal medicine 2001;134(8):629-36.

51. Solomon SD, Rice MM, K AJ, et al. Renal function and effectiveness of angiotensin-converting enzyme inhibitor therapy in patients with chronic stable coronary disease in the Prevention of Events with ACE inhibition (PEACE) trial. Circulation 2006;114(1):26-31.

52. Marre M, Lievre M, Chatellier G, et al. Effects of low dose ramipril on cardiovascular and renal outcomes in patients with type 2 diabetes and raised excretion of urinary albumin: randomised, double blind, placebo controlled trial (the DIABHYCAR study). BMJ (Clinical research ed) 2004;328(7438):495.

53. EUROPA Investigators. Efficacy of perindopril in reduction of cardiovascular events among patients with stable coronary artery disease: randomised, double-blind, placebo-controlled, multicentre trial (the EUROPA study). . Lancet (British edition) 2003;362(9386):755-57.

54. MacMahon S, Sharpe N, Gamble G, et al. Randomized, placebo-controlled trial of the angiotensin-converting enzyme inhibitor, ramipril, in patients with coronary or other occlusive arterial disease. Journal of the American College of Cardiology 2000;36(2):438-43.

55. PROGRESS Collaborative Group. Randomised trial of a perindopril-based blood-pressure-lowering regimen among 6,105 individuals with previous stroke or transient ischaemic attack. Lancet 2001;358(9287):1033-41.

56. Asselbergs FW, Diercks GF, Hillege HL, et al. Effects of fosinopril and pravastatin on cardiovascular events in subjects with microalbuminuria. Circulation 2004;110(18):2809-16.

57. Patel A, MacMahon S, Chalmers J, et al. Effects of a fixed combination of perindopril and indapamide on macrovascular and microvascular outcomes in patients with type 2 diabetes mellitus (the ADVANCE trial): a randomised controlled trial. Lancet 2007;370(9590):829-40.

58. Pitt B, Byington RP, Furberg CD, et al. Effect of amlodipine on the progression of atherosclerosis and the occurrence of clinical events. PREVENT Investigators. Circulation 2000;102(13):1503-10.

59. Estacio RO, Jeffers BW, Hiatt WR, et al. The effect of nisoldipine as compared with enalapril on cardiovascular outcomes in patients with non-insulin-dependent diabetes and hypertension. The New England journal of medicine 1998;338(10):645-52.

60. Hansson L, Zanchetti A, Carruthers SG, et al. Effects of intensive blood-pressure lowering and low-dose aspirin in patients with hypertension: principal results of the Hypertension Optimal Treatment (HOT) randomised trial. The Lancet 1998;351(9118):1755-62.

61. Group UKPDS. Tight Blood Pressure Control and Risk of Macrovascular and Microvascular Complications in Type 2 Diabetes: UKPDS 38. BMJ: British Medical Journal 1998;317(7160):703-13.

62. The Allhat Officers Coordinators for the Allhat Collaborative Research Group. Major Outcomes in High-Risk Hypertensive Patients Randomized to Angiotensin-Converting Enzyme Inhibitor or Calcium Channel Blocker vs DiureticThe Antihypertensive and Lipid-Lowering Treatment to Prevent Heart Attack Trial (ALLHAT). JAMA 2002;288(23):2981-97.

63. Wing LM, Reid CM, Ryan P, et al. A comparison of outcomes with angiotensin-converting--enzyme inhibitors and diuretics for hypertension in the elderly. The New England journal of medicine 2003;348(7):583-92.

64. Hansson L, Lindholm LH, Niskanen L, et al. Effect of angiotensin-converting-enzyme inhibition compared with conventional therapy on cardiovascular morbidity and mortality in hypertension: the Captopril Prevention Project (CAPPP) randomised trial. Lancet 1999;353(9153):611-16.

65. Hansson L, Lindholm LH, Ekbom T, et al. Randomised trial of old and new antihypertensive drugs in elderly patients: cardiovascular mortality and morbidity the Swedish Trial in Old Patients with Hypertension-2 study. Lancet 1999;354(9192):1751-56.

66. Group UKPDS. Efficacy of Atenolol and Captopril in Reducing Risk of Macrovascular and Microvascular Complications in Type 2 Diabetes: UKPDS 39. BMJ: British Medical Journal 1998;317(7160):713-20.

67. Zanchetti A, Bond MG, Hennig M, et al. Calcium Antagonist Lacidipine Slows Down Progression of Asymptomatic Carotid Atherosclerosis. Circulation 2002;106(19):2422-27.

68. National Intervention Cooperative Study in Elderly Hypertensives Group. Randomized Double-Blind Comparison of a Calcium Antagonist and a Diuretic in Elderly Hypertensives. Hypertension 1999;34(5):1129-33.

69. Hansson L, Hedner T, Lund-Johansen P, et al. Randomised trial of effects of calcium antagonists compared with diuretics and beta-blockers on cardiovascular morbidity and mortality in hypertension: the Nordic Diltiazem (NORDIL) study. Lancet 2000;356(9227):359.

70. Yui Y, Sumiyoshi T, Kodama K, et al. Comparison of Nifedipine Retard with Angiotensin Converting Enzyme Inhibitors in Japanese Hypertensive Patients with Coronary Artery Disease: The Japan Multicenter Investigation for Cardiovascular Diseases-B (JMIC-B) Randomized Trial. Hypertension Research 2004;27(3):181-91.

71. Ogawa H, Nakayama M, Morimoto T, et al. Low-dose aspirin for primary prevention of atherosclerotic events in patients with type 2 diabetes: a randomized controlled trial. Jama 2008;300(18):2134-41.

72. Juul-Moller S, Edvardsson N, Sorensen S, et al. Double-blind trial of aspirin in primary prevention of myocardial infarction in patients with stable chronic angina pectoris. The Lancet 1992;340(8833):1421-25.

73. Belch J, MacCuish A, Campbell I, et al. The prevention of progression of arterial disease and diabetes (POPADAD) trial: factorial randomised placebo controlled trial of aspirin and antioxidants in patients with diabetes and asymptomatic peripheral arterial disease. BMJ (Clinical research ed) 2008;337:a1840.

74. Fowkes FG, Price JF, Stewart MC, et al. Aspirin for prevention of cardiovascular events in a general population screened for a low ankle brachial index: a randomized controlled trial. Jama 2010;303(9):841-8.

75. Schaake-Koning C, van den Bogaert W, Dalesio O, et al. Effects of concomitant cisplatin and radiotherapy on inoperable non-small-cell lung cancer. The New England journal of medicine 1992;326(8):524-30.

76. Blanke C, Ansari R, Mantravadi R, et al. Phase III trial of thoracic irradiation with or without cisplatin for locally advanced unresectable non-small-cell lung cancer: a Hoosier Oncology Group protocol. Journal of clinical oncology : official journal of the American Society of Clinical Oncology 1995;13(6):1425-9.

77. Trovò MG, Zanelli GD, Minatel E, et al. Radiotherapy versus radiotherapy enhanced by cisplatin in stage III non-small cell lung cancer. International journal of radiation oncology, biology, physics 1992;24(3):573-4.

78. Jeremic B, Shibamoto Y, Acimovic L, et al. Randomized trial of hyperfractionated radiation therapy with or without concurrent chemotherapy for stage III non-small-cell lung cancer. Journal of clinical oncology : official journal of the American Society of Clinical Oncology 1995;13(2):452-8.

79. Ball D, Bishop J, Smith J, et al. A randomised phase III study of accelerated or standard fraction radiotherapy with or without concurrent carboplatin in inoperable non-small cell lung cancer: final report of an Australian multi-centre trial. Radiotherapy and oncology : journal of the European Society for Therapeutic Radiology and Oncology 1999;52(2):129-36.

80. Jeremic B, Shibamoto Y, Acimovic L, et al. Hyperfractionated radiation therapy with or without concurrent low-dose daily carboplatin/etoposide for stage III non-small-cell lung cancer: a randomized study. Journal of clinical oncology : official journal of the American Society of Clinical Oncology 1996;14(4):1065-70.

81. Clamon G, Herndon J, Cooper R, et al. Radiosensitization with carboplatin for patients with unresectable stage III non-small-cell lung cancer: a phase III trial of the Cancer and Leukemia Group B and the Eastern Cooperative Oncology Group. Journal of clinical oncology : official journal of the American Society of Clinical Oncology 1999;17(1):4-11.

82. Groen HJ, van der Leest AH, Fokkema E, et al. Continuously infused carboplatin used as radiosensitizer in locally unresectable non-small-cell lung cancer: a multicenter phase III study. Annals of oncology : official journal of the European Society for Medical Oncology 2004;15(3):427-32.

83. King SB, Kosinski AS, Guyton RA, et al. Eight-year mortality in the Emory Angioplasty versus Surgery Trial (EAST). Journal of the American College of Cardiology 2000;35(5):1116-21.

84. Rodriguez AE, Baldi J, Fernández Pereira C, et al. Five-year follow-up of the Argentine randomized trial of coronary angioplasty with stenting versus coronary bypass surgery in patients with multiple vessel disease (ERACI II). Journal of the American College of Cardiology 2005;46(4):582-8.

85. Kaehler J, Koester R, Billmann W, et al. 13-year follow-up of the German angioplasty bypass surgery investigation. European Heart Journal 2005;26(20):2148-53.

86. Hueb W, Lopes NH, Gersh BJ, et al. Five-year follow-up of the Medicine, Angioplasty, or Surgery Study (MASS II): a randomized controlled clinical trial of 3 therapeutic strategies for multivessel coronary artery disease. Circulation 2007;115(9):1082-9.

87. Henderson RA, Pocock SJ, Sharp SJ, et al. Long-term results of RITA-1 trial: clinical and cost comparisons of coronary angioplasty and coronary-artery bypass grafting. Randomised Intervention Treatment of Angina. Lancet 1998;352(9138):1419-25.

88. Booth J, Clayton T, Pepper J, et al. Randomized, controlled trial of coronary artery bypass surgery versus percutaneous coronary intervention in patients with multivessel coronary artery disease: six-year follow-up from the Stent or Surgery Trial (SoS). Circulation 2008;118(4):381-8.

89. Seyfarth M, Sibbing D, Bauer I, et al. A randomized clinical trial to evaluate the safety and efficacy of a percutaneous left ventricular assist device versus intra-aortic balloon pumping for treatment of cardiogenic shock caused by myocardial infarction. Journal of the American College of Cardiology 2008;52(19):1584-8.

90. Thiele H, Sick P, Boudriot E, et al. Randomized comparison of intra-aortic balloon support with a percutaneous left ventricular assist device in patients with revascularized acute myocardial infarction complicated by cardiogenic shock. Eur Heart J 2005;26(13):1276-83.

91. Thiele H, Zeymer U, Neumann FJ, et al. Intraaortic balloon support for myocardial infarction with cardiogenic shock. The New England journal of medicine 2012;367(14):1287-96.

92. Groeneveld GJ, Veldink JH, van der Tweel I, et al. A randomized sequential trial of creatine in amyotrophic lateral sclerosis. Annals of neurology 2003;53(4):437-45.

93. Rosenfeld J, King RM, Jackson CE, et al. Creatine monohydrate in ALS: effects on strength, fatigue, respiratory status and ALSFRS. Amyotroph Lateral Scler 2008;9(5):266-72.

94. Bill PA, Vigonius U, Pohlmann H, et al. A double-blind controlled clinical trial of oxcarbazepine versus phenytoin in adults with previously untreated epilepsy. Epilepsy Res 1997;27(3):195-204.

95. Guerreiro MM, Vigonius U, Pohlmann H, et al. A double-blind controlled clinical trial of oxcarbazepine versus phenytoin in children and adolescents with epilepsy. Epilepsy Res 1997;27(3):205-13.

96. Leaf A, Albert CM, Josephson M, et al. Prevention of fatal arrhythmias in high-risk subjects by fish oil n-3 fatty acid intake. Circulation 2005;112(18):2762-8.

97. Raitt MH, Connor WE, Morris C, et al. Fish oil supplementation and risk of ventricular tachycardia and ventricular fibrillation in patients with implantable defibrillators: a randomized controlled trial. Jama 2005;293(23):2884-91.

98. Brouwer IA, Zock PL, Camm AJ, et al. Effect of fish oil on ventricular tachyarrhythmia and death in patients with implantable cardioverter defibrillators: the Study on Omega-3 Fatty Acids and Ventricular Arrhythmia (SOFA) randomized trial. Jama 2006;295(22):2613-9.

99. de Silva M, MacArdle B, McGowan M, et al. Randomised comparative monotherapy trial of phenobarbitone, phenytoin, carbamazepine, or sodium valproate for newly diagnosed childhood epilepsy. The Lancet 1996;347(9003):709-13.

100. Heller AJ, Chesterman P, Elwes RD, et al. Phenobarbitone, phenytoin, carbamazepine, or sodium valproate for newly diagnosed adult epilepsy: a randomised comparative monotherapy trial. Journal of neurology, neurosurgery, and psychiatry 1995;58(1):44-50.

101. Bonner JA, McGinnis WL, Stella PJ, et al. The possible advantage of hyperfractionated thoracic radiotherapy in the treatment of locally advanced nonsmall cell lung carcinoma. Cancer 1998;82(6):1037-48.

102. BARI Investigators. The final 10-year follow-up results from the BARI randomized trial. Journal of the American College of Cardiology 2007;49(15):1600-06.

103. Merville P, Bergé F, Deminière C, et al. Lower incidence of chronic allograft nephropathy at 1 year post-transplantation in patients treated with mycophenolate mofetil. American journal of transplantation : official journal of the American Society of Transplantation and the American Society of Transplant Surgeons 2004;4(11):1769-75.

104. Miladipour AH, Ghods AJ, Nejadgashti H. Effect of mycophenolate mofetil on the prevention of acute renal allograft rejection. Transplantation proceedings 2002;34(6):2089-90.

105. Weimer R, Süsal C, Yildiz S, et al. Post-transplant sCD30 and neopterin as predictors of chronic allograft nephropathy: impact of different immunosuppressive regimens. American journal of transplantation : official journal of the American Society of Transplantation and the American Society of Transplant Surgeons 2006;6(8):1865-74.

106. Wlodarczyk Z, Walaszewski J, Perner F, et al. Freedom from rejection and stable kidney function are excellent criteria for steroid withdrawal in tacrolimus-treated kidney transplant recipients. Annals of transplantation 2002;7(3):28-31.

Appendix 4: Search strategies used in Chapter 4

Database: CDSR// DARE/Method studies

#1 MeSH descriptor: [Proportional Hazards Models] explode all trees

#2 (proportion* near/1 hazard*)

#3 (cox near/2 proportion* near/2 hazard*)

#4 (Cox near/2 regress* near/2 model)

#5 #1 or #2 or #3 or #4

#6 (Hazard* near/2 ratio*)

#7 (surviv* near/2 model*)

#8 MeSH descriptor: [Survival Analysis] explode all trees

#9 MeSH descriptor: [Disease-Free Survival] explode all trees

#10 ((assess* or analys*) near/4 surviv*)

#11 MeSH descriptor: [Kaplan-Meier Estimate] explode all trees

#12 kaplan-meier

#13 "time to event"

#14 "time to mortalit*"

#15 "time to progress* diseas*"

#16 parmar*

#17 #6 or #7 or #8 or #9 or #10 or #11 or #12 or #13 or #14 or #15 or #16

#18 #5 and #17 Publication Year from 2005 to 2015

Database: Medline

Strategy used:

Searches Results Search Type

1 proportional hazards models/
2 (proportion* adj1 hazard*).tw.
3 (cox adj2 proportion* adj2 hazard*).tw.
4 (Cox adj2 regress* adj2 model).tw.
5 or/1-4
6 (Hazard* adj2 ratio*).tw.
7 (surviv* adj2 model*).tw.
8 survival analysis/ or disease-free survival/
9 ((assess* or analys*) adj4 surviv*).tw.
10 kaplan-meier estimate/
11 kaplan-meier.tw.
12 "time to event".tw.
13 "time to mortalit*".tw.
14 "time to progress* diseas*".tw.
15 parmar*.tw.
16 or/6-15
17 5 and 16
18 limit 17 to yr="2005 -Current"
19 ((systematic* adj3 (review* or overview*)) or (methodologic* adj3 (review* or overview*))).ti,ab.
20 ((quantitative adj3 (review* or overview* or synthes*)) or (research adj3 (integrati* or overview*))).ti,ab.
21 ((integrative adj3 (review* or overview*)) or (collaborative adj3 (review* or overview*)) or (pool* adj3 analy*)).ti,ab.
22 (data synthes* or data extraction* or data abstraction*).ti,ab.
23 (handsearch* or hand search*).ti,ab.
24 (medline or cochrane or pubmed or medlars or embase or cinahl).ti,ab,hw.
25 (cochrane or (health adj2 technology assessment) or evidence report).jw.
26 systematic review.tw.
27 or/19-26
28 18 and 27

Database: Embase

Strategy used:

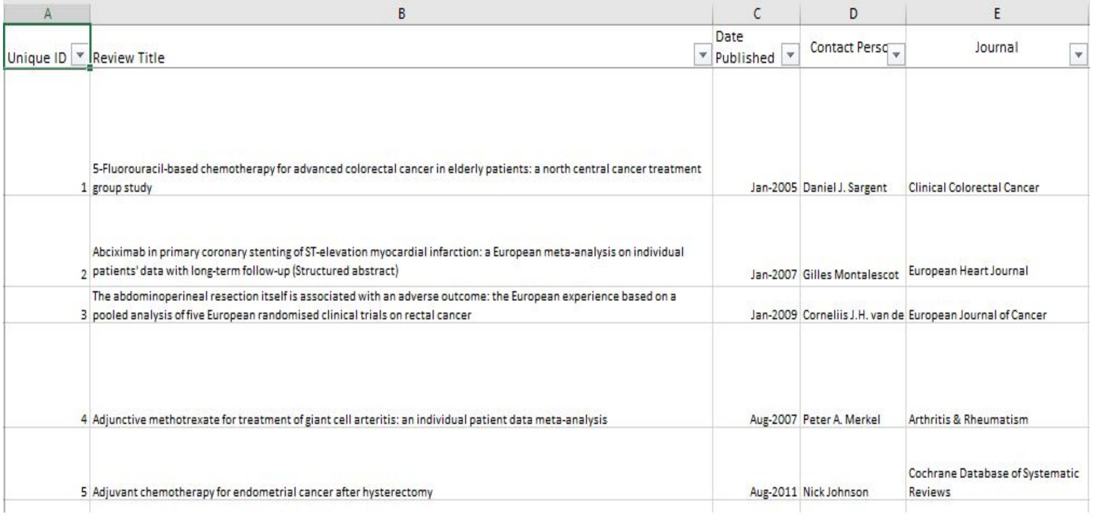
# ▲	Searches
1	proportional hazards models/
2	(proportion* adj1 hazard*).tw.
3	(cox adj2 proportion* adj2 hazard*).tw.
4	(Cox adj2 regress* adj2 model).tw.
5	or/1-4
6	(Hazard* adj2 ratio*).tw.
7	(surviv* adj2 model*).tw.
8	survival analysis/ or disease-free survival/
9	((assess* or analys*) adj4 surviv*).tw.
10	kaplan-meier estimate/
11	kaplan-meier.tw.
12	"time to event".tw.
13	"time to mortalit*".tw.
14	"time to progress* diseas*".tw.
15	parmar*.tw.
16	or/6-15
17	5 and 16
18	limit 17 to yr="2005 -Current"
19	((systematic* adj3 (review* or overview*)) or (methodologic* adj3 (review* or overview*))).ti,ab.
20	((quantitative adj3 (review* or overview* or synthes*)) or (research adj3 (integrati* or overview*))).ti,ab.
21	((integrative adj3 (review* or overview*)) or (collaborative adj3 (review* or overview*)) or (pool* adj3 analy*)).ti,ab.
22	(data synthes* or data extraction* or data abstraction*).ti,ab.
23	(handsearch* or hand search*).ti,ab.
24	(medline or cochrane or pubmed or medlars or embase or cinahl).ti,ab,hw.
25	(cochrane or (health adj2 technology assessment) or evidence report).jw.
26	systematic review.tw.
27	or/19-26
28	**18 and 27**

Database: PubMed

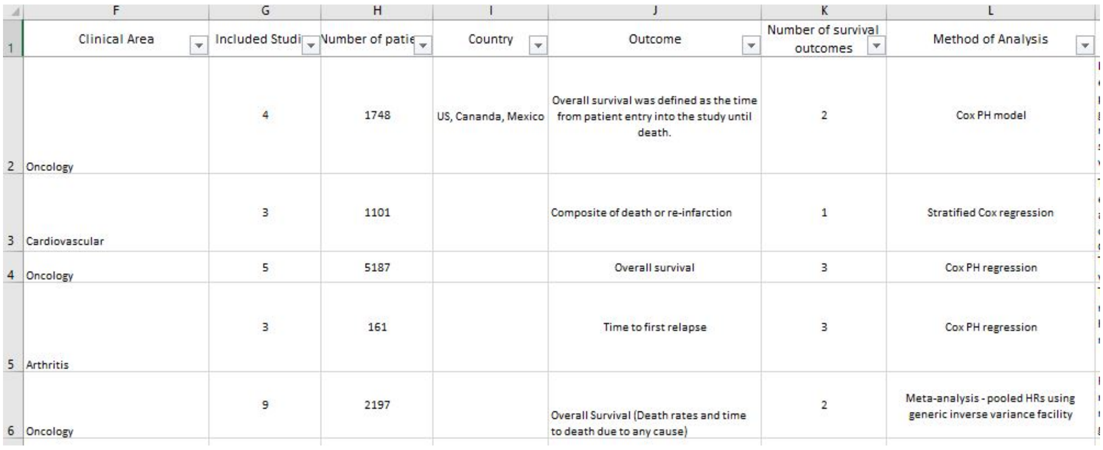
Strategy used:

#1	Search (((proportion* hazard*) OR cox proportion* hazard*) OR Cox regress* model)
#2	Search (((((((((Hazard* ratio*) OR surviv* model*) OR (assess* or analys*)) AND surviv*) OR kaplan-meier) AND "time to event") OR "time to mortalit*") OR "time to progress* diseas*") OR parmar*)
#3	Search (#1 and #2)
#4	Search ("2014/01/01"[Date - Entrez] : "3000"[Date - Entrez])
#5	Search (#3 and #4)
#6	Search (#3 and #4) Filters: Systematic Reviews
#7	Search (#3 and #4) Filters: Systematic Reviews; Review
#8	Search (#3 and #4) Filters: Systematic Reviews; Review; Meta-Analysis

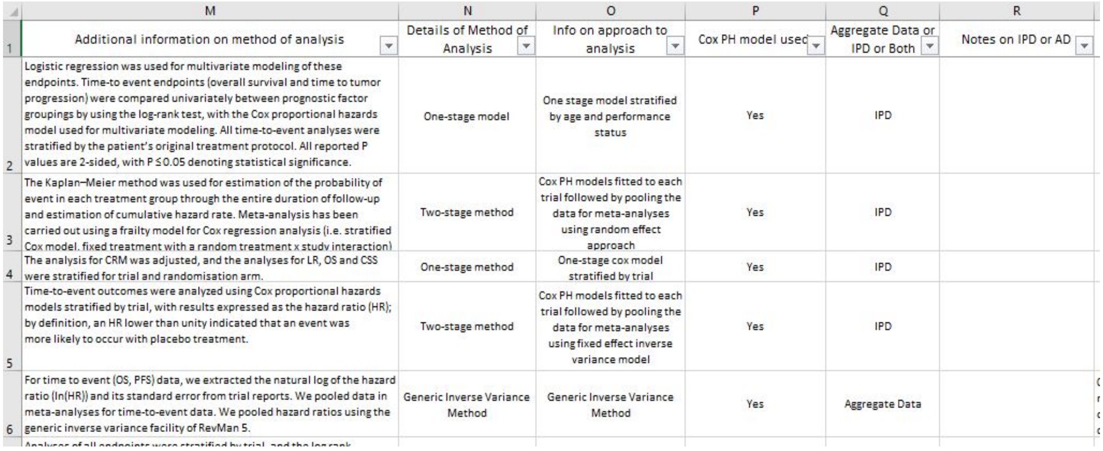
Appendix 5: Screenshot of Microsoft Excel Database used in Chapter 4



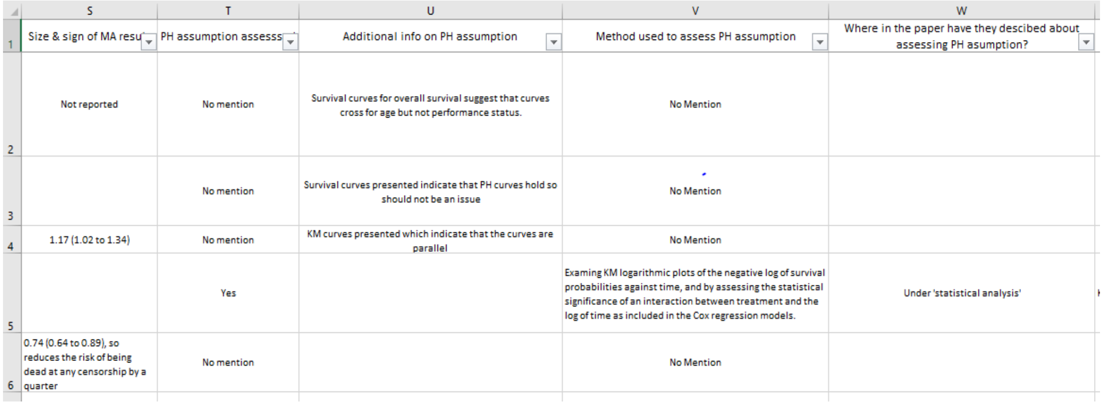
A	B	C	D	E
Unique ID	Review Title	Date Published	Contact Perso	Journal
1	5-Fluorouracil-based chemotherapy for advanced colorectal cancer in elderly patients: a north central cancer treatment group study	Jan-2005	Daniel J. Sargent	Clinical Colorectal Cancer
2	Abciximab in primary coronary stenting of ST-elevation myocardial infarction: a European meta-analysis on individual patients' data with long-term follow-up (Structured abstract)	Jan-2007	Gilles Montalescot	European Heart Journal
3	The abdominoperineal resection itself is associated with an adverse outcome: the European experience based on a pooled analysis of five European randomised clinical trials on rectal cancer	Jan-2009	Corneliis J.H. van de	European Journal of Cancer
4	Adjunctive methotrexate for treatment of giant cell arteritis: an individual patient data meta-analysis	Aug-2007	Peter A. Merkel	Arthritis & Rheumatism
5	Adjuvant chemotherapy for endometrial cancer after hysterectomy	Aug-2011	Nick Johnson	Cochrane Database of Systematic Reviews



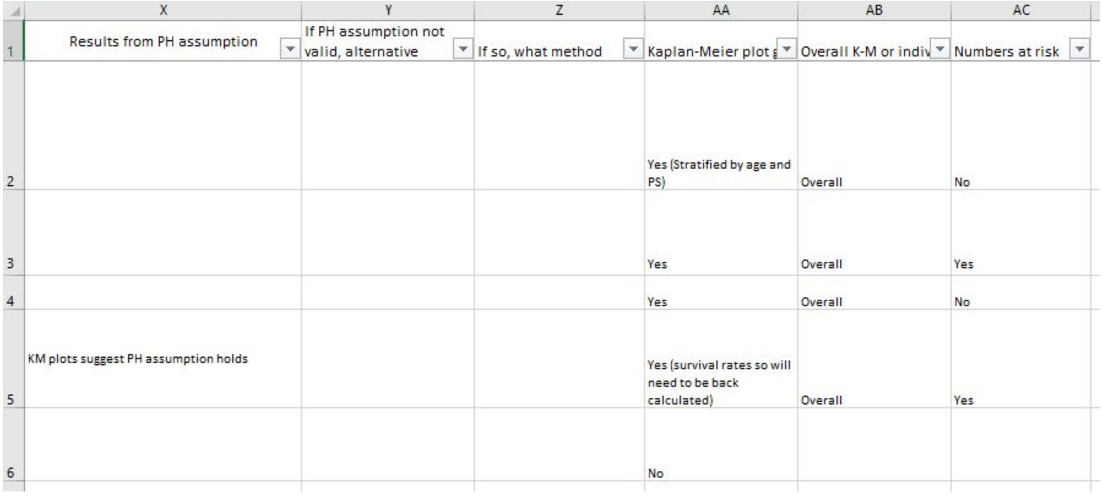
	F	G	H	I	J	K	L
1	Clinical Area	Included Studi	Number of patie	Country	Outcome	Number of survival outcomes	Method of Analysis
2	Oncology	4	1748	US, Cananda, Mexico	Overall survival was defined as the time from patient entry into the study until death.	2	Cox PH model
3	Cardiovascular	3	1101		Composite of death or re-infarction	1	Stratified Cox regression
4	Oncology	5	5187		Overall survival	3	Cox PH regression
5	Arthritis	3	161		Time to first relapse	3	Cox PH regression
6	Oncology	9	2197		Overall Survival (Death rates and time to death due to any cause)	2	Meta-analysis - pooled HRs using generic inverse variance facility



	M	N	O	P	Q	R
1	Additional information on method of analysis	Details of Method of Analysis	Info on approach to analysis	Cox PH model used	Aggregate Data or IPD or Both	Notes on IPD or AD
2	Logistic regression was used for multivariate modeling of these endpoints. Time-to event endpoints (overall survival and time to tumor progression) were compared univariately between prognostic factor groupings by using the log-rank test, with the Cox proportional hazards model used for multivariate modeling. All time-to-event analyses were stratified by the patient's original treatment protocol. All reported P values are 2-sided, with P ≤0.05 denoting statistical significance.	One-stage model	One stage model stratified by age and performance status	Yes	IPD	
3	The Kaplan–Meier method was used for estimation of the probability of event in each treatment group through the entire duration of follow-up and estimation of cumulative hazard rate. Meta-analysis has been carried out using a frailty model for Cox regression analysis (i.e. stratified Cox model, fixed treatment with a random treatment x study interaction)	Two-stage method	Cox PH models fitted to each trial followed by pooling the data for meta-analyses using random effect approach	Yes	IPD	
4	The analysis for CRM was adjusted, and the analyses for LR, OS and CSS were stratified for trial and randomisation arm.	One-stage method	One-stage cox model stratified by trial	Yes	IPD	
5	Time-to-event outcomes were analyzed using Cox proportional hazards models stratified by trial, with results expressed as the hazard ratio (HR); by definition, an HR lower than unity indicated that an event was more likely to occur with placebo treatment.	Two-stage method	Cox PH models fitted to each trial followed by pooling the data for meta-analyses using fixed effect inverse variance model	Yes	IPD	
6	For time to event (OS, PFS) data, we extracted the natural log of the hazard ratio (ln(HR)) and its standard error from trial reports. We pooled data in meta-analyses for time-to-event data. We pooled hazard ratios using the generic inverse variance facility of RevMan 5.	Generic Inverse Variance Method	Generic Inverse Variance Method	Yes	Aggregate Data	
	Analyses of all endpoints were stratified by trial, and the log rank					



	S	T	U	V	W
1	Size & sign of MA resu	PH assumption assesss	Additional info on PH assumption	Method used to assess PH assumption	Where in the paper have they descibed about assessing PH asumption?
2	Not reported	No mention	Survival curves for overall survival suggest that curves cross for age but not performance status.	No Mention	
3		No mention	Survival curves presented indicate that PH curves hold so should not be an issue	No Mention	
4	1.17 (1.02 to 1.34)	No mention	KM curves presented which indicate that the curves are parallel	No Mention	
5		Yes		Examing KM logarithmic plots of the negative log of survival probabilities against time, and by assessing the statistical significance of an interaction between treatment and the log of time as included in the Cox regression models.	Under 'statistical analysis'
6	0.74 (0.64 to 0.89), so reduces the risk of being dead at any censorship by a quarter	No mention		No Mention	



	X	Y	Z	AA	AB	AC
1	Results from PH assumption	If PH assumption not valid, alternative	If so, what method	Kaplan-Meier plot g	Overall K-M or indiv	Numbers at risk
2				Yes (Stratified by age and PS)	Overall	No
3				Yes	Overall	Yes
4				Yes	Overall	No
5	KM plots suggest PH assumption holds			Yes (survival rates so will need to be back calculated)	Overall	Yes
6				No		

Appendix 6: References of 123 systematic reviews included in Chapter 4

1. D'Andre S, Sargent DJ, Cha SS, et al. 5-Fluorouracil–Based Chemotherapy for Advanced Colorectal Cancer in Elderly Patients: A North Central Cancer Treatment Group Study. Clinical Colorectal Cancer 2005;4(5):325-31.

2. Montalescot G, Antoniucci D, Kastrati A, et al. Abciximab in primary coronary stenting of ST-elevation myocardial infarction: a European meta-analysis on individual patients' data with long-term follow-up. European Heart Journal 2007;28(4):443-49.

3. den Dulk M, Putter H, Collette L, et al. The abdominoperineal resection itself is associated with an adverse outcome: The European experience based on a pooled analysis of five European randomised clinical trials on rectal cancer. European Journal of Cancer 2009;45(7):1175-83.

4. Mahr AD, Jover JA, Spiera RF, et al. Adjunctive methotrexate for treatment of giant cell arteritis: An individual patient data meta-analysis. Arthritis & Rheumatism 2007;56(8):2789-97.

5. Johnson N, Bryant A, Miles T, et al. Adjuvant chemotherapy for endometrial cancer after hysterectomy. Cochrane Database Syst Rev 2011;2011(10):CD003175-CD75.

6. Adjuvant chemotherapy in invasive bladder cancer: a systematic review and meta-analysis of individual patient data Advanced Bladder Cancer (ABC) Meta-analysis Collaboration. European urology 2005;48(2):189-99; discussion 99-201.

7. Fredrickson Fanzca MJ, Danesh-Clough TK, White R. Adjuvant dexamethasone for bupivacaine sciatic and ankle blocks: results from 2 randomized placebo-controlled trials. Regional anesthesia and pain medicine 2013;38(4):300-7.

8. Shepperd S, Doll H, Angus RM, et al. Admission avoidance hospital at home. Cochrane Database Syst Rev 2008(4):Cd007491.

9. Delaney CP, Wolff BG, Viscusi ER, et al. Alvimopan, for postoperative ileus following bowel resection: a pooled analysis of phase III studies. Annals of surgery 2007;245(3):355-63.

10. Gaitskell K, Martinek I, Bryant A, et al. Angiogenesis inhibitors for the treatment of ovarian cancer. Cochrane Database of Systematic Reviews 2011(9)

11. Iacovelli R, Altavilla A, Procopio G, et al. Are post-docetaxel treatments effective in patients with castration-resistant prostate cancer and performance of 2? A meta-analysis of published trials. Prostate cancer and prostatic diseases 2013;16(4):323-7.

12. Thoonsen H, Richard E, Bentham P, et al. Aspirin in Alzheimer's disease: increased risk of intracerebral hemorrhage: cause for concern? Stroke 2010;41(11):2690-2.

13. Iqbal N, Parker A, Frederich R, et al. Assessment of the cardiovascular safety of saxagliptin in patients with type 2 diabetes mellitus: pooled analysis of 20 clinical trials. Cardiovascular Diabetology 2014;13(1):33.

14. Chan K, Shah K, Lien K, et al. A Bayesian meta-analysis of multiple treatment comparisons of systemic regimens for advanced pancreatic cancer. PloS one 2014;9(10):e108749.

15. Ben-Aharon I, Vidal L, Rizel S, et al. Bisphosphonates in the adjuvant setting of breast cancer therapy--effect on survival: a systematic review and meta-analysis. PloS one 2013;8(8):e70044.

16. Blood pressure lowering and major cardiovascular events in people with and without chronic kidney disease: meta-analysis of randomised controlled trials. BMJ : British Medical Journal 2013;347:f5680.

17. Sin DD, Tashkin D, Zhang X, et al. Budesonide and the risk of pneumonia: a meta-analysis of individual patient data. Lancet 2009;374(9691):712-19.

18. Bristow SM, Bolland MJ, MacLennan GS, et al. Calcium supplements and cancer risk: a meta-analysis of randomised controlled trials. The British journal of nutrition 2013;110(8):1384-93.

19. Nevitt SJ, Marson AG, Tudur Smith C. Carbamazepine versus phenobarbitone monotherapy for epilepsy: an individual participant data review. Cochrane Database of Systematic Reviews 2018(10)

20. White WB, Pratley R, Fleck P, et al. Cardiovascular safety of the dipetidyl peptidase-4 inhibitor alogliptin in type 2 diabetes mellitus. Diabetes, obesity & metabolism 2013;15(7):668-73.

21. Baujat B, Audry H, Bourhis J, et al. Chemotherapy as an adjunct to radiotherapy in locally advanced nasopharyngeal carcinoma. Cochrane Database Syst Rev 2006(4):Cd004329.

22. Kelly K, Chansky K, Mack PC, et al. Chemotherapy outcomes by histologic subtypes of non-small-cell lung cancer: analysis of the southwest oncology group database for antimicrotubule-platinum therapy. Clinical lung cancer 2013;14(6):627-35.

23. Di Maio M, Gridelli C, Gallo C, et al. Chemotherapy-induced neutropenia and treatment efficacy in advanced non-small-cell lung cancer: a pooled analysis of three randomised trials. The Lancet Oncology 2005;6(9):669-77.

24. Ardizzoni A, Boni L, Tiseo M, et al. Cisplatin- versus carboplatin-based chemotherapy in first-line treatment of advanced non-small-cell lung cancer: an individual patient data meta-analysis. J Natl Cancer Inst 2007;99(11):847-57.

25. Caixeta A, Lansky AJ, Serruys PW, et al. Clinical follow-up 3 years after everolimus- and paclitaxel-eluting stents: a pooled analysis from the SPIRIT II (A Clinical Evaluation of the XIENCE V Everolimus Eluting Coronary Stent System in the Treatment of Patients With De Novo Native Coronary Artery Lesions) and SPIRIT III (A Clinical Evaluation of the Investigational Device XIENCE V Everolimus Eluting Coronary Stent System [EECSS] in the Treatment of Subjects With De Novo Native Coronary Artery Lesions) randomized trials. JACC Cardiovascular interventions 2010;3(12):1220-8.

26. Halabi S, Vogelzang NJ, Ou SS, et al. Clinical outcomes by age in men with hormone refractory prostate cancer: a pooled analysis of 8 Cancer and Leukemia Group B (CALGB) studies. The Journal of urology 2006;176(1):81-6.

27. Abbate A, Kontos MC, Abouzaki NA, et al. Comparative safety of interleukin-1 blockade with anakinra in patients with ST-segment elevation acute myocardial infarction (from the VCU-ART and VCU-ART2 pilot studies). The American journal of cardiology 2015;115(3):288-92.

28. Papakostas GI, Montgomery SA, Thase ME, et al. Comparing the rapidity of response during treatment of major depressive disorder with bupropion and the SSRIs: a pooled survival analysis of 7 double-blind, randomized clinical trials. The Journal of clinical psychiatry 2007;68(12):1907-12.

29. Douillard JY, Laporte S, Fossella F, et al. Comparison of docetaxel- and vinca alkaloid-based chemotherapy in the first-line treatment of advanced non-small cell lung cancer: a

meta-analysis of seven randomized clinical trials. Journal of thoracic oncology : official publication of the International Association for the Study of Lung Cancer 2007;2(10):939-46.

30. O'Meara S, Cullum N, Nelson EA, et al. Compression for venous leg ulcers. Cochrane Database Syst Rev 2012;11(11):Cd000265.

31. Aupérin A, Le Péchoux C, Pignon JP, et al. Concomitant radio-chemotherapy based on platin compounds in patients with locally advanced non-small cell lung cancer (NSCLC): a meta-analysis of individual data from 1764 patients. Annals of oncology : official journal of the European Society for Medical Oncology 2006;17(3):473-83.

32. Hlatky MA, Boothroyd DB, Bravata DM, et al. Coronary artery bypass surgery compared with percutaneous coronary interventions for multivessel disease: a collaborative analysis of individual patient data from ten randomised trials. Lancet 2009;373(9670):1190-7.

33. Pastula DM, Moore DH, Bedlack RS. Creatine for amyotrophic lateral sclerosis/motor neuron disease. Cochrane Database Syst Rev 2012;12:Cd005225.

34. Roll S, Müller-Nordhorn J, Keil T, et al. Dacron vs. PTFE as bypass materials in peripheral vascular surgery: systematic review and meta-analysis. BMC Surg 2008;8:22-22.

35. Halkes PH, Gray LJ, Bath PM, et al. Dipyridamole plus aspirin versus aspirin alone in secondary prevention after TIA or stroke: a meta-analysis by risk. Journal of neurology, neurosurgery, and psychiatry 2008;79(11):1218-23.

36. Sinnaeve PR, Simes J, Yusuf S, et al. Direct thrombin inhibitors in acute coronary syndromes: effect in patients undergoing early percutaneous coronary intervention. European Heart Journal 2005;26(22):2396-403.

37. Klotz L, Miller K, Crawford ED, et al. Disease Control Outcomes from Analysis of Pooled Individual Patient Data from Five Comparative Randomised Clinical Trials of Degarelix Versus Luteinising Hormone-releasing Hormone Agonists. European urology 2014;66(6):1101-08.

38. Van Gelder IC, Wyse DG, Chandler ML, et al. Does intensity of rate-control influence outcome in atrial fibrillation? An analysis of pooled data from the RACE and AFFIRM studies. EP Europace 2006;8(11):935-42.

39. De Luca G, Dirksen MT, Spaulding C, et al. Drug-eluting vs bare-metal stents in primary angioplasty: a pooled patient-level meta-analysis of randomized trials. Archives of internal medicine 2012;172(8):611-21.

40. Li X, Xu SN, Qin DB, et al. Effect of adding gemtuzumab ozogamicin to induction chemotherapy for newly diagnosed acute myeloid leukemia: a meta-analysis of prospective randomized phase III trials. Annals of oncology : official journal of the European Society for Medical Oncology 2014;25(2):455-61.

41. Su Y, Yang W-B, Li S, et al. Effect of Angiogenesis Inhibitor Bevacizumab on Survival in Patients with Cancer: A Meta-Analysis of the Published Literature. PloS one 2012;7(4):e35629.

42. Troughton RW, Frampton CM, Brunner-La Rocca H-P, et al. Effect of B-type natriuretic peptide-guided treatment of chronic heart failure on total mortality and hospitalization: an individual patient meta-analysis. European Heart Journal 2014;35(23):1559-67.

43. Bolland MJ, Avenell A, Baron JA, et al. Effect of calcium supplements on risk of myocardial infarction and cardiovascular events: meta-analysis. BMJ (Clinical research ed) 2010;341:c3691.

44. Rothwell PM, Fowkes FG, Belch JF, et al. Effect of daily aspirin on long-term risk of death due to cancer: analysis of individual patient data from randomised trials. Lancet 2011;377(9759):31-41.

45. Pearse RM, Belsey JD, Cole JN, et al. Effect of dopexamine infusion on mortality following major surgery: individual patient data meta-regression analysis of published clinical trials. Critical care medicine 2008;36(4):1323-9.

46. Brouwer IA, Raitt MH, Dullemeijer C, et al. Effect of fish oil on ventricular tachyarrhythmia in three studies in patients with implantable cardioverter defibrillators. European Heart Journal 2009;30(7):820-26.

47. Ladoire S, Dalban C, Roché H, et al. Effect of obesity on disease-free and overall survival in node-positive breast cancer patients in a large French population: a pooled analysis of two randomised trials. European journal of cancer (Oxford, England : 1990) 2014;50(3):506-16.

48. Jonat W, Gnant M, Boccardo F, et al. Effectiveness of switching from adjuvant tamoxifen to anastrozole in postmenopausal women with hormone-sensitive early-stage breast cancer: a meta-analysis. The Lancet Oncology 2006;7(12):991-6.

49. Perez MV, Wang PJ, Larson JC, et al. Effects of postmenopausal hormone therapy on incident atrial fibrillation: the Women's Health Initiative randomized controlled trials. Circulation Arrhythmia and electrophysiology 2012;5(6):1108-16.

50. Bohlius J, Schmidlin K, Brillant C, et al. Erythropoietin or Darbepoetin for patients with cancer-meta-analysis based on individual patient data. Cochrane Database Syst Rev 2009;2009(3):Cd007303.

51. Andre F, Broglio K, Pusztai L, et al. Estrogen receptor expression and docetaxel efficacy in patients with metastatic breast cancer: a pooled analysis of four randomized trials. The oncologist 2010;15(5):476-83.

52. Andre F, Broglio K, Roche H, et al. Estrogen receptor expression and efficacy of docetaxel-containing adjuvant chemotherapy in patients with node-positive breast cancer: results from a pooled analysis. Journal of clinical oncology : official journal of the American Society of Clinical Oncology 2008;26(16):2636-43.

53. Bongartz T, Warren FC, Mines D, et al. Etanercept therapy in rheumatoid arthritis and the risk of malignancies: a systematic review and individual patient data meta-analysis of randomised controlled trials. Annals of the rheumatic diseases 2009;68(7):1177-83.

54. Jiang Y, Yin W, Zhou L, et al. First efficacy results of capecitabine with anthracycline- and taxane-based adjuvant therapy in high-risk early breast cancer: a meta-analysis. PloS one 2012;7(3):e32474-e74.

55. Stein EA, Corsini A, Gimpelewicz CR, et al. Fluvastatin treatment is not associated with an increased incidence of cancer. International Journal of Clinical Practice 2006;60(9):1028-34.

56. O'Meara S, Tierney J, Cullum N, et al. Four layer bandage compared with short stretch bandage for venous leg ulcers: systematic review and meta-analysis of randomised controlled trials with data from individual patients. BMJ (Clinical research ed) 2009;338:b1344-b44.

57. Takagi H, Goto SN, Matsui M, et al. A further meta-analysis of population-based screening for abdominal aortic aneurysm. Journal of vascular surgery 2010;52(4):1103-8.

58. Santangeli P, Pelargonio G, Dello Russo A, et al. Gender differences in clinical outcome and primary prevention defibrillator benefit in patients with severe left ventricular dysfunction: a systematic review and meta-analysis. Heart rhythm 2010;7(7):876-82.

59. Kokka F, Bryant A, Brockbank E, et al. Hysterectomy with radiotherapy or chemotherapy or both for women with locally advanced cervical cancer. Cochrane Database Syst Rev 2015(4):Cd010260.

60. Cranney A, Wells GA, Yetisir E, et al. Ibandronate for the prevention of nonvertebral fractures: a pooled analysis of individual patient data. Osteoporosis international : a journal established as result of cooperation between the European Foundation for Osteoporosis and the National Osteoporosis Foundation of the USA 2009;20(2):291-7.

61. Halpin DM, Peterson S, Larsson TP, et al. Identifying COPD patients at increased risk of mortality: predictive value of clinical study baseline data. Respiratory medicine 2008;102(11):1615-24.

62. Shao N, Wang Y, Jiang WY, et al. Immunotherapy and endothelin receptor antagonists for treatment of castration-resistant prostate cancer. International journal of cancer 2013;133(7):1743-50.

63. Haller DG, O'Connell MJ, Cartwright TH, et al. Impact of age and medical comorbidity on adjuvant treatment outcomes for stage III colon cancer: a pooled analysis of individual patient data from four randomized, controlled trials. Annals of oncology : official journal of the European Society for Medical Oncology 2015;26(4):715-24.

64. Rao SV, O'Grady K, Pieper KS, et al. Impact of bleeding severity on clinical outcomes among patients with acute coronary syndromes. The American journal of cardiology 2005;96(9):1200-6.

65. van der Hage JA, Mieog JS, van de Velde CJ, et al. Impact of established prognostic factors and molecular subtype in very young breast cancer patients: pooled analysis of four EORTC randomized controlled trials. Breast cancer research : BCR 2011;13(3):R68.

66. Claessen BE, Smits PC, Kereiakes DJ, et al. Impact of lesion length and vessel size on clinical outcomes after percutaneous coronary intervention with everolimus- versus paclitaxel-eluting stents pooled analysis from the SPIRIT (Clinical Evaluation of the XIENCE V Everolimus Eluting Coronary Stent System) and COMPARE (Second-generation everolimus-eluting and paclitaxel-eluting stents in real-life practice) Randomized Trials. JACC Cardiovascular interventions 2011;4(11):1209-15.

67. Ben-Josef E, Moughan J, Ajani JA, et al. Impact of overall treatment time on survival and local control in patients with anal cancer: a pooled data analysis of Radiation Therapy Oncology Group trials 87-04 and 98-11. Journal of clinical oncology : official journal of the American Society of Clinical Oncology 2010;28(34):5061-6.

68. Blanke CD, Bot BM, Thomas DM, et al. Impact of young age on treatment efficacy and safety in advanced colorectal cancer: a pooled analysis of patients from nine first-line phase III chemotherapy trials. Journal of clinical oncology : official journal of the American Society of Clinical Oncology 2011;29(20):2781-6.

69. Stockley RA, Whitehead PJ, Williams MK. Improved outcomes in patients with chronic obstructive pulmonary disease treated with salmeterol compared with placebo/usual therapy: results of a meta-analysis. Respiratory research 2006;7(1):147.

70. Oba K, Kobayashi M, Matsui T, et al. Individual patient based meta-analysis of lentinan for unresectable/recurrent gastric cancer. Anticancer research 2009;29(7):2739-45.

71. Buyse M, Squifflet P, Lange BJ, et al. Individual patient data meta-analysis of randomized trials evaluating IL-2 monotherapy as remission maintenance therapy in acute myeloid leukemia. Blood 2011;117(26):7007-13.

72. Templeton AJ, Ace O, Amir E, et al. Influence of censoring on conclusions of trials for women with metastatic breast cancer. European journal of cancer (Oxford, England : 1990) 2015;51(6):721-4.

73. Saw J, Bhatt DL, Moliterno DJ, et al. The influence of peripheral arterial disease on outcomes: a pooled analysis of mortality in eight large randomized percutaneous coronary intervention trials. Journal of the American College of Cardiology 2006;48(8):1567-72.

74. Unverzagt S, Buerke M, de Waha A, et al. Intra-aortic balloon pump counterpulsation (IABP) for myocardial infarction complicated by cardiogenic shock. Cochrane Database Syst Rev 2015(3):Cd007398.

75. Faron M, Pignon JP, Malka D, et al. Is primary tumour resection associated with survival improvement in patients with colorectal cancer and unresectable synchronous metastases? A pooled analysis of individual data from four randomised trials. European journal of cancer (Oxford, England : 1990) 2015;51(2):166-76.

76. Mauri L, Massaro JM, Jiang S, et al. Long-Term Clinical Outcomes With Zotarolimus-Eluting Versus Bare-Metal Coronary Stents. JACC: Cardiovascular Interventions 2010;3(12):1240-49.

77. Dimopoulos MA, Chen C, Spencer A, et al. Long-term follow-up on overall survival from the MM-009 and MM-010 phase III trials of lenalidomide plus dexamethasone in patients with relapsed or refractory multiple myeloma. Leukemia 2009;23(11):2147-52.

78. Ng SS, Lee JF, Yiu RY, et al. Long-term oncologic outcomes of laparoscopic versus open surgery for rectal cancer: a pooled analysis of 3 randomized controlled trials. Annals of surgery 2014;259(1):139-47.

79. Fox KA, Clayton TC, Damman P, et al. Long-term outcome of a routine versus selective invasive strategy in patients with non-ST-segment elevation acute coronary syndrome a meta-analysis of individual patient data. Journal of the American College of Cardiology 2010;55(22):2435-45.

80. Stone GW, Ellis SG, Colombo A, et al. Long-term safety and efficacy of paclitaxel-eluting stents final 5-year analysis from the TAXUS Clinical Trial Program. JACC Cardiovascular interventions 2011;4(5):530-42.

81. Theophilus M, Platell C, Spilsbury K. Long-term survival following laparoscopic and open colectomy for colon cancer: a meta-analysis of randomized controlled trials. Colorectal disease : the official journal of the Association of Coloproctology of Great Britain and Ireland 2014;16(3):O75-81.

82. Siegfried N, Muller M, Deeks JJ, et al. Male circumcision for prevention of heterosexual acquisition of HIV in men. Cochrane Database Syst Rev 2009(2):Cd003362.

83. Dangas GD, Serruys PW, Kereiakes DJ, et al. Meta-analysis of everolimus-eluting versus paclitaxel-eluting stents in coronary artery disease: final 3-year results of the SPIRIT clinical trials program (Clinical Evaluation of the Xience V Everolimus Eluting Coronary Stent System in the Treatment of Patients With De Novo Native Coronary Artery Lesions). JACC Cardiovascular interventions 2013;6(9):914-22.

84. Tudur Smith C, Marson AG, Chadwick DW, et al. Multiple treatment comparisons in epilepsy monotherapy trials. Trials 2007;8:34-34.

85. Knight SR, Russell NK, Barcena L, et al. Mycophenolate mofetil decreases acute rejection and may improve graft survival in renal transplant recipients when compared with azathioprine: a systematic review. Transplantation 2009;87(6):785-94.

86. Zhou S-w, Huang Y-y, Wei Y, et al. No survival benefit from adding cetuximab or panitumumab to oxaliplatin-based chemotherapy in the first-line treatment of metastatic colorectal cancer in KRAS wild type patients: a meta-analysis. PloS one 2012;7(11):e50925-e25.

87. Spagnolo P, Del Giovane C, Luppi F, et al. Non-steroid agents for idiopathic pulmonary fibrosis. Cochrane Database Syst Rev 2010(9):Cd003134.

88. Lawrenson JG, Evans JR. Omega 3 fatty acids for preventing or slowing the progression of age-related macular degeneration. Cochrane Database Syst Rev 2015;2015(4):Cd010015.

89. Mercado N, Wijns W, Serruys PW, et al. One-year outcomes of coronary artery bypass graft surgery versus percutaneous coronary intervention with multiple stenting for multisystem disease: a meta-analysis of individual patient data from randomized clinical trials. The Journal of thoracic and cardiovascular surgery 2005;130(2):512-9.

90. Qi WX, Shen Z, Lin F, et al. Overall survival benefits for irinotecan-containing regimens as first-line treatment for advanced gastric cancer: an updated meta-analysis of ten randomized controlled trials. International journal of cancer 2013;132(2):E66-73.

91. Nevitt SJ, Tudur Smith C, Marson AG. Oxcarbazepine versus phenytoin monotherapy for epilepsy: an individual participant data review. Cochrane Database Syst Rev 2018;10(10):Cd003615.

92. DIPART Group. Patient level pooled analysis of 68 500 patients from seven major vitamin D fracture trials in US and Europe. BMJ : British Medical Journal 2010;340:b5463.

93. Staropoli N, Ciliberto D, Botta C, et al. Pegylated liposomal doxorubicin in the management of ovarian cancer: a systematic review and metaanalysis of randomized trials. Cancer biology & therapy 2014;15(6):707-20.

94. Weis S, Franke A, Berg T, et al. Percutaneous ethanol injection or percutaneous acetic acid injection for early hepatocellular carcinoma. Cochrane Database Syst Rev 2015;1(1):Cd006745.

95. Ronellenfitsch U, Schwarzbach M, Hofheinz R, et al. Perioperative chemo(radio)therapy versus primary surgery for resectable adenocarcinoma of the stomach, gastroesophageal junction, and lower esophagus. Cochrane Database Syst Rev 2013(5):Cd008107.

96. Zhou M, Yu P, Qu X, et al. Phase III trials of standard chemotherapy with or without bevacizumab for ovarian cancer: a meta-analysis. PloS one 2013;8(12):e81858-e58.

97. Nolan SJ, Tudur Smith C, Pulman J, et al. Phenobarbitone versus phenytoin monotherapy for partial onset seizures and generalised onset tonic-clonic seizures. Cochrane Database Syst Rev 2013(1):Cd002217.

98. Nolan SJ, Marson AG, Pulman J, et al. Phenytoin versus valproate monotherapy for partial onset seizures and generalised onset tonic-clonic seizures. Cochrane Database Syst Rev 2013(8):Cd001769.

99. Barta SK, Lee JY, Kaplan LD, et al. Pooled analysis of AIDS malignancy consortium trials evaluating rituximab plus CHOP or infusional EPOCH chemotherapy in HIV-associated non-Hodgkin lymphoma. Cancer 2012;118(16):3977-83.

100. Seidman AD, Chan S, Wang J, et al. A pooled analysis of gemcitabine plus docetaxel versus capecitabine plus docetaxel in metastatic breast cancer. The oncologist 2014;19(5):443-52.

101. Blum JL, Barrios CH, Feldman N, et al. Pooled analysis of individual patient data from capecitabine monotherapy clinical trials in locally advanced or metastatic breast cancer. Breast cancer research and treatment 2012;136(3):777-88.

102. Bischoff-Ferrari HA, Willett WC, Orav EJ, et al. A Pooled Analysis of Vitamin D Dose Requirements for Fracture Prevention. New England Journal of Medicine 2012;367(1):40-49.

103. Huang YY, Yang Q, Zhou SW, et al. Postoperative chemoradiotherapy versus postoperative chemotherapy for completely resected gastric cancer with D2 Lymphadenectomy: a meta-analysis. PloS one 2013;8(7):e68939.

104. PORT Meta-analysis Trialists Group. Postoperative radiotherapy for non-small cell lung cancer. Cochrane Database Syst Rev 2005(2):Cd002142.

105. Ndrepepa G, Neumann FJ, Richardt G, et al. Prognostic value of access and non-access sites bleeding after percutaneous coronary intervention. Circulation Cardiovascular interventions 2013;6(4):354-61.

106. Lehert P, Chéron G, Calatayud GA, et al. Racecadotril for childhood gastroenteritis: an individual patient data meta-analysis. Digestive and liver disease : official journal of the Italian Society of Gastroenterology and the Italian Association for the Study of the Liver 2011;43(9):707-13.

107. Bohlius J, Schmidlin K, Brillant C, et al. Recombinant human erythropoiesis-stimulating agents and mortality in patients with cancer: a meta-analysis of randomised trials. Lancet 2009;373(9674):1532-42.

108. D'Amico AV, Denham JW, Bolla M, et al. Short- vs long-term androgen suppression plus external beam radiation therapy and survival in men of advanced age with node-negative high-risk adenocarcinoma of the prostate. Cancer 2007;109(10):2004-10.

109. Wang J, Zou ZH, Xia HL, et al. Strengths and weaknesses of immunotherapy for advanced non-small-cell lung cancer: a meta-analysis of 12 randomized controlled trials. PloS one 2012;7(3):e32695.

110. Filardo G, Powell JT, Martinez MA, et al. Surgery for small asymptomatic abdominal aortic aneurysms. Cochrane Database Syst Rev 2015;2015(2):Cd001835.

111. Soon YY, Tham IW, Lim KH, et al. Surgery or radiosurgery plus whole brain radiotherapy versus surgery or radiosurgery alone for brain metastases. Cochrane Database Syst Rev 2014;2014(3):Cd009454.

112. Sladden MJ, Balch C, Barzilai DA, et al. Surgical excision margins for primary cutaneous melanoma. Cochrane Database Syst Rev 2009(4):Cd004835.

113. Shen A, Tang C, Wang Y, et al. A systematic review of sorafenib in Child-Pugh A patients with unresectable hepatocellular carcinoma. Journal of clinical gastroenterology 2013;47(10):871-80.

114. Traut U, Brügger L, Kunz R, et al. Systemic prokinetic pharmacologic treatment for postoperative adynamic ileus following abdominal surgery in adults. Cochrane Database Syst Rev 2008(1):Cd004930.

115. Hart MG, Garside R, Rogers G, et al. Temozolomide for high grade glioma. Cochrane Database Syst Rev 2013;2013(4):Cd007415.

116. Cooper CB, Anzueto A, Decramer M, et al. Tiotropium reduces risk of exacerbations irrespective of previous use of inhaled anticholinergics in placebo-controlled clinical trials. International journal of chronic obstructive pulmonary disease 2011;6:269-75.

117. Morriss R, Vinjamuri I, Faizal MA, et al. Training to recognise the early signs of recurrence in schizophrenia. Cochrane Database Syst Rev 2013(2):Cd005147.

118. Benatar M, Kurent J, Moore DH. Treatment for familial amyotrophic lateral sclerosis/motor neuron disease. Cochrane Database Syst Rev 2009;2009(1):Cd006153.

119. Abraham WT, Anand IS, Klapholz M, et al. Treatment of anemia with darbepoetin alfa in heart failure. Congestive heart failure (Greenwich, Conn) 2010;16(3):87-95.

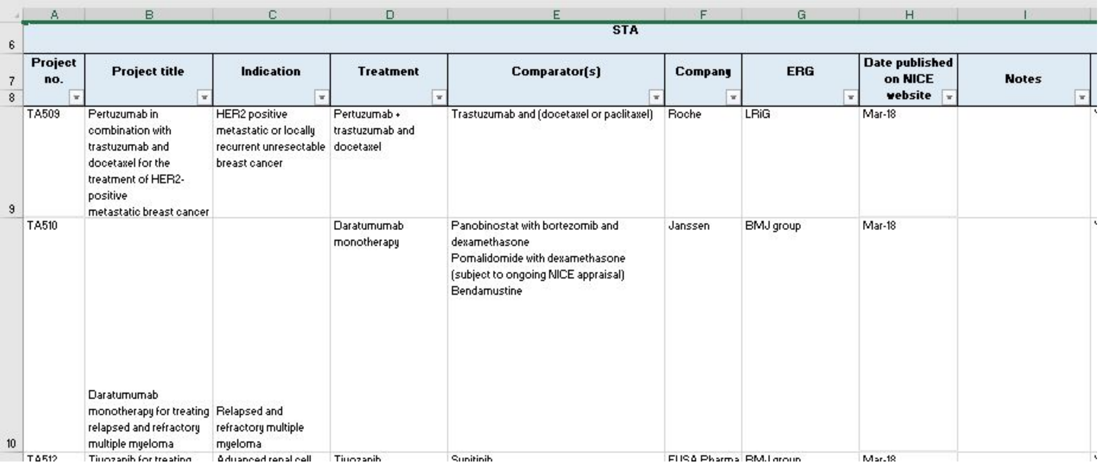
120. Franko J, Shi Q, Goldman CD, et al. Treatment of colorectal peritoneal carcinomatosis with systemic chemotherapy: a pooled analysis of north central cancer treatment group phase III trials N9741 and N9841. Journal of clinical oncology : official journal of the American Society of Clinical Oncology 2012;30(3):263-7.

121. Maier PC, Funk J, Schwarzer G, et al. Treatment of ocular hypertension and open angle glaucoma: meta-analysis of randomised controlled trials. BMJ (Clinical research ed) 2005;331(7509):134.

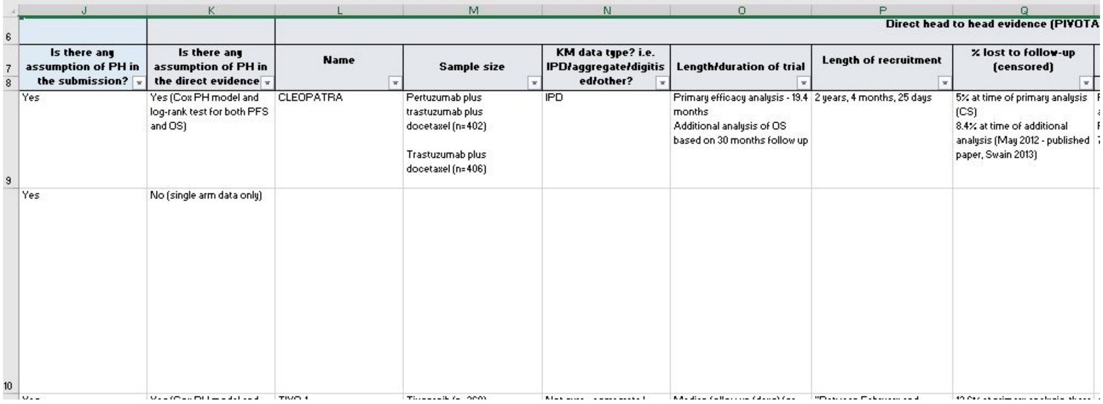
122. Rejnmark L, Avenell A, Masud T, et al. Vitamin D with calcium reduces mortality: patient level pooled analysis of 70,528 patients from eight major vitamin D trials. The Journal of clinical endocrinology and metabolism 2012;97(8):2670-81.

123. van Oostrom SH, Driessen MT, de Vet HC, et al. Workplace interventions for preventing work disability. Cochrane Database Syst Rev 2009(2):Cd006955.

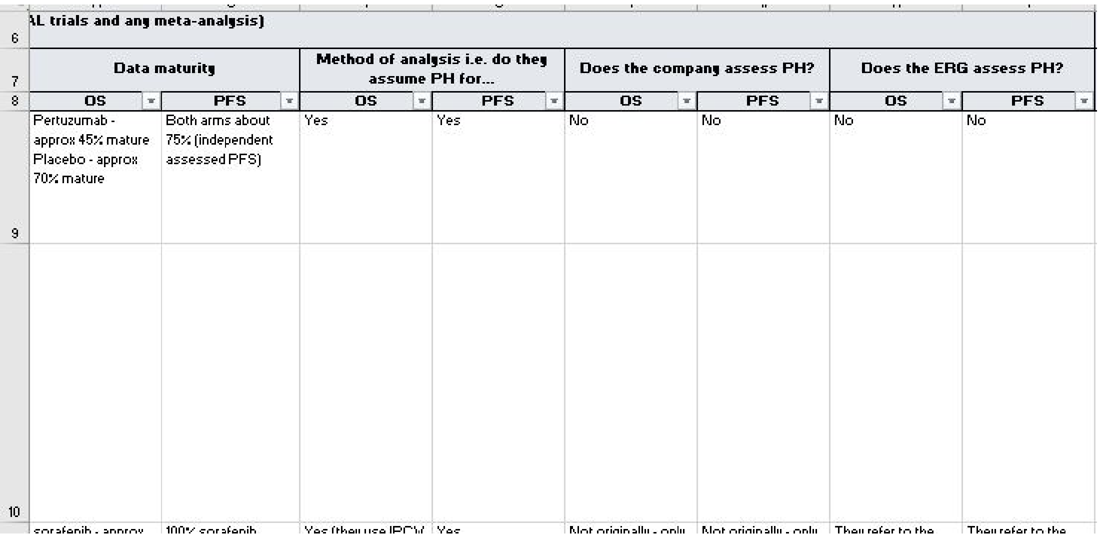
Appendix 7 : Screenshot of Microsoft Excel Database used in Chapter 6



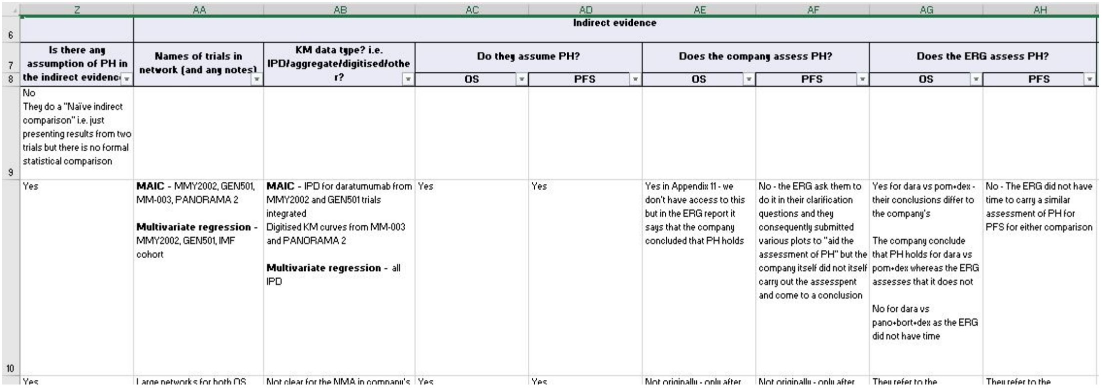
				STA				
Project no.	**Project title**	**Indication**	**Treatment**	**Comparator(s)**	**Company**	**ERG**	**Date published on NICE website**	**Notes**
TA509	Pertuzumab in combination with trastuzumab and docetaxel for the treatment of HER2-positive metastatic breast cancer	HER2 positive metastatic or locally recurrent unresectable breast cancer	Pertuzumab + trastuzumab and docetaxel	Trastuzumab and (docetaxel or paclitaxel)	Roche	LRiG	Mar-18	
TA510	Daratumumab monotherapy for treating relapsed and refractory multiple myeloma	Relapsed and refractory multiple myeloma	Daratumumab monotherapy	Panobinostat with bortezomib and dexamethasone Pomalidomide with dexamethasone (subject to ongoing NICE appraisal) Bendamustine	Janssen	BMJ group	Mar-18	
TA512	Tivozanib for treating	Advanced renal cell	Tivozanib	Sunitinib	EUSA Pharma	BMJ group	Mar-18	



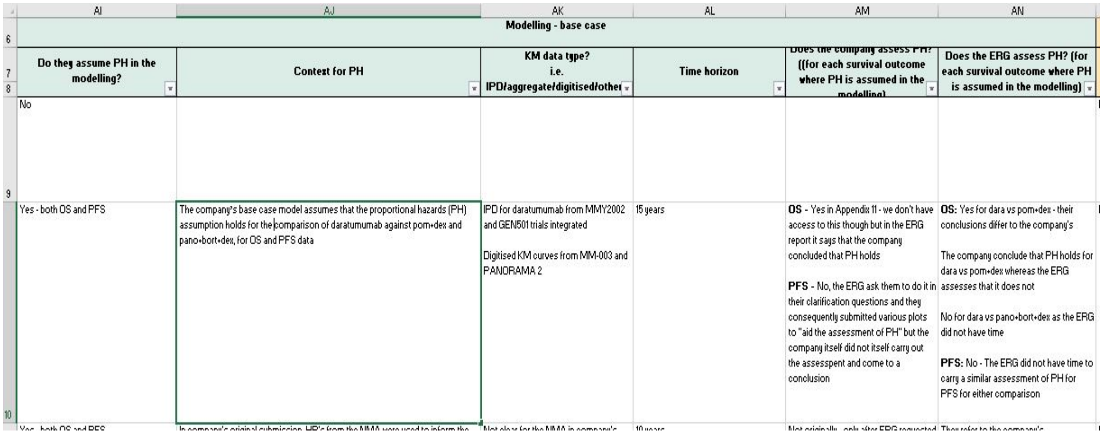
						Direct head to head evidence (PIVOTA	
Is there any assumption of PH in the submission?	**Is there any assumption of PH in the direct evidence**	**Name**	**Sample size**	**KM data type? i.e. IPD/aggregate/digitised/other?**	**Length/duration of trial**	**Length of recruitment**	**% lost to follow-up (censored)**
Yes	Yes (Cox PH model and log-rank test for both PFS and OS)	CLEOPATRA	Pertuzumab plus trastuzumab plus docetaxel (n=402) Trastuzumab plus docetaxel (n=406)	IPD	Primary efficacy analysis - 19.4 months Additional analysis of OS based on 30 months follow up	2 years, 4 months, 25 days	5% at time of primary analysis (CS) 8.4% at time of additional analysis (May 2012 - published paper, Swain 2013)
Yes	No (single arm data only)						
Yes	Yes (Cox PH model and	TIVO-1	Tivozanib (n=260)	Not sure - aggregate /	Median follow up (days) (as	"Between February and	12.9% at primary analysis, these



AL trials and any meta-analysis)							
Data maturity		**Method of analysis i.e. do they assume PH for...**		**Does the company assess PH?**		**Does the ERG assess PH?**	
OS	**PFS**	**OS**	**PFS**	**OS**	**PFS**	**OS**	**PFS**
Pertuzumab - approx 45% mature Placebo - approx 70% mature	Both arms about 75% (independent assessed PFS)	Yes	Yes	No	No	No	No
sorafenib - approx	100% sorafenib	Yes (they use IPCW	Yes	Not originally - only	Not originally - only	They refer to the	They refer to the



	Z	AA	AB	AC	AD	AE	AF	AG	AH
6		Indirect evidence							
7	Is there any assumption of PH in the indirect evidence	Names of trials in network (and any notes)	KM data type? i.e. IPD/aggregate/digitised/other?	Do they assume PH?		Does the company assess PH?		Does the ERG assess PH?	
8				OS	PFS	OS	PFS	OS	PFS
9	No They do a "Naïve indirect comparison" i.e. just presenting results from two trials but there is no formal statistical comparison								
10	Yes	**MAIC** - MMY2002, GEN501, MM-003, PANORAMA 2 **Multivariate regression** - MMY2002, GEN501, IMF cohort	**MAIC** - IPD for daratumumab from MMY2002 and GEN501 trials integrated Digitised KM curves from MM-003 and PANORAMA 2 **Multivariate regression** - all IPD	Yes	Yes	Yes in Appendix 11 - we don't have access to this but in the ERG report it says that the company concluded that PH holds	No - the ERG ask them to do it in their clarification questions and they consequently submitted various plots to "aid the assessment of PH" but the company itself did not itself carry out the assesspent and come to a conclusion	Yes for dara vs pom+dex - their conclusions differ to the company's The company conclude that PH holds for dara vs pom+dex whereas the ERG assesses that it does not No for dara vs pano+bort+dex as the ERG did not have time	No - The ERG did not have time to carry a similar assessment of PH for PFS for either comparison
	Yes	Large networks for both OS	Not clear for the NMA in company's	Yes	Yes	Not originally - only after	Not originally - only after	They refer to the	They refer to the



	AI	AJ	AK	AL	AM	AN
6	Modelling - base case					
7	Do they assume PH in the modelling?	Context for PH	KM data type? i.e. IPD/aggregate/digitised/other	Time horizon	Does the company assess PH? ((for each survival outcome where PH is assumed in the modelling)	Does the ERG assess PH? (for each survival outcome where PH is assumed in the modelling)
9	No					
10	Yes - both OS and PFS	The company's base case model assumes that the proportional hazards (PH) assumption holds for the comparison of daratumumab against pom+dex and pano+bort+dex, for OS and PFS data	IPD for daratumumab from MMY2002 and GEN501 trials integrated Digitised KM curves from MM-003 and PANORAMA 2	15 years	**OS** - Yes in Appendix 11 - we don't have access to this though but in the ERG report it says that the company concluded that PH holds **PFS** - No, the ERG ask them to do it in their clarification questions and they consequently submitted various plots to "aid the assessment of PH" but the company itself did not itself carry out the assesspent and come to a conclusion	**OS:** Yes for dara vs pom+dex - their conclusions differ to the company's The company conclude that PH holds for dara vs pom+dex whereas the ERG assesses that it does not No for dara vs pano+bort+dex as the ERG did not have time **PFS:** No - The ERG did not have time to carry a similar assessment of PH for PFS for either comparison
	Yes - both OS and PFS	In company's original submission, HR's from the NMA were used to inform the	Not clear for the NMA in company's	10 years	Not originally - only after ERG requested	They refer to the company's

Appendix 8 : References of 31 Single Technology Appraisals included in Chapter 6

1. Riemsma R, Büyükkaramikli N, De Groot S, Fayter D, Armstrong N, Wei C-Y, et al. Ribociclib in combination with an aromatase inhibitor for previously untreated advanced or metastatic hormone receptor-positive, HER2-negative breast cancer: a Single Technology Assessment. York: Kleijnen Systematic Reviews Ltd, 2017

2. Fleeman N, Bagust A, Richardson M, Boland A, Krishan A, Beale S, et al. Nivolumab for previously treated locally advanced or metastatic squamous-cell non-small cell lung cancer [ID811]: A Single Technology Appraisal. LRiG, University of Liverpool, 2015

3. Dickson R, Hounsome J, Stainthorpe A, Abdulla A, Bagust A, Richardson M, et al. Nivolumab for previously treated locally advanced or metastatic non-squamous non-small cell lung cancer [ID900]: A Single Technology Appraisal. LRiG, University of Liverpool, 2016

4. Venetoclax for treating chronic lymphocytic leukaemia: A Single Technology Appraisal. Warwick Evidence, 2016.

5. Jones-Hughes T, Dunham J, Robinson S, Napier M, Hoyle M. Regorafenib for previously treated unresectable or metastatic gastrointestinal stromal tumours: A Single Technology Appraisal. Peninsula Technology Assessment Group (PenTAG), 2017.

6. Armstrong N, Ramaekers BLT, Pouwels X, Zaim R, Wolff RF, Riemsma RR, et al. Nivolumab for treating recurrent or metastatic squamous-cell carcinoma of the head and neck after platinum-based chemotherapy: a Single Technology Assessment. York: Kleijnen Systematic Reviews Ltd, 2016.

7. Tappenden P, Carroll C, Stevens J, Simpson E, Thokala P, Sanderson J, et al. Ibrutinib for treating Waldenström's macroglobulinaemia: A Single Technology Appraisal. School of Health and Related Research (ScHARR), 2016.

8. Greenhalgh J, Bagust A, Stainthorpe A, Richardson M, Boland A, Beale S, et al. Paclitaxel as albumin-bound nanoparticles with gemcitabine for untreated metastatic pancreatic cancer [ID1058]: A Single Technology Appraisal. LRiG, University of Liverpool, 2017

9. Edwards SJ, Wakefield V, Cain P, Jhita T, Masento N, Salih F, et al. Cabozantinib for previously treated advanced renal cell carcinoma: A Single Technology Appraisal. BMJ Technology Assessment Group, 2016.

10. Tikhonova I, Jones-Hughes T, Dunham J, Warren, F, Robinson S, Stephens P, et al. Olaratumab in combination with doxorubicin for treating advanced soft tissue sarcoma: A Single Technology Appraisal. Peninsula Technology Assessment Group (PenTAG), 2017.

11. Edwards SJ, Wakefield V, Bacelar M, Salih F, Masento N, Karner C. Vismodegib for treating basal cell carcinoma: A Single Technology Appraisal. BMJ-TAG, 2017.

12. Boyers D, Cruickshank M, Jacobsen E, Cooper D, Fraser C, Culligan D, et al. Brentuximab vedotin for relapsed or refractory systemic anaplastic large cell lymphoma. Aberdeen HTA Group, 2017.

13. Edwards SJ, Bacelar M , Barton S, Karner C, Masento N, Salih F. Dartumumab for treating relapsed and refractory multiple myeloma: A Single Technology Appraisal. BMJ-TAG, 2017.

14. Edwards SJ, Kew KM, Jhita T, Barton S, Salih F, Masento N. Tivozanib for treating renal cell carcinoma: A Single Technology Appraisal. BMJ Technology Assessment Group, 2017.

15. Riemsma R, Corro Ramos I, Thielen F, Fayter D, Armstrong N, Wei C-Y, et al. Obinutuzumab for untreated advanced follicular lymphoma: a Single Technology Assessment. York: Kleijnen Systematic Reviews Ltd, 2017.

16. Stevenson M., Tappenden P, Rawdin A. Regorafenib for previously treated unresectable hepatocellular carcinoma: A Single Technology Appraisal – Rapid Review. School of Health and Related Research (ScHARR), 2018.

17. Eribulin for treating locally advanced or metastatic breast cancer after chemotherapy [ID1072]: A Single Technology Appraisal. LRiG, University of Liverpool, 2017

18. Ixazomib citrate in combination with lenalidomide and dexamethasone for relapsed refractory multiple myeloma: A Single Technology Appraisal. Warwick Evidence, 2017.

19. Edwards SJ, Karner C, Cain P, Masento N, Salih F, Wakefield V. Lenvatinib with everolimus for previously treated advanced renal cell carcinoma: A Single Technology Appraisal. BMJ-TAG, 2017.

20. Claxton L, Woolacott N, O'Connor J, Wright K, Hodgson R. Ceritinib for untreated anaplastic lymphoma kinase-positive advanced non-small-cell lung cancer: A Single Technology Appraisal. CRD and CHE Technology Assessment Group, 2017.

21. Tappenden P, Simpson E, Sanderson J, Pollard D, Clowes M, Kaltenthaler E, et al. Ibrutinib for treating relapsed or refractory mantle cell lymphoma: A Single Technology Appraisal. School of Health and Related Research (ScHARR), 2016.

22. Cooper, K, Kalita, N, Harris, P, Onyimadu, O, Gaisford, W, Picot, J. Fulvestrant for untreated hormone-receptor positive locally advanced or metastatic breast cancer: A Single Technology Appraisal. Southampton Health Technology Assessments Centre (SHTAC), 2017.

23. Cooper K, Kalita N, Onyimadu O, Colquitt JL, Loveman E, Frampton GK. Atezolizumab for treating locally advanced or metastatic urothelial carcinoma: A Single Technology Appraisal.

Southampton Health Technology Assessments Centre, 2017.

24. Fleeman N, Stainthorpe A, Bagust A, Richardson M, Nolan S, Houten R, et al. Palbociclib in combination with an aromatase inhibitor for previously untreated metastatic, hormone receptor-positive, HER2- breast cancer [ID915]: A Single Technology Appraisal. LRiG, University of Liverpool, 2016

25. Cooper, K, Rose, M, Harris P, Chorozoglou, M, Picot, J. Nivolumab for treating relapsed or refractory classical Hodgkin lymphoma: A Single Technology Appraisal. Southampton Health

Technology Assessments Centre (SHTAC), 2017

26. Brentuximab vedotin for treating CD30-positive Hodgkin lymphoma: A Single Technology Appraisal. BMJ Technology Assessment Group, 2017.

27. Greenhalgh J, Mahon J, Boland A, Beale S, Krishan A, Abdulla A, et al. Pembrolizumab for untreated PD-L1 positive metastatic non-small cell lung cancer [ID990]: A Single Technology Appraisal. LRiG, University of Liverpool, 2016.

28. Blinatumomab for treating Philadelphiachromosome-negative relapsed or refractory acute lymphoblastic leukaemia. A single technology appraisal. Warwick Evidence, February 2017.

29. Pandor A, Stevenson M, Martyn-St James M, Stevens J, Hamilton J, Rawdin A, et al. Ponatinib for treating chronic myeloid leukaemia: A Single Technology Appraisal. School of Health and Related Research (ScHARR), 2016.

30. Fleeman N, Abdulla A, Bagust A, Beale S, Richardson M, Stainthorpe A, et al. Pegylated liposomal irinotecan hydrochloride trihydrate for treating pancreatic cancer after gemcitabine [ID778]: A Single Technology Appraisal. LRiG, University of Liverpool, 2016

31. Pertuzumab in combination with trastuzumab and docetaxel for the treatment of HER2 positive metastatic or locally recurrent unresectable breast cancer [ID523]: A Single Technology Appraisal. LRiG, University of Liverpool, 2017

Appendix 9 : Mean and Range of Censoring level used for each DGM

DGM	Mean Censoring Level (%)	Range of Censoring Level (%)
1	28.2	25.4 to 31.0
2	26.6	23.6 to 29.8
3	20.7	17.5 to 23.4
4	68.4	65.5 to 71.0
5	67.5	64.5 to 70.6
6	62.6	59.7 to 65.6
7	28.0	25.3 to 30.7
8	26.6	23.5 to 29.6
9	20.4	18.0 to 23.0
10	68.1	65.5 to 71.0
11	68.0	65.1 to 70.7
12	63.5	60.4 to 67.3