Analysis of treatment effects and event rates that change over time in clinical trials

Kim Maree Jachno

BSc(Hons), MBiostat

A thesis submitted for the degree of
Doctor of Philosophy at Monash University in 2021
School of Public Health and Preventive Medicine

Copyright notice

Contents

I. Abstract

Introduction: Time-to-event analysis, or survival analysis, is the most widely utilized analytical method applied to outcomes of clinical trials. Most clinical trials with time-to-event outcomes are designed assuming constant event rates and proportional hazards however, nonproportional hazards are seen increasingly frequently in trials. The impact of non-constant event rates and nonproportional hazards and the interplay between these two factors on the design, use of statistical methodology and reporting of trials has not been evaluated.

Aims: The aims of this thesis are : to assess whether non-constant event rates and non-proportional hazards were allowed for in the design, analysis and reporting of trials, to investigate the impact of non-constant event rates in the presence of non-proportionality, to illustrate the potential gains in understanding and clinical insight that may be possible using analysis methods which allow for time-dependence of treatment effects, and to improve the awareness and reporting of treatment effects that change over time through visual presentations.

Methods: A scoping review was undertaken to assess how non-constant event rates and non-proportional treatment effects were allowed for in the design of trials, to determine the main methodological approaches used, and assess the reporting and presentation quality of trial findings. A simulation study was performed to investigate the impact of non-constant event rates in the presence of non-proportionality using statistical methods informed by the review for analysing time-to-event data. An application of regression-based methods which allow for time-dependent treatment effects was used to illustrate the potential for increased clinical insight into treatment effects and interactions in a trial. Finally, graphical means to improve the visual presentation of treatment effect measures was proposed as a way of improving the awareness of time-dependent treatment effects and provide impetus to more fully report and investigate trial findings.

Results: The review confirmed that when designing trials constant event rates and proportional hazards are typically assumed, that methods assuming proportional hazards are the predominant method to analyse trial results and that reporting of the key assumption was lacking. The simulation showed that even modest departures from non-constant event rates could further augment the loss in power to detect treatment effects depending on the nature of any nonproportionality. Through a re-examination of endpoints, we found evidence for nonproportionality, time-dependent treatment effects and treatment interaction effects not previously reported. We developed a series of recommendations to improve the reporting of clinical trials through the use of treatment effect plots.

Conclusions: The research in this thesis demonstrates that allowing for non-constant event rates and nonproportionality in the design, analysis and reporting of clinical trials can still be improved. Nonproportionality is being observed more frequently due to the mechanistic nature of new interventions and because of increased regulatory oversight requiring the conduct of larger, longer trials. Illustrating the increased insight and clinical understanding that can be obtained through the use of more recently developed analysis approaches combined with our proposed presentations of complementary graphs should provide the impetus to more fully report clinical trial findings.

II. Thesis including published works declaration

I hereby declare that this thesis contains no material which has been accepted for the award of any other degree or diploma at any university or equivalent institution and that, to the best of my knowledge and belief, this thesis contains no material previously published or written by another person, except where due reference is made in the text of the thesis.

This thesis includes one original paper published in peer reviewed journals and three submitted publications. The core theme of the thesis is the analysis of time-dependent treatment effects in the presence of non-constant event rates in clinical trials. The ideas, development and writing up of all the papers in the thesis were the principal responsibility of myself, the student, working within the School of Public Health and Preventive Medicine under the supervision of Professor Rory Wolfe and Professor Stephane Heritier.

The inclusion of co-authors reflects the fact that the work came from active collaboration between researchers and acknowledges input into team-based research.

In the case of chapters two to five my contribution to the work involved the following:

Thesis Chapter	**Publication Title**	**Status** *(published, in press, accepted or returned for revision, submitted)*	**Nature and % of student contribution**	**Co-author name(s) Nature and % of Co-author's contribution***	**Co-author(s), Monash student Y/N***
2	Are non-constant event rates and non-proportional treatment effects accounted from in the design and analysis of randomised controlled trials? A review of current practice	Published in peer reviewed journal: BMC Medical Research Methodology	80%. Led the concept and design of the review and data extraction. Created the database collection tool and analysis code. Wrote first draft and subsequent revisions based on critical review from co-authors.	1) Rory Wolfe 12% Contributed to the design of the study and drafting of the manuscript. Provided critical review of the manuscript. 2) Stephane Heritier 8% Contributed to the design of the study and drafting of the manuscript. Provided critical review of the manuscript.	No for all

Thesis Chapter	Publication Title	**Status** *(published, in press, accepted or returned for revision, submitted)*	Nature and % of student contribution	Co-author name(s) Nature and % of Co-author's contribution*	Co-author(s), Monash student Y/N*
3	Impact of a non-constant baseline hazard on detection of time-dependent treatment effects: A simulation study	Returned for revison from peer reviewed journal: BMC Medical Research Methodology	80%. Led the design of the study, selected methods and scenarios. Designed the computer code, ran and analysed the simulations. Wrote first draft and all subsequent revisions based on critical review from co-authors.	1) Rory Wolfe 12% Contributed to the design of the study and drafting of the manuscript. Provided critical review of the manuscript. 2) Stephane Heritier 8% Contributed to the design of the study and drafting of the manuscript. Provided critical review of the manuscript.	No for all
4	Examining evidence for time-dependent treatment effects using alternative regression-based methods in clinical trials	Under review in peer reviewed journal: Pharmaceutical Statistics	80%. Led the design of the study, and obtained authorisations to secure data sharing platform. Designed the computer code, ran and analysed the simulations. Wrote first draft and all subsequent revisions based on critical review from co-authors.	1) Rory Wolfe 5% Contributed to the design of the study and drafting of the manuscript. Provided critical review of the manuscript. 2) Stephane Heritier 3% Provided critical review of the manuscript. 3) Robyn L Woods 2% Provided critical review of the manuscript. 4) Suzanne Mahady 2% Provided critical review of the manuscript. 5) Andrew T Chan 2% Provided critical review of the manuscript. 6) Andrew Tonkin 2% Provided critical review of the manuscript. 7) Anne Murray 2% Provided critical review of the manuscript. 8) John J McNeil 2% Provided critical review of the manuscript.	No to all

Thesis Chapter	Publication Title	Status *(published, in press, accepted or returned for revision, submitted)*	Nature and % of student contribution	Co-author name(s) Nature and % of Co-author's contribution*	Co-author(s), Monash student Y/N*
5		Submitted to peer reviewed journal: Trials	80%. Led the design of the study, obtained authorisations to secure data sharing platform. Designed the computer code, ran and analysed the simulations. Wrote first draft and all subsequent revisions based on critical review from co-authors.	1) Rory Wolfe 12% Contributed to the design of the study and drafting of the manuscript. Provided critical review of the manuscript. 2) Stephane Heritier 8% Contributed to the design of the study and drafting of the manuscript. Provided critical review of the manuscript.	No for all

I have renumbered sections of submitted or published papers in order to generate a consistent presentation within the thesis.

Student name: Kim Jachno

Student signature: **Date: 28/06/2021**

I hereby certify that the above declaration correctly reflects the nature and extent of the student's and co-authors' contributions to this work. In instances where I am not the responsible author I have consulted with the responsible author to agree on the respective contributions of the authors.

Main Supervisor name: Rory Wolfe

Main Supervisor signature: **Date: 18/07/2021**

III. List of research outputs

Listed below are the candidate's first author publications and conference proceedings that are relevant to the period of candidature

Publications relevant to the thesis

Jachno Kim, Heritier Stephane, Wolfe Rory. Are non-constant rates and non-proportional treatment effects accounted for in the design and analysis of randomised controlled trials? A review of current practice. BMC Med Res Methodol. 2019;19(1):103.

Jachno Kim, Heritier Stephane, Wolfe Rory. Impact of a non-constant baseline hazard on detection of time-dependent treatment effects: a simulation study. Submitted to BMC Med Res Methodol. Accepted, undergoing editorial revision

Jachno Kim, Heritier Stephane, Woods Robyn, Mahady Suzanne, Chan Andrew T., Tonkin, Andrew, Murray Anne, McNeil John J., Wolfe Rory. Examining evidence for time-dependent treatment effects using alternative regression-based methods in clinical trials. Submitted to Pharmaceutical Statistics. Under review

Jachno Kim, Heritier Stephane, Wolfe Rory. Complementing the Kaplan-Meier plot to enable assessment of treatment effect consistency with proportional hazards. Submitted to Trials. Under consideration

Conference proceedings:

Jachno Kim, Heritier Stephane, Wolfe Rory. Accounting for non-constant rates and time-dependent treatment effects when designing, analysing and reporting clinical trials: a review. Oral presentation at the Young Statisticians Conference, 1-2 October 2019, Canberra, Australia

Jachno Kim, Heritier Stephane, Wolfe Rory. Impact of the hazard rate on pre-specified methods of analysis in the presence of time-dependent treatment effects. Presented at the 5th International Clinical Trials Methodology Conference, 6-9th October 2019, Brighton, United Kingdom

Jachno Kim, Heritier Stephane, Wolfe Rory. Design and analysis of clinical trials with treatment effects and event rates that change over time. Invited speaker for the "Topics in Innovative Clinical Trials" session of the Biostatistics and Bioinformatics Section at Australia and New Zealand Statistical Conference 2020 (cancelled due to Covid-19)

IV. Acknowledgements

I would like to thank my supervisors, Professor Rory Wolfe and Professor Stephane Heritier for their help, patience and guidance during this PhD journey. They have been consistently generous with their time and support over the last few years and have enabled me to explore options and areas of research interest.

I am also very grateful to the co-authors of one publication: Associate Professor Robyn Woods, Dr Suzanne Mahady, Professor Andrew T. Chan, Professor Andrew Tonkin, Professor Anne Murray and Professor John McNeil. They provided excellent feedback and advice from a clinician's perspective during the publication drafts and were very encouraging of the impact of the final submission.

I would also like to thank my colleagues from the Biostatistics Unit of the School of Public Health and Preventive Medicine for their support – including but not limited to - Cath Smith, Sam Brilleman, Sarah Arnup, Jess Kazka, Pam Simpson, Kelsey Grantham, Simon Turner, Lizzie Korevaar, Matt Page and Miranda Cumpston. Special thanks to the Graduate Research Office for all their support and administrative guidance over the past few years, in particular the incomparable Kathryn Daly, Dr Liz Douglas and Professor Sally Green.

This research was supported in part by an Australian Government Research Training Program (RTP) Stipend and RTP Fee-Offset through Federation University Australia and top up funding from a National Health and Medical Research Council of Australia grant (APP1128222).

Finally, thanks to my family: firstly, to my three greatest achievements, Patrick, Emily and Alistair, who provided the motivation to undertake this study in the first place, and secondly but most importantly, to my wonderful partner David who has picked up the slack of all the parts of our lives that I wasn't able to fit in during the past four tumultuous years; he is the mainstay that has made it possible for me to persevere and complete this work. I'm looking forward to getting to spend so much more time with you all in the future.

V. List of Abbreviations

∆RMST	Difference in restricted mean survival time
∆S(t)	Difference in survival curve probability
AFT	Accelerated failure time
ASPREE	ASPirin in Reducing Events in the Elderly
CI	Confidence interval
CONSORT	Consolidated Standards of Reporting Trials
df	Degrees of freedom
DGM	Data-generating model
FH	Fleming-Harrington
FPM	Flexible parametric model
HR	Hazard ratio
HR(t)	Time-dependent hazard ratio plot
h(t)	Hazard function
H(t)	Cumulative hazard function
ICMJE	International Committee of Medical Journal Editors
IPD	Individual patient dataset
KM	Kaplan-Meier
LM	Landmark
LR	Logrank
MCSE	Monte Carlo standard error
MST	Mean survival time
PE	Piecewise exponential
PH	Proportional hazards
p-obs	Pseudo-observations
PRISMA	Preferred Reporting Items for Systematic reviews and Meta-Analyses
S(t)	Survival function
RCT	Randomized controlled trial
RMST	Restricted mean survival time
RP	Royston-Parmar
SSC	Sample size calculation
STE	Scaled treatment effect
TD	Time dependent/time dependence
TR	Time ratio

Chapter 1 Introduction

Randomised controlled trials provide the highest level of evidence on which to base decisions regarding the use of health interventions in humans. Time-to event analysis, or survival analysis has been the most widely utilised analytical method in research articles in leading general medical journals over the past two decades [1]. Most trials with time-to-event outcomes are designed assuming proportional hazards of the treatment effect and the hazard ratio from a Cox proportional hazards model has become ubiquitous as the method for quantifying treatment effects [2].

Proportional hazards (PH) implies that the effect of treatment - or any covariate - is constant at all times during the trial such that a fixed magnitude estimate obtained by taking the ratio of the two hazards ie the hazard ratio (HR) is an appropriate way to summarise the treatment effect. Nonproportionality – or treatment effects that may vary over time - is detected in larger trials, trials with long term follow up and trials that study treatments with novel mechanisms of action, characteristics that have become more commonplace [3–5]. With the advent of immunotherapies for cancer treatments, many new treatments exhibit a delay prior to any beneficial effect as activation of the immune system is required. Within a trial, hazards may also not be proportional for different observed subgroups of patients. Unobserved disease susceptibility or frailty that varies between individuals can also result in hazards that are not proportional as those with greater outcome susceptibility are likely to experience the event of interest earlier. As a result, comparisons of treatment groups later in the trial can differ from comparisons earlier in the trial. Ignoring time-varying effects and estimating “average” hazard ratios can result in misleading conclusions [6].

The reliance on the Cox model as the method for time-to-event analyses in clinical trials can be restricting. Information about the underlying event rate is also of interest to trialists and while this can be recovered for the Cox model, it is not directly estimated in a standard approach. In addition, when designing trials where the Cox model is used for analysis, an assumption of constant event rates is typically used, mainly for simplicity. Sample size calculations assuming constant or piecewise constant event rates are applied even when prior information on the shape of the underlying event rate is available [7]. When nonproportionality of treatment effects could be anticipated, there is limited research on the impacts of non-constant event rates on the Cox PH model HR estimand or other estimate of treatment effect [8].

The aims of this thesis are set out at the end of this chapter but briefly summarised are: (i) to assess whether non-constant event rates and nonproportional hazards were allowed for in the design, analysis and reporting of clinical trials in medical research literature, (ii) to investigate the impact of non-constant event rates on the power to detect treatment effects in the presence of nonproportionality, (iii) to illustrate the potential gains in understanding and clinical insight that may be possible using analysis methods which allow for time-dependence (TD) of treatment effects, and (iv) to improve the awareness and reporting of treatment effects that change over time through visual presentations.

This first chapter sets the context for the research presented in this thesis by providing a brief overview of time-to-event analysis covering the key functions and estimands, different estimation and sample size calculation approaches and outlining the existing guidance and reviews for designing, analysing and presenting time-to-event outcomes. This introduction concludes with the detailed aims and objectives of the research and presents an outline of the thesis structure.

1.1 Key functions and measures in time-to-event analysis

Time-to-event analysis refers to the statistical methods which analyse the time it takes for an event of interest to occur from some reference or baseline origin time. The analysis of observations of time-to-event data is also commonly referred to as survival analysis as early work using these methods often used death as the event occurrence. Time-to-event data is unique because the outcome under investigation is a time to event of interest and for some study participants that event may not have occurred during the period of time they were under observation, i.e. the data have been censored.

The most common type of censoring is right censoring, where up to a certain time point some participants have not yet experienced the event but are no longer followed. We may expect that sometime in the future the event can be observed but within the time period of observation, the event is not experience by such participants. Time-to-event analysis techniques utilise the partial information provided by each participant with censored data to obtain unbiased estimates of measures of importance.

There are three fundamental functions which describe the relationship between the event time T and the event of interest. These are the hazard function, the cumulative hazard function and the survival function. The hazard function $h(t)$ represents the instantaneous event rate at time t knowing that the participant has not experience this event so far, where

$$h(t) = \lim_{\delta \to 0} \left(\frac{Pr(t \leq T < t + \delta | T \geq t)}{\delta} \right)$$

The cumulative hazard function $H(t)$ is a measure of the accumulation of instantaneous risk of event occurrence, and is obtained by integration over the hazard function up to any time t

$$H(t) = \int_0^t h\,(u)\,du$$

The survival function $S(t)$ is the probability of the event of interest occurring after time t or alternatively the probability of being event-free at time t

$$S(t) = \Pr(T > t).$$

The survival function is monotonic decreasing and the cumulative hazard function is monotonic increasing. The hazard function can be any non-negative function, able to both increase and decrease over time. Mathematically, the three measures are related and can be written in terms of one another as

$$h(t) = -\frac{d \log S(t)}{dt}$$

$$S(t) = \exp\left[-\int_0^t h\,(u)\,du\right] = \exp[-H(t)]$$

From these functions, different measures can be constructed to enable quantification of covariate effects such as assignment to treatment group in a randomised trial.

Hazard Ratio (HR)

The hazard ratio is obtained by comparing the instantaneous event rate in the treatment group ($h_1(t)$, group code = 1) to the control group ($h_0(t)$, group code = 0). The effect of treatment is measured as the ratio of hazards in the treatment group to the control group. A typical assumption is that this ratio is constant over time, ie

$$\mathrm{HR} = \frac{h_1(t)}{h_0(t)} = \exp(\beta)$$

A generalisation of this idea will lead to the Cox model also known as the proportional hazards model or a parametric counterpart such as the Weibull model or flexible parametric models (these models are introduced in Sections 1.2.2 and 1.2.3). Of course, this constant PH assumption may not be true in practice paving the way for the definition of other measures of effect.

The HR does not have a clear interpretation when nonproportionality is observed and a large body of literature has been devoted to the development and use of alternative estimands [8–12]. The restricted mean survival time (RMST) is an example of a robust and clinically meaningful summary measure of survival time distribution that does not rely on the concept of hazard. A test of the difference of RMST (ΔRMST) between treatment groups may be more appropriate than a hazard ratio to determine treatment effects in the presence of non-proportionality of hazards between groups.

Difference in Restricted Mean Survival Time (ΔRMST)

The RMST μ of a time-to-event random variable T is the mean of $\min(T, t^*)$ where the cut off time t^* is greater than zero. The RMST can be derived as the area under the survival curve $S(t) = P(T > t)$ from $t = 0$ to $t = t^*$. In a two-group randomised trial with survival functions $S_{X_T}(t)$ and $S_{X_C}(t)$ for the treatment group and the control group respectively, the difference in RMST between groups can be defined as

$$\Delta\text{RMST} = \int_0^{t^*} S_{X_T}(t) - S_{X_C}(t)\, dt$$

An estimate of the ΔRMST can be obtained in a number of ways, including a method that is the focus in this thesis, by fitting a flexible parametric model either under the assumption of PH (ie equivalent to a time-fixed treatment effect) or allowing for non-PH (time-dependent treatment effects).

1.2 Analysis approaches

1.2.1 Non-parametric estimation of survival

Non-parametric approaches do not rely on assumptions about the shape or form of parameters in the underlying population. They are used to describe the data by estimating the survival function $S(t)$ and provide estimates of the median and centiles of survival time. These descriptive statistics, due to censoring, cannot be calculated directly from the data by ordering the observed event times and choosing the corresponding quantile.

The Kaplan-Meier (KM) estimator is the most used non-parametric method to estimate the survival function. It works by breaking up the estimation of $S(t)$ into intervals based on observed event times. Study participants contribute to the estimation of $S(t)$ until either the event occurs or the observation is censored. The KM estimate of the survival function is

$$\hat{S}(t) = \prod_{j|t_j \leq t} \left(\frac{n_j - d_j}{n_j}\right)$$

where $t_j, j = 1 \ldots K$ are the distinct ordered failure times, n_{t_j} is the number of participants at risk before time t_j and d_{t_j} the number of events observed at time t_j. Between events, the estimated survival probability is constant therefore the curve is a step function where vertical drops indicate the occurrence of one or more events.

Confidence intervals (CIs) which provide an estimate of the range of plausible values of the survival probability in the population which study participants represent can be calculated as $\hat{S}(t) \pm$

$z_{1-\alpha/2}$ se$[\hat{S}(t)]$ where the standard error (se) is calculated from the Greenwood formula for variance

$$\widehat{\mathrm{Var}}(\hat{S}(t)) = \hat{S}^2(t) \sum_{j|t_j \le t} \frac{d_j}{n_j(n_j - d_j)},$$

$z_{1-\alpha/2}$ is the $1 - \alpha/2$ critical value of the standard normal distribution and α is the nominated significance level.

Tests of survival curve difference

To test whether two or more groups of study participants have different survival time distributions, a hypothesis testing procedure can be employed to compare survival curves. Rank-based tests are well established for this purpose. There are several versions of these rank-based tests, which differ in the weight given to each time point in the calculation of the test statistic. The most common rank-based test utilised in the medical research literature is the logrank test which gives each event equal weighting.

Logrank test

The logrank test assesses the null hypothesis that there is no difference between the survival curves of two (or more) groups in the probability of an event at any time point over the total survival time period under consideration. This test compares observed (*O*) and expected (*E*) numbers of events across time and between groups. The analysis is based on the sum of differences of the estimated hazard function at each observed event time t_j with an implicit equal weighting of one for all event times. The test statistic is defined as

$$T_{LR} = \frac{\left[\sum_{j=1}^{K}(O_j - E_j)\right]^2}{\sum_{j=1}^{K} V_j} \sim \chi_1^2$$

where V_j is the variance of $\mathrm{Var}(O_j - E_j)$ for $j = 1, \dots, K$ event times.

Rank-based tests are subject to the assumptions that censoring is independent to outcome and group, that events happened at the times specified and survival probabilities are the same at all times. As a result of these assumptions, rank-based tests are maximally powerful to detect treatment effects when hazards are proportional.

Weighted logrank tests

When nonproportionality is anticipated, the logrank test can lose power to detect treatment differences with the magnitude of the loss of power dependent on the configuration of the nonproportionality. Variations of the logrank test include the Wilcoxon test, which weights each time point by the number of subjects at risk. Based on this weight, the Wilcoxon test is more sensitive to differences between curves early in the follow-up, when more subjects are at risk. Other tests, like the Peto-Prentice test, use weights with magnitude in between those of the logrank and Wilcoxon tests.

Fleming and Harrington [13] proposed a family of weighted tests, the extended $G^{\rho,\gamma}$ which can be expressed as

$$G^{\rho,\gamma} = \sqrt{\frac{n_1 + n_2}{n_1 n_2}} \int_0^{\infty} \{\hat{S}(t-)\}^{\rho}\{1 - \hat{S}(t-)^{\gamma}\} \frac{\overline{Y_1}(t)\overline{Y_2}(t)}{\overline{Y_1}(t) + \overline{Y_2}(t)} \left\{ \frac{d\overline{N_1}(t)}{\overline{Y_1}(t)} - \frac{d\overline{N_2}(t)}{\overline{Y_2}(t)} \right\}$$

where $\hat{S}(t-)$ is the Kaplan-Meier estimate of the survival rate based on the pooled data from the two treatment groups, $\overline{Y}_i(t)$ is the number of patients at risk in group i at time t, and $\overline{N}_i(t)$ is the number of events in group i up to and included time t. When $\rho = 0, \gamma = 0$ then $G^{0,0}$ corresponds to the logrank test with equal weights. When $\rho > \gamma$, the test gives more weight to earlier failures than to later ones, and when $\rho = 1, \gamma = 0$ corresponds to the generalised Wilcoxon test. When $\rho < \gamma$ more weight is given to later failures than to earlier ones. Commonly utilised Fleming-Harrington (FH) tests are $G^{1,0}$, $G^{1,1}$ and $G^{0,1}$ which preferentially weigh early, middle and latter events respectively.

When follow up duration is long, nonproportionality can occur. In this setting these rank-based tests have the limitation that they may be under powered to detect differences between groups under the assumption of PH. This drawback may be exacerbated if the nonproportionality is so marked that the survival curves cross.

Omnibus tests

Many tests of difference between two survival curves have been proposed that aim to achieve acceptable power under PH and under anticipated non-PH patterns whilst maintaining type I error rates close to the nominal level. Omnibus or global tests may be derived by combining some members within a class or across classes of test statistics. This can be useful in the presence of nonproportional hazards. A combined test assessed in this thesis utilises information from the logrank test and a test of difference in the mean survival time between treatment groups [14]. The motivation for the development of the combined test was to capitalise on the optimal power of the logrank test when the assumption of PH is met, and to provide some insurance should nonproportionality be present. Another omnibus test used in this thesis is the versatile test proposed by Karrison [15]. The default comparison test considers $\mathbf{Z_m} = \max(|Z_1|, |Z_2|, |Z_3|)$ where Z_1, Z_2 and Z_3 are Z statistics from $G^{0,0}$, $G^{1,0}$ and $G^{0,1}$ extended FH family, $\mathbf{Z}_m \sim \mathrm{N}_3(\mu, \Sigma)$ an asymptotic, trivariate normal distribution with $\boldsymbol{\mu}$ the vector of means and Σ the variance-covariance matrix. This combination of Z statistics was selected to provide relatively good coverage across the range of likely scenarios encompassing proportional hazards, early difference and late difference configurations.

1.2.2 Semi-parametric estimation of survival

The Cox proportional hazards model is the most common survival model and is formulated as

$$h_i(t) = h_0(t)\exp(X_i\beta)$$

with $h_0(t)$ the baseline hazard function, ie the hazard function when all covariates are equal to zero, X_i represents covariates, β are the estimated coefficients. In using partial maximum likelihood to fit the Cox PH model, only the coefficients β need to be estimated, not the baseline hazard so no absolute effects are estimated directly. Relative effects expressed as hazard ratios can be obtained by exponentiating the coefficients. The logrank statistic can be derived as the score test for the Cox PH model comparing two groups. It is therefore asymptotically equivalent to testing the β coefficient for treatment in the Cox PH model.

Nonproportional hazards can be introduced by including an interaction term between time and the covariate which is expected to have a time-dependent effect. Another possible way is to add a time-dependent function of time in the model, alone or as an interaction term with a specific covariate. However, it may be easier to estimate these effects when making some assumption about the shape of the underlying hazards through the use of parametric models.

1.2.3 Parametric estimation of survival

Weibull distribution

Parametric survival models offer many advantages over semi-parametric models. They provide smooth estimates of the hazard and survival functions for any combination of covariates. It is easier to include time-dependent effects and model on different scales. It is also easier to extrapolate and obtain out-of-sample predictions with parametric survival models compared to the Cox model. A commonly used distribution function assumed for the baseline hazard is the Weibull distribution which assumes the baseline hazard function $h_0(t) = \lambda\gamma t^{\gamma-1}$ with λ and γ positive valued parameters that determine the scale and shape of the distribution respectively. When $\gamma = 1$, a constant hazard is assumed and this corresponds to the exponential distribution. The Weibull distribution can capture a variety of increasing and decreasing event rate scenarios.

Assuming a Weibull distribution for the baseline hazard, a Weibull PH model can be written as

$$h_i(t) = \lambda\gamma t^{\gamma-1}\exp(X_i\beta)$$

For the Weibull PH model, the effect of treatment is obtained as

$$HR = \frac{\lambda\gamma t^{\gamma-1}\exp(\beta_C + \beta_T)}{\lambda\gamma t^{\gamma-1}\exp(\beta_C)} = \exp(\beta_T)$$

Flexible parametric models

A more flexible alternative to parametric regression models uses restricted cubic splines to model the baseline hazard first proposed by Royston and Parmar [16, 17]. These flexible parametric models (FPMs) are formulated by modelling survival times on the log cumulative hazard scale under an assumption of proportional hazards

$$\log H_i(t) = \log H_0(t) + X_i\beta = s(\log(t)|\boldsymbol{\gamma}_s, \mathbf{k}_0) + X_i\beta$$

where $s(\log(t)|\boldsymbol{\gamma}_s, \mathbf{k}_0)$ is the restricted cubic spline function with parameters $\boldsymbol{\gamma}_s$ for the baseline cumulative hazard with a vector of $\mathbf{k}_0$ knots. By derivation, the baseline hazards can then be estimated. In this way, the attraction of the Cox model - allowing the shape of the baseline hazard to be flexible through the absence of any distributional assumptions – can be achieved by allowing the basis function of cubic splines to flexibly fit the baseline hazard. FPMs have the additional appeal as parametric models of standard estimation options and interpretability, providing both relative and absolute estimates of treatment effect.

Restricted cubic splines are piecewise cubic polynomials joined together at knots locations with smoothing constraints placed on the knots, the restriction coming from imposing linear terms beyond the first and last knots. This restriction ensures that an overall smooth function is fitted and that the fit is not unduly affected by extreme observations. In general, FPMs are implemented on the log cumulative hazard scale using one set of spline variables with predefined knot positions based on centiles of uncensored log survival times depending on the number of knots, with boundary knots at the minimum and maximum uncensored log survival times. The number of knots used to model the baseline hazard can be guided by clinical input and model selection criteria.

FPMs can be generalised to accommodate nonproportional hazards. Time-dependent effects can be modelled using a different set of spline variables for each covariate of interest, possibly using a different number of knots in potentially different locations than the spline variables used to model the baseline hazard. Defining $\mathbf{k_0}$ to denote the number of knots for the baseline hazard function, $\mathbf{k}_j$ to denote the knots for the jth TD effect with associated parameters, $\boldsymbol{\delta}_j$ when there are J covariates with TD effects, the log cumulative hazard model is

$$\log H_i(t) = s\{\log(t)|\boldsymbol{\gamma}, \mathbf{k}_0\} + \sum_{j=1}^{J} s\,\{\log(t)|\boldsymbol{\delta}_j, \mathbf{k}_j\}\mathbf{x}_j + \mathbf{x}\boldsymbol{\beta}$$

1.3 Designing clinical trials - sample size and power

When designing a trial with a time-to-event outcome, where information is based on the number of events rather than the number of participants, there is importance in correct specification of the baseline hazard rate. Trials typically have fixed lengths of conduct, often composed of an accrual phase during which recruitment occurs, and a follow up phase where there is continued observation of event occurrence in the recruited participants. In trials of fixed duration, the interplay between the possibility of withdrawal and administrative censoring along with event rates needs to be taken into consideration in order to ensure that the chosen duration is sufficient to observe the required number of events.

Sample size formulae for comparing two survival distributions using the logrank test [18, 19], or the exponential survivor function [20, 21] assume constant event rates and proportional hazards. Almost equivalently, the sample size formula using the beta co-efficient in the Cox model [22] assumes proportional hazards. These are the most widely used methods to determine the number of patients needed in a trial. Under PH, the shape of the baseline hazard has no effect on power nor on the magnitude of estimated treatment effects using standard analytical approaches. However, in a non-PH context the appropriateness of analytical approaches can depend on the shape of the underlying hazard.

In the past two decades, there have been several proposed methods of sample size calculation (SSC) that acknowledge that the PH assumption may be too restrictive. These have included incorporating Fleming-Harrington weights into the SSC [23, 24], allowing for nonproportionality to be specified as a series of piecewise exponential ‘stages’ within a trial [25, 26], calculations that address specific types of nonproportionality such as lag to effect [24], using parametric modelling approaches to allow for changing event rates such as the Weibull distribution [27, 28] or the generalized gamma distribution [29], or using SSCs calculated assuming alternative model-free approach such as restricted mean survival time [14]. Alternatively, it is possible to use simulation strategies to determine the sample size required [30].

Whilst there has been some assessment of the adequacy of sample size reporting in general [31, 32], the uptake of alternative methods of sample size calculations for time-to-event endpoints into widespread usage has not been assessed to date. Reviews of adequacy of the event rate parameters used in sample size calculations compared to that observed in the trial have found that event rates were often underestimated or poorly estimated with large discrepancy between anticipated and observed event rates [31, 33]. Even when knowledge about non-constant event rates is available, sample size calculations assuming constant, or at the most, piecewise constant event rates are generally applied [7]. There has been little research into the effect of non-constant event rates when nonproportionality of treatment effects would be anticipated.

1.4 Analysis and reporting characteristics of trials

The work and publications of the Consolidated Standards of Reporting Trials (CONSORT) group have encouraged the adoption of guidelines to reporting trials and other research designs [34–36]. Previous reviews of trials involving time-to-event primary outcomes in the past twenty years have commonly assessed the adequacy and completeness of the reporting [3, 37–40], with review specific focus on the presentation of the survival plots [38], the completeness of the endpoint reporting [39], the implications of using summary statistics for inclusion into meta-analysis [40], or assessing for nonproportionality [3]. Recently published re-examinations of oncology trials have highlighted how prevalent time-dependent treatment effects may be, and that the use of standard analytical approaches assuming time-fixed treatment effects may underestimate the magnitude of, or miss completely, treatment effects that provide substantial survival benefits [41, 42].

Guidelines for presenting trial results graphically have been a priority for regulatory bodies. Kaplan-Meier plots are the predominant means in which to display the results of time-to-event outcomes [43] in the absence of competing events. They provide information about the survival experience of the groups presented, and a visual indication of the difference between the survival probabilities and quantiles of survival time over time. However, Kaplan-Meier plots do not provide direct information about measures of treatment effect despite such measures usually being the key focus of a clinical trial. Because the information to detect survival curve differences comes from the number of events occurring in each group relative to the number of participants available, trying to infer the strength of treatment effect differences from survival curves can be difficult and caution has been advised [44]. There can be a disconnect between the visual impression of when survival curves differ and the evidence for statistical assessment of difference. Some measure of treatment effect such as a logrank statistic or a HR estimated under the assumption of PH normally accompanies a Kaplan-Meier plot. Since both the logrank test and the HR are maximally powerful under PH, ideally assessment for any nonproportionality that may be present should be conducted and reported alongside a Kaplan-Meier plots.

1.5 Research aims and objectives

Nonproportionality of hazards is increasingly being observed and is a pressing issue that should not be ignored in the design or analysis phases of a trial. If nonproportionality is anticipated then the sample size and pre-specified statistical analysis plan should take this into account.

A constant event rate is another simplifying assumption commonly employed at the design phase of trials. If the assumption of proportional hazards holds, then the timing of event occurrences during a trial has no effect on the magnitude of the treatment effect and hence the power for a given number of events. However, in the presence of nonproportionality, the underlying event rate can be anticipated to impact on the performance characteristics of the treatment effect measures such as the magnitude of the treatment effect and the power and coverage compared to that anticipated at the design stage. To date, little attention has been paid to the interplay of non-proportionality and non-constant event rates.

We planned to review which regression-based methods were currently utilised to analyse time-to-event outcomes in clinical trials and whether there was allowance for anticipated nonproportionality. When analysis methods assuming - implicitly or explicitly - proportional hazards were used, we reviewed the awareness of the importance of testing for proportional hazards and the adequacy of reporting of the test results. We documented use of regression-based approaches that allowed for time-dependent treatment effects or non-constant event rates and recorded when alternative estimands to the HR were used.

The overall aims of the research in this thesis were to

- conduct a review of all clinical trial reports with primary time-to-event outcomes in four major medical research journals during the first six months of 2017, documenting design, planned analysis and testing approaches to accommodate anticipated non-constant treatment effect or event rates when planning clinical trials, and assessing the adequacy of reporting against checklists based on CONSORT guidelines

- undertake a simulation study using identified analytical approaches from the review to investigate the impact of non-constant event rates and nonproportionality on detection of time-dependent treatment effects

- illustrate how the use of flexible modelling approaches and the use of alternative estimands can bring new insights to answer clinical research questions using selected outcomes from a long running clinical trial

- propose the presentation of a complementary plot of treatment effect to accompany Kaplan-Meier plots which visual assessment of the strength and pattern of any time-dependent treatment effect.

1.6 Outline of the thesis

This thesis includes four manuscripts, one of which has been published with the other three submitted to peer-review journals and either under review or in the editorial stages of being accepted for publication.

Chapter 2 presents the results of the review of current practice for all original reports from four high impact journals for the first six months of 2017, examining characteristics of the design, analysis and reporting of clinical trials with primary time-to-event outcomes. This review has been published in the peer-review journal BMC Medical Research Methodology.

Chapter 3 describes the methods and results of a simulation study investigating the impact of non-constant event rates on estimated treatment effect measures in the presence of non-proportionality. A revision of the paper is currently undergoing the editorial process for acceptance in the BMC Medical Research Methodology journal.

Chapter 4 presents the results of an examination of the evidence for time-dependent treatment effects in ASPREE, a long running community-based trial assessing the evidence for preventive effects of low dose daily aspirin in older people. The regression-based approaches used in this chapter provide more clinical insight on the data and allow for alternative estimands to be computed. Estimation of time-dependent treatment effects was informed by the results from the simulation study (Chapter 3) and a more complex modelling approach was vindicated for some of the endpoints. This paper has been submitted to Pharmaceutical Statistics and is currently under review.

Chapter 5 provides recommendations for improving the visual presentation of results from time-to-event outcomes by providing a plot of treatment effect that is complementary to the Kaplan-Meier plot. Published Kaplan-Meier plots from the trials in the earlier review (Chapter 2) were used to illustrate the utility of this proposal and feedback from clinicians (including co-authors of the paper presented in Chapter 4) helped improve visual presentation and refine recommendations. This paper has been submitted to the Trials journal and is under consideration.

Chapter 6 presents a summary of the thesis findings, discusses limitations of the work and presents suggestions for further research.

Chapter 2

2.1 Manuscript introduction: Are non-constant rates and non-proportional treatment effects accounted for in the design and analysis of randomised controlled trials? A review of current practice

This chapter presents the results of a review of trials with a time-to-event primary outcome published during the first half of the 2017 year in four high impact medical journals. At the time of undertaking this research, two previous reviews of time-to-event methodology had found that awareness and reporting of the PH assumption when using the Cox model had been lacking [37, 39] and one review highlighted the extent of nonproportionality in oncology clinical trials that was not being evaluated [3]. Previous reviews of reporting of sample size calculations for continuous, binary and time-to-event outcomes found that there were inadequacies in the assumptions reported and post hoc modifications of sample size parameters were frequent [31, 32].

The review presented in this chapter extended the previous research specifically for time-to-event outcomes with (i) the first assessment of the uptake, if any, of recently developed theoretical or empirical methods of sample size calculations allowing for non-constant event rates and/or nonproportionality, (ii) a more in depth recording of all modelling approaches planned or used, (iii) details of the methods for assessing departures from proportionality planned and implemented and reported when hazard ratios from the Cox PH model were used, and (iv) assessment of the graphical presentation methods used to present the trial findings. We also illustrated the potential of regulatory guidelines in conjunction with journal editorial boards to improve the quality of reporting of trials. This was demonstrated by the increased timeliness of trial registrations before and after the introduction of a policy requiring pre-trial public registration as a condition of publication of trial findings.

In the next section is presented a manuscript as published in the journal *BMC Medical Research Methodology* [45]. The supplementary materials for the paper consisting of a citation listing of the review trials and the dataset underpinning the findings of the review are provided in Appendices A and B of the thesis.

Jachno *et al. BMC Medical Research Methodology* (2019) 19:103
https://doi.org/10.1186/s12874-019-0749-1

BMC Medical Research Methodology

RESEARCH ARTICLE Open Access

Are non-constant rates and non-proportional treatment effects accounted for in the design and analysis of randomised controlled trials? A review of current practice

Kim Jachno[*], Stephane Heritier and Rory Wolfe

Abstract

Background: Most clinical trials with time-to-event primary outcomes are designed assuming constant event rates and proportional hazards over time. Non-constant event rates and non-proportional hazards are seen increasingly frequently in trials. The objectives of this review were firstly to identify whether non-constant event rates and time-dependent treatment effects were allowed for in sample size calculations of trials, and secondly to assess the methods used for the analysis and reporting of time-to-event outcomes including how researchers accounted for non-proportional treatment effects.

Methods: We reviewed all original reports published between January and June 2017 in four high impact medical journals for trials for which the primary outcome involved time-to-event analysis. We recorded the methods used to analyse and present the main outcomes of the trial and assessed the reporting of assumptions underlying these methods. The sample size calculation was reviewed to see if the effect of either non-constant hazard rates or anticipated non-proportionality of the treatment effect was allowed for during the trial design.

Results: From 446 original reports we identified 66 trials with a time-to-event primary outcome encompassing trial start dates from July 1995 to November 2014. The majority of these trials (73%) had sample size calculations that used standard formulae with a minority of trials (11%) using simulation for anticipated changing event rates and/or non-proportional hazards. Well-established analytical methods, Kaplan-Meier curves (98%), the log rank test (88%) and the Cox proportional hazards model (97%), were used almost exclusively for the main outcome. Parametric regression models were considered in 11% of the reports. Of the trials reporting inference from the Cox model, only 11% reported any results of testing the assumption of proportional hazards.

Conclusions: Our review confirmed that when designing trials with time-to-event primary outcomes, methodologies assuming constant event rates and proportional hazards were predominantly used despite potential efficiencies in sample size needed or power achieved using alternative methods. The Cox proportional hazards model was used almost exclusively to present inferential results, yet testing and reporting of the pivotal assumption underpinning this estimation method was lacking.

Keywords: Randomised controlled trial, Time-to-event outcome, Proportional hazards, Event rates, Trial reporting, Sample size calculation

* Correspondence: kim.jachno@monash.edu.au
School of Public Health and Preventive Medicine, Monash University, Level 4, 553 St. Kilda Road, Melbourne 3004, Australia

Background

Time-to-event analysis, or survival analysis, has become the most widely utilized analytical method in research articles in leading general medical journals over the past two decades [1]. These analytical methods compare the duration of time until an event of interest occurs between different intervention groups. Randomised controlled trials (RCTs) provide the highest level of evidence on which to base decisions regarding the use of health interventions in humans. The Cox proportional hazards (PH) model [2] has become ubiquitous as the primary method for assessing treatment effects in RCTs with time-to-event outcomes. Its usage is matched only by the log rank test and Kaplan-Meier curves. Despite the popularity of the Cox PH model to estimate treatment effects, consideration of the fundamental assumption of proportional hazards is not always considered and reported [3].

Over the past two decades, the work and publications of the Consolidated Standards of Reporting Trials (CONSORT) group have encouraged the adoption of guidelines to report RCTs and other research designs [4–6]. Concurrently, there has been a range of policies issued by funding bodies and medical research publishers to enhance the quality, accountability and transparency of clinical trial design and reporting [7, 8]. In September 2004, the International Committee of Medical Journal Editors (ICMJE) disseminated a policy that pre-registration in a public trials registry would be required as a condition of consideration for publication for any trial starting from July 2005 [8]. Partly as a result of these improvements in regulatory oversight, trials are generally larger, and treatment effects are being evaluated for longer [9, 10] and as a consequence non-proportional hazards are detected more frequently [11]. Additionally, trials investigating different therapy modalities, such as immunotherapy compared to chemotherapy, or surgical compared to nonsurgical approaches [12], and the increased use of composite endpoints could also be reasons to anticipate treatment effects that vary over time. The summary hazard ratio (HR) effect measure from the Cox PH model may be less than ideal for decision making when treatment effects change over time [13]. By assuming the effect of treatment is always in the same direction, the HR from the Cox model has the potential to over or underestimate the magnitude of the treatment effect at any given time. Of more concern, if the effect of treatment changes direction over time then the true efficacy of a treatment, or safety issues with the treatment may be missed entirely if a summary HR is relied on.

When designing trials with time-to-event outcomes, sample size formulae exist to inform the required number of events needed to compare two survival distributions with a target effect size and desired power. The number of participants needed to be recruited is then calculated using expected event rates (the hazard), length of recruitment and follow up stages, any loss to follow up, administrative censoring and other logistical considerations in order to observe the number of events required. The most widely used sample size calculation methods to determine the number of events needed are based on the non-parametric log rank test [14, 15] which is most powerful for detecting alternative hypotheses when the hazards are proportional but makes no assumption about the distribution of the baseline hazard function. Alternative methods are based on the difference between two exponential survival functions [16, 17] which assumes proportional hazards as well as the more restrictive assumption of a constant baseline hazard function. Almost equivalently, the sample size formula derived for the HR from a Cox model [18] assumes proportional hazards between the different arms of the trial, but does not make any assumptions about the shape of baseline hazard function. While the Cox model does not assume a constant baseline hazard function, the sample size calculations based upon it yield almost equivalent number of events required to calculations assuming exponential survival rates. However, the shape of the hazard will influence the times at which those events are observed, and hence this needs to be considered together with other logistical considerations such as censoring rates in order to ascertain how many participants need to be recruited to the trial.

In the past two decades, several sample size methods have been proposed that acknowledge that the assumptions of proportional hazards and constant event rates may be too restrictive. These have included incorporating Fleming-Harrington weights [19, 20], allowing for non-proportionality to be specified as a series of piecewise exponential ‘stages’ within a trial [21], or sample size calculations that address specific types of non-proportionality such as lag to effect [20]. Parametric modelling approaches that allow for non-constant event rates such as the Weibull distribution [22, 23] or the generalized gamma distribution [24] have also been proposed. Simulation strategies can be used to empirically determine the sample size required and this approach enables either or both of (i) event rates assumed to change over time and (ii) anticipated non-proportionality of the treatment effect [25]. However, simulation requires a higher degree of programming skill and prior specification of more parameters in order to arrive at a final sample size. The uptake in trial practice of these alternative theoretical or empirical methods of sample size calculation has not been assessed to date.

There are three main approaches to analyzing time-to-event data involving non-parametric, semi-parametric and parametric models. Non-parametric methods such as the Kaplan-Meier method [26], or the method of Nelson [27] and Aalen [28] account for censoring and other characteristics of time-to-event data without making assumptions about the distribution of the event times through the hazard function or how the covariates affect event occurrence. The semi-parametric Cox model makes no assumption about the shape of the hazard function but covariates are assumed to have a multiplicative effect on the hazard. Parametric modelling alternatives to the Cox model such as the exponential-, Weibull- and Gompertz-distributed models assume a specific form for the hazard function as well as making the PH assumption. Other parametric models such as accelerated failure time models utilizing the Weibull and log-logistic distributions, or more recently developed fully flexible spline-based approaches [29, 30] are alternatives to semi-parametric modelling which may enable more clinically useful measures of absolute, as well as relative risk and measures of treatment effect that can be presented as either risk-based (hazard) or time based measures such as the absolute difference in mean survival time due to treatment. Models with a fully specified hazard function also enable easier accounting for, and presentation of time-dependent effects [31].

Previous reviews of survival analysis methodology have found that awareness and reporting of the proportional hazards assumption when using the Cox model has been lacking [32, 33]. Current methods for assessing the validity of the PH assumption include visual assessments and analytical tests. Graphical methods to assess proportionality involve inspection of log-transformed cumulative hazard functions [34] or scaled Schoenfeld residuals [35] against log-transformed time to observe equal slopes or horizontal lines when the PH assumption holds. Scaled Schoenfeld residuals can also be used in an analytical test for trend of non-zero slope against time - the Grambsch and Therneau test [36]. Another analytical method for assessing departures from proportionality is to create an interaction of treatment and time and inspect the significance of that time-dependent covariate [2] when included in a Cox model. However, all of these methods for assessing non-proportionality have some limitations, lacking power to detect some non-linear trends, or involving subjectivity or a particular form of departure from the PH assumption in the process [37].

The aims of this review were to assess the methods currently utilized to (i) accommodate anticipated non-constant treatment effects or event rates during the design phase, and (ii) account for non-proportional treatment effects over time during the analysis phase of trials involving time-to-event outcomes. When Cox models were used, we aimed to document whether there was evidence of an awareness of the underlying PH assumption, along with the any planned or reported PH testing, in either the main trial report or supplementary documentation. With the increased emphasis on improving the adequacy of reporting of results from trials over the past two decades, we also examined whether guidelines or policies may have had an impact on trial conduct.

Methods

All original reports published between January and June 2017 in three high impact general medical journals, the *New England Journal of Medicine*, the *British Medical Journal* and *The Lancet*, and one high impact specialized oncology journal, the *Journal of Clinical Oncology*, were considered. Initial screening excluded reports that were not based on data obtained from RCTs such as case reports and cohort studies, genomic and exomic analyses, systematic reviews, special reports or meta-analyses. Secondary screening then excluded articles that were reports early in the pipeline of drug development primarily investigating safety, pharmacokinetics and pharmacodynamics (Phase I and II trials), and reports of RCT data that were follow up or secondary reports (Phase IV trials). Finally, Phase III RCTs where the primary outcome was not a time-to-event endpoint, and reports requiring specialized trial design and analysis methodologies such as cluster randomised trials, or those involving crossover designs were excluded (see Fig. 1).

For each included trial we (KJ) recorded methodological approaches to calculating the sample size, and the clarity and completeness of the reporting of the assumptions that underpinned the sample size calculation. We noted time-to-event methods used for analysis and presentation. For trials using the Cox PH model, we recorded whether the PH assumption was acknowledged and investigated, the test(s) used and whether results of these investigations were detailed anywhere in the main report, attached protocols or other supplementary information. Trial registration information was collected for all trials and the information from the appropriate registry was used in addition to dates provided in the report to determine nominated trial start and end dates for the primary outcome. The publication date used was the issue publication date.

Results

There were 446 original reports published in the four selected journals during the review period and 66 of these reports were trials with a primary time-to-event outcome (Fig. 1). A citation listing of the final 66 trials is provided as additional material (see Additional file 1). The dataset of the final categories determined for the statistical approaches used in the trials is also provided (see Additional file 2).

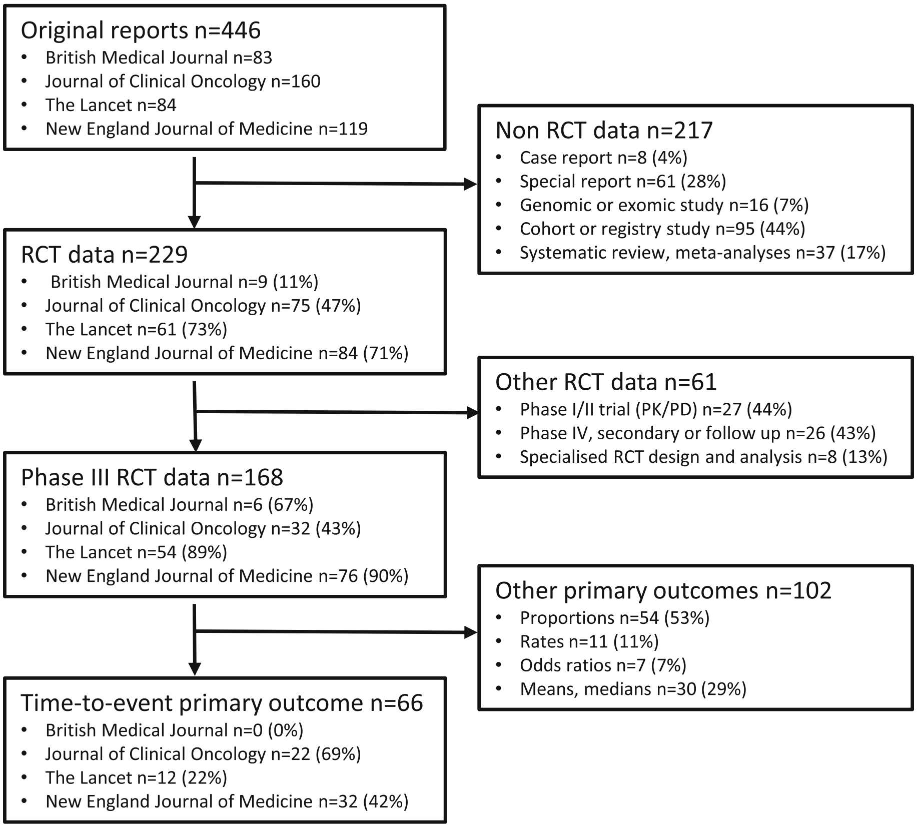

Original reports n=446
• British Medical Journal n=83
• Journal of Clinical Oncology n=160
• The Lancet n=84
• New England Journal of Medicine n=119
Non RCT data n=217
• Case report n=8 (4%)
• Special report n=61 (28%)
• Genomic or exomic study n=16 (7%)
• Cohort or registry study n=95 (44%)
• Systematic review, meta-analyses n=37 (17%)
RCT data n=229
• British Medical Journal n=9 (11%)
• Journal of Clinical Oncology n=75 (47%)
• The Lancet n=61 (73%)
• New England Journal of Medicine n=84 (71%)
Other RCT data n=61
• Phase I/II trial (PK/PD) n=27 (44%)
• Phase IV, secondary or follow up n=26 (43%)
• Specialised RCT design and analysis n=8 (13%)
Phase III RCT data n=168
• British Medical Journal n=6 (67%)
• Journal of Clinical Oncology n=32 (43%)
• The Lancet n=54 (89%)
• New England Journal of Medicine n=76 (90%)
Other primary outcomes n=102
• Proportions n=54 (53%)
• Rates n=11 (11%)
• Odds ratios n=7 (7%)
• Means, medians n=30 (29%)
Time-to-event primary outcome n=66
• British Medical Journal n=0 (0%)
• Journal of Clinical Oncology n=22 (69%)
• The Lancet n=12 (22%)
• New England Journal of Medicine n=32 (42%)

Fig. 1 Study design and primary focus of original reports included in this review. The boxes on the left side contain a listing of the classification of the 446 original reports divided into the numbers (n) from each of the four journals reviewed. Percentages in the subsequent boxes use the journal-specific number (n) from the previous box as the reference. The boxes on the right side are the different exclusion criteria applied to the original reports to obtain the final cohort of 66 Phase III RCTs with time to event primary outcomes reviewed. Percentages in each exclusion criteria box use the total number (n) of exclusions at that step as the reference

Description and summary findings of the statistical approaches used in trials

The statistical method characteristics of the trials in this review are summarized in Table 1. For the design phase of the trials, sample size approaches based on formulae involving a time-to-event outcome were categorized as either the log rank test, exponential survival distributions, the Cox PH model or simulation categories. Sample size approaches based on formulae involving a binary outcome at a pre-specified time point such as detecting a difference in proportions of event occurrence between the different arms of the trial were categorized as difference in proportion.

For the analysis phase of the trial, the time-to-event methods that were identified included the use of the non-parametric log rank test, the semiparametric Cox PH model, parametric regression models and landmark analysis approaches for providing multiple estimates of treatment effect. For trials where the Cox PH model was used, there was a further assessment of any acknowledgement of the underlying proportional hazards assumption, and details, if provided, about the method(s) planned to test the assumption.

Figure 2 presents a summary of our findings. Trial duration and time between trial completion and publication are represented by the lighter and darker horizontal bars respectively. The trials had start dates or registration dates in public databases stretching over a period of nearly two decades from July 1995 through November 2014, providing a means to assess if there have been any changes in trial design and reporting over that period. Trial registration timing relative to the start of recruitment is indicated by the triangles. Following the policy adopted by most major medical journal requiring trials to be prospectively registered, changes in timeliness of the trial registration process is evident. No trials which began prior to July 2005 had been registered prior to the nominated start date of the trial, with the clear majority of trials after July 2005 being registered prior to, or in a timely manner after, the nominated start date of the trial.

Table 1 Reported characteristics of the trials

Reported trial characteristic	*N* (%)
Sample size calculation approach	
Log rank test	40 (61%)
Cox model beta coefficient	4 (6%)
Exponentially distributed survival	4 (6%)
Simulation	7 (11%)
Difference in proportions	6 (9%)
Unclear	5 (6%)
Time-to-event analytical methods[a]	
Non-parametric log rank test	58 (88%)
Cox PH model	64 (97%)
Parametric regression	7 (11%)
Landmark analysis	7 (11%)
Proportional hazards (PH) assumption[b]	
PH assumption acknowledged	34 (53%)
PH testing methods documented	31 (48%)
Analytical test methods	10 (16%)
Visual assessment methods	6 (9%)
Visual and analytical methods	7 (11%)
Unspecified	8 (13%)

[a] Trials typically presented more than one analytical method
[b] for the 64 studies where Cox PH model used

There was no discernible pattern of change of trials reporting efficacy of primary outcome over time with the 38 (58%) RCTs reporting significant primary outcome findings being evenly spread throughout the two decades' starting time encapsulated within this review (Fig. 2, column E).

Designing trials - sample size calculations

There were 7/66 (11%) calculations based on simulation for predicted non-constant event rates over the course of the trial or to allow for an anticipated cure proportion or other non-proportional treatment effect in the trial. Methods that explicitly assume PH, or are maximally powerful under a PH assumption, were used in the majority ($n = 48/66$; 73%) of the sample size calculations. Among these, calculation based on the log rank test was most common ($n = 40/48$; 83%) noting that this utilizes ordered event times and is derived assuming a constant treatment effect over time. Other calculations were based on methods assuming PH for the treatment effect - either through assuming a difference between exponential survival distributions ($n = 4/48$; 8%) with the additional assumption of constant hazard functions, or the beta coefficient (HR) of a Cox model ($n = 4/48$; 8%) which does not make any assumptions about the shape of the baseline hazard function.

There were six trials which used a sample size calculation based on analysis of a difference in proportions of event occurrence in the different arms of the trial at a pre-specified fixed time. For three of these trials, this was justified by specified dual aims for the primary endpoint, (i) to show non-inferiority at a pre-specified time point using a difference in proportions, and (ii) to show superiority of the experimental treatment of interest using time-to-event methods. There were five reports where the basis for the sample size calculation was unclear.

Methods for the presentation and inference of results

For the graphical presentation of the primary outcome results, in 65/66 trials (98%) there was either a Kaplan-Meier survival plot or its reciprocal, a cumulative incidence plot. The Cox PH model was reported in 64/66 trials (97%) and the non-parametric log rank test was reported in 58/66 trials (88%; see Table 1). The dominance of the Cox PH model as a means to assess time-to-event outcomes, and in particular as the main inferential finding of the reports in this review is evident in Fig. 2 (columns U and I).

There were seven trials that planned to use parametric regression-based modelling approaches that could account for treatment effects changing over time (Table 1 and Fig. 2, column P/L). Six trials used parametric methods as well as the Cox PH method and one trial used parametric regression as the only inferential method. Regression approaches used were Weibull and flexible spline-based regression models that accounted explicitly for event rates being dependent on time, and exponential regression models using a dichotomous change point to allow for the effect of treatment to differ in two pre-specified stages. Seven trials out of 66 (11%) used the Cox model and also performed secondary 'landmark' analyses of the primary outcome presenting multiple estimates of the treatment effect for subsets of patients contingent on reaching intermediate event indicators, such as survival to one year or complete response in a biomarker assay.

Awareness of the PH assumption

About half of the reports (34/64; 53%) using the Cox model indicated an awareness of the importance of the PH assumption (Table 1 and Fig. 2, column A), and a similar proportion (31/64; 48%) included details of planned testing to check for any departures from proportionality in either the main report, attached supplementary information or any additional published protocols or statistical analysis plans referenced by the report. Analytical tests (17/64; 27%), either a time by

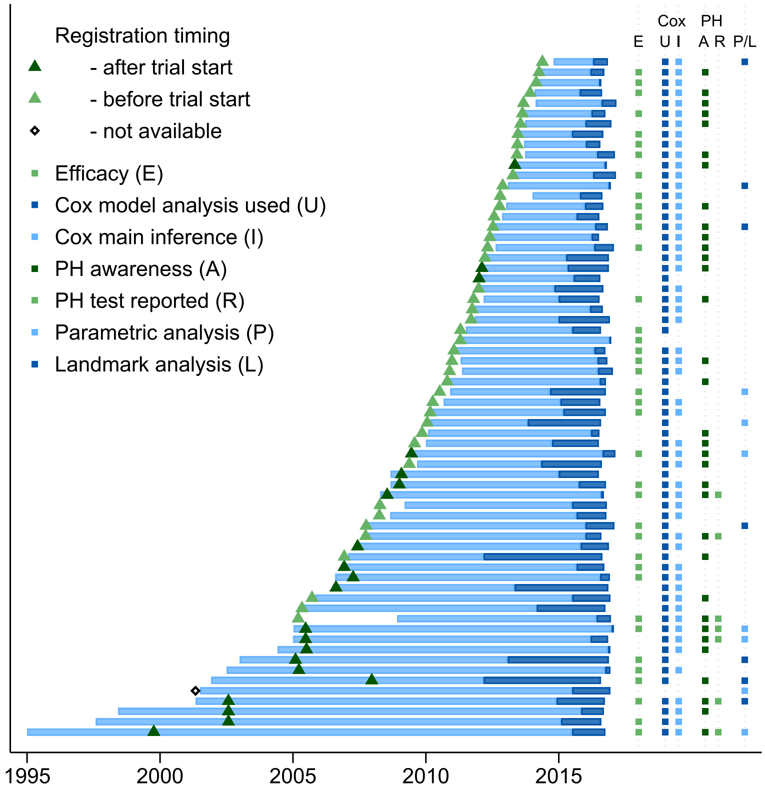


Fig. 2 Summary presentation of the findings of the review. Trial duration (years), between nominated start date and completion date, is indicated by the lighter shaded horizontal bars. Duration of time between completion and publication data is indicated by the darker shaded horizontal bars. Time of trial registration is shown by the triangles with lighter and darker shading indicating registration before and after nominated trial start date. Columns on the right side represent the determinations of trial characteristics for this review, including a trial reporting efficacy (E) of the primary outcome, the Cox PH model usage (U) in the report and presentation of the hazard ratio as the main inferential (I) finding. For trials using Cox analysis, the determinations of the awareness (A) and reporting (R) of the proportional hazards assumption for each trial is presented. Planned or presented usage of alternative regression models to the Cox PH model such as parametric or landmark (P/L) analysis is shown in the final column

treatment interaction in the Cox model or the Grambsch-Therneau test, were the most planned method of assessing for potential changing treatment effects over time, followed by visual means (13/64; 20%). Only seven reports (11%) explicitly presented the results of either visual or analytical tests of the assumption (Fig. 2, column R).

Influences on reporting assessment of the PH assumption

Comprehensive reporting of the PH assumption was more likely to occur when statistically significant results were being presented. Six of the seven trials reporting results of the PH testing also reported a statistically significant effect of treatment on the primary outcome. Of the 27 trials where there was an awareness but not reporting of the PH assumption, 22 trials (81%) used the Cox model as the main inferential finding with half of these presenting significant findings (Fig. 2, column I). In the 30 trials where there was no mention of the PH assumption, 24 trials (80%) presented the Cox model as the main inferential result, with 14 of these significant findings and 10 non-significant findings.

We expected that guidelines such as the CONSORT statement and improved regulatory oversight would have led to an increased consideration to plan and report investigations of the PH assumption over time. Unexpectedly, reporting of PH assumption test results was only seen in trials that commenced prior to June 2009. This might be explained by trials of longer planned duration having a greater awareness of the potential for time-dependent treatment effects to manifest, and hence be more likely to explicitly report results of tests of the PH assumption. However, it is of concern that there was no evidence of increased awareness and reporting of investigation of the PH assumption in trials initiated more recently, irrespective of the planned duration of the trial.

Discussion

This review assessed design and analysis of RCTs with time-to-event primary outcomes in an era in which non-constant event rates and non-proportional treatment effects are encountered more frequently. Our findings are now discussed alongside previous reviews of reporting of RCTs involving time-to-event primary outcomes and other relevant literature.

Sample size calculations – adequacy of reporting

Previous reviews have assessed the sample size calculations for a mix of continuous and binary as well as time-to-event outcomes [38, 39]. These reviews concluded that whilst reporting of sample size calculations has improved over time as a result of more stringent requirements imposed by journals and the provision of guidelines such as the CONSORT statement, there were still inadequacies in the assumptions reported and that post hoc modification of sample size parameters was frequent. In our review we too found that initial sample size calculations could have been more adequately reported: the number of participants in the trial was often adjusted for appropriate reasons such as interim analysis, important secondary analysis, or loss to follow up without clear demarcation between the number of events required using the sample size formula and the number of participants to be recruited. We found encouraging signs that researchers are beginning to anticipate the impacts of non-proportional hazards and changing event rates on sample size calculations evidenced by seven trials using simulation-based procedures for their determination of sample size. No trials in our review used more recently proposed modified sample size calculations to allow for anticipated cure proportions [40] or lag times until full treatment effect [20, 41] as could be anticipated in many of the immunotherapy-based treatments under assessment in oncology trials.

Modelling approaches – changes in recent years

Our review highlights a gradual change over recent decades in the modelling approaches used by general medical and oncology researchers to assess treatment effects on time-to-event outcomes. A review of survival analyses in four cancer journals published during 1991 [32], reported that the log rank test was used to assess treatment differences in 84/113 (74%) whereas only 4/113 (4%) trials used the Cox PH model. No parametric models were used to assess the treatment effect in that review. Over a decade later, another review of 274 trials in major cancer journals published during 2004 [33] found that the log rank test was used in 63% of studies with the Cox model being used in 51% of studies to report the treatment effect. Again, no parametric models were used. Similarly, a review of reports published in five oncology journals during 2015 found that the log rank test was used in 66% of studies with the Cox model being used in 88% of studies to report the treatment effect, and there was no reported use of parametric modelling approaches [42]. In our review, the log rank test was used in 88% of studies, the Cox model in 97% of studies, and parametric modelling approaches were proposed or used in 11% of trials. We also noted that additional landmark analysis was used in 11% of the trials, indicating recognition by the authors that one summary measure of treatment effect did not fully describe the trial findings.

Assessing for treatment effects that are over time-dependent

Despite the widespread use of the Cox proportional hazards model in medical research, awareness and testing for non-proportionality has not yet become systematic. In the 1995 review of four cancer journals, only 2 (5%) of 43 papers which used the Cox model mentioned that the PH assumption was verified whilst in 2004, one of 64 (2%) usages of a Cox model reported verifying the PH assumption [32, 33]. More recently, a review of trials from five journals published during 2014 [3] found that there was evidence of non-proportionality in 13/54 trials (24%) determined by digitally recreating the individual patient data from the published Kaplan–Meier curves; however, there was no indication of the number of trials in which the PH assumption was assessed in the original reports for that review. A review of survival analysis reporting in the same or similar journals [42] published in 2015 found that only 2/32 (7%) trials using the Cox PH model reported testing for the PH assumption. Our review found the highest reporting rate of 7/64 (11%) which suggests that guidelines to improve the reporting of results may be having an effect but there is still considerable room for improvement.

Success of guidelines and policies for improving the quality of reporting

The success of journal guidelines and requirements for improving the quality of the reporting of trials is evident in the change in timeliness of trial registrations in our review. The four reviewed journals are either members of the ICMJE or adopted the July 2004 policy requiring pre-trial public registration as a condition of publication for trials commencing from July 2005 with trials beginning prior to that date able to register under an exemption clause by September 2005. No trials which began prior to July 2005 had been registered prior to the nominated start date of the trial, whereas the clear majority of trials after July 2005 had been registered prior to, or shortly after the nominated start date of the trial (Fig. 2). This success stands in contrast to the assessment and reporting of the PH assumption in Cox models, resulting in renewed calls made by others [43], and echoed here by us, for the reviewers, journal editors, regulators and funders of research to demand enhanced content in reports and associated supplementary documentation in order to improve trial reproducibility and interpretation.

Conclusions

In this review, we explored whether researchers account for non-constant event rates and non-proportional treatment effects during the design, analysis and reporting phases of randomised trials. The insights we derive are timely as health research has entered an era in which trials are being conducted for longer durations and are often adequately powered to evaluate the durability of treatment effects over time. Longer trials make the PH assumption increasingly unrealistic over the entire study duration. In addition, treatment effects that change over time are more likely to be encountered in trials due to the increased use of composite endpoints, and due to the nature of interventions that are now employed in late stage oncology trials. The journals included in this review were all high impact journals that have emphasized the CONSORT guidelines as part of their submission requirements yet the quality of the reporting over the past two decades has been consistently less than optimal. These major medical journals have rigorous statistical review policies and require protocols and other supplementary documents to accompany their original reports of RCTs. This enhanced comprehensiveness of reporting gives investigators adequate scope for completeness and precision in the reporting of trial results.

Additional files

Additional file 1: Listing of the sixty-six randomised clinical trials in this review. A citation listing by journal. (DOCX 27 kb)

Additional file 2: Determination of the characteristics of the sixty-six randomised clinical trials in this review. Dataset containing the final determinations of trial characteristics. (XLS 63 kb)

Abbreviations

CONSORT: Consolidated Standards of Reporting Trials; HR: Hazard ratio; ICMJE: International Committee of Medical Journal Editors; PH: Proportional hazards; RCT: Randomised controlled trial

Acknowledgements

Not applicable.

Funding

KJ was supported in part by an Australian Government Research Training Program (RTP) Stipend and RTP Fee-Offset Scholarship through Federation University Australia and a National Health and Medical Research Council of Australia grant (APP1128222). The funding bodies had no role in the design of the study, the collection, analysis and interpretation of data or in the writing of the manuscript.

Availability of data and materials

The trials that were reviewed for this publication (Additional file 1) and the dataset supporting the conclusions of this article (Additional file 2) are included in the supplementary information.

Authors' contributions

KJ extracted and reviewed the reports, and drafted the manuscript. RW conceived the review, resolved any uncertainty encountered by KJ and helped with the drafting of the manuscript. SH revised draft versions of the manuscript. All authors read and approved the final manuscript.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 7 December 2018 Accepted: 7 May 2019
Published online: 16 May 2019

References

1. Sato Y, Gosho M, Nagashima K, Takahashi S, Ware JH, Laird NM. Statistical methods in the journal — an update. N Engl J Med. 2017;376(11):1086–7.
2. Cox DR. Regression models and life-tables. J R Stat Soc Ser B Methodol. 1972;34(2):187–220.
3. Trinquart L, Jacot J, Conner SC, Porcher R. Comparison of treatment effects measured by the Hazard ratio and by the ratio of restricted mean survival times in oncology randomized controlled trials. J Clin Oncol. 2016;34(15): 1813–9.
4. Moher D, Schulz KF, Altman D, Group ftC. The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomized trials. JAMA. 2001;285(15):1987–91.
5. Moher D, Hopewell S, Schulz KF, Montori V, Gøtzsche PC, Devereaux PJ, et al. CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. BMJ. 2010;340.
6. Begg C, Cho M, 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.
7. Bhatt A. Quality of clinical trials: a moving target. Perspect Clin Res. 2011; 2(4):124–8.
8. De Angelis C, Drazen JM, Frizelle FA, Haug C, Hoey J, Horton R, et al. Clinical trial registration: a statement from the international committee of Medical Journal Editors. CMAJ : Can Med Assoc J. 2004;171(6):606–7.
9. International conference on harmonisation of technical requirements for Pharmaceuticals for Human use. *ICH Harmonised Tripartite Guidelines: Statistical Principles for Clinical Trials E9*. London, England: European Medicines Agency 1998.
10. Booth CM, Cescon DW, Wang L, Tannock IF, Krzyzanowska MK. Evolution of the randomized controlled trial in oncology over three decades. J Clin Oncol. 2008;26(33):5458–64.
11. Royston P, Parmar MKB. An approach to trial design and analysis in the era of non-proportional hazards of the treatment effect. Trials. 2014;15:314.
12. Howard G, Chambless LE, Kronmal RA. Assessing differences in clinical trials comparing surgical vs nonsurgical therapy: using common (statistical) sense. JAMA. 1997;278(17):1432–6.
13. Hernán MA. The hazards of Hazard ratios. Epidemiology (Cambridge, Mass). 2010;21(1):13–5.
14. Schoenfeld DA, Richter JR. Nomograms for calculating the number of patients needed for a clinical trial with survival as an endpoint. Biometrics. 1982;38(1):163–70.
15. Freedman LS. Tables of the number of patients required in clinical trials using the logrank test. Stat Med. 1982;1(2):121–9.
16. Lachin JM. Introduction to sample size determination and power analysis for clinical trials. Control Clin Trials. 1981;2(2):93–113.
17. Lachin JM, Foulkes MA. Evaluation of sample size and power for analyses of survival with allowance for nonuniform patient entry, losses to follow-up, noncompliance, and stratification. Biometrics. 1986;42(3):507–19.

18. Hsieh FY, Lavori PW. Sample-size calculations for the Cox proportional hazards regression model with nonbinary covariates. Control Clin Trials. 2000;21(6):552–60.
19. Hasegawa T. Sample size determination for the weighted log-rank test with the Fleming–Harrington class of weights in cancer vaccine studies. Pharm Stat. 2014;13(2):128–35.
20. Sit T, Liu M, Shnaidman M, Ying Z. Design and analysis of clinical trials in the presence of delayed treatment effect. Stat Med. 2016;35(11):1774–9.
21. Barthel FMS, Babiker A, Royston P, Parmar MK. Evaluation of sample size and power for multi-arm survival trials allowing for non-uniform accrual, non-proportional hazards, loss to follow-up and cross-over. Stat Med. 2006;25(15):2521–42.
22. Heo M, Faith MS, Allison DB. Power and sample size for survival analysis under the Weibull distribution when the whole lifespan is of interest. Mech Ageing Dev. 1998;102(1):45–53.
23. Wu J. Power and sample size for randomized phase III survival trials under the Weibull model. J Biopharm Stat. 2015;25(1):16–28.
24. Phadnis MA, Wetmore JB, Mayo MS. A clinical trial design using the concept of proportional time using the generalized gamma ratio distribution. Stat Med. 2017;36:4121–40.
25. Hooper R. Versatile sample-size calculation using simulation. Stata J. 2013;13(1):21–38.
26. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. J Am Stat Assoc. 1958;53(282):457–81.
27. Nelson W. Theory and applications of Hazard plotting for censored failure data. Technometrics. 1972;14(4):945–66.
28. Aalen O. Nonparametric inference for a family of counting processes. Ann Stat. 1978;6(4):701–26.
29. Royston P, Parmar MKB. Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. Stat Med. 2002;21(15):2175–97.
30. Crowther MJ, Lambert PC. Simulating biologically plausible complex survival data. Stat Med. 2013;32(23):4118–34.
31. Royston P, Lambert PC. Flexible parametric survival analysis using Stata: beyond the Cox model: Stata Press; 2011.
32. Altman DG, De Stavola BL, Love SB, Stepniewska KA. Review of survival analyses published in cancer journals. Br J Cancer. 1995;72(2):511–8.
33. Mathoulin-Pelissier S, Gourgou-Bourgade S, Bonnetain F, Kramar A. Survival end point reporting in randomized Cancer clinical trials: a review of major journals. J Clin Oncol. 2008;26(22):3721–6.
34. Kalbfleisch JD, Prentice RL. The statistical analysis of failure time data. New York: Wiley; 1980.
35. Schoenfeld D. Partial residuals for the proportional hazards regression model. Biometrika. 1982;69(1):239–41.
36. Grambsch PM, Therneau TM. Proportional hazards tests and diagnostics based on weighted residuals. Biometrika. 1994;81(3):515–26.
37. Austin PC. Statistical power to detect violation of the proportional hazards assumption when using the Cox regression model. J Stat Comput Simul. 2018;88(3):533–52.
38. Bariani GM, de Celis Ferrari ACR, Precivale M, Arai R, Saad ED, Riechelmann RP. Sample size calculation in oncology trials: quality of reporting and implications for clinical Cancer research. Am J Clin Oncol. 2015;38(6):570.
39. Charles P, Giraudeau B, Dechartres A, Baron G, Ravaud P. Reporting of sample size calculation in randomised controlled trials: review. BMJ. 2009;338:b1732.
40. Wu J. Sample size calculation for testing differences between cure rates with the optimal log-rank test. J Biopharm Stat. 2017;27(1):124–34.
41. Zhang D, Quan H. Power and sample size calculation for log-rank test with a time lag in treatment effect. Stat Med. 2009;28(5):864–79.
42. 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.
43. Gamble C, Krishan A, Stocken D, Lewis S, Juszczak E, Caroline D, et al. Guidelines for the content of statistical analysis plans in clinical trials. JAMA. 2017;318(23):2337–43.

Chapter 3

3.1 Introduction: Impact of a non-constant baseline hazard on detection of time-dependent treatment effects: a simulation study

As described in Chapter 2, the majority of trials with time-to-event outcomes use analytical approaches that are maximally powerful under an assumption of proportional hazards implying a time-independent or ‘fixed’ magnitude treatment effect is the estimand of interest. The sample size calculation for a time-to-event outcome determines first the number of *events* required to be observed in order to detect a pre-specified treatment effect with a nominated power and significance level. An additional assumption of constant event rates - constant baseline hazards - is then typically applied to determine the number of *participants* that need to be recruited given logistical considerations of total trial duration and anticipated accrual and withdrawal rates. This chapter presents the results of a simulation study which investigated the interplay of relaxation of these two assumptions of constant event rates and proportional hazards and explored the implications for clinical trial design.

Oncology trials exhibiting time-dependent treatment effects due to the advent of immunotherapy-based drug regimens provided the motivation for the two forms of nonproportionality assessed in the simulation study - a time lag until treatment becomes effective and an early effect of treatment that ceases. The impact of clinically plausible non-constant event rates was evaluated both when there was no time-dependent treatment effect ie proportional hazards, and when time-dependent treatment effects were present. The power of commonly utilised regression-based measures of treatment effect and tests of survival curve difference were compared. The suitability of three measures of treatment effect - the hazard ratio, the difference in restricted mean survival time and the time ratio - were evaluated in terms of the magnitude of treatment effect and coverage properties relative to the values stipulated at the design phase.

In the next section is presented a manuscript which has been accepted pending final editorial revisions by the *BMC Medical Research Methodology* journal. Supplementary methods and results for the manuscript are available in Appendix C, and example Stata code to create and analyse the simulated datasets on which the findings of the manuscript are based can be found in Appendix D.

RESEARCH ARTICLE

IMPACT OF A NON-CONSTANT BASELINE HAZARD ON DETECTION OF TIME-DEPENDENT TREATMENT EFFECTS: A SIMULATION STUDY

Kim Jachno*[1], Stephane Heritier[1], Rory Wolfe[1]

[1]School of Public Health and Preventive Medicine, Monash University, Melbourne, Victoria, Australia.

*Correspondence to Kim Jachno

Abstract

Background: Non-proportional hazards are common with time-to-event data but the majority of randomised clinical trials (RCTs) are designed and analysed using approaches which assume the treatment effect follows proportional hazards (PH). Recent advances in oncology treatments have identified two forms of non-PH of particular importance - a time lag until treatment becomes effective, and an early effect of treatment that ceases after a period of time. In sample size calculations for treatment effects on time-to-event outcomes where information is based on the number of events rather than the number of participants, there is crucial importance in correct specification of the baseline hazard rate amongst other considerations. Under PH, the shape of the baseline hazard has no effect on the resultant power and magnitude of treatment effects using standard analytical approaches. However, in a non-PH context the appropriateness of analytical approaches can depend on the shape of the underlying hazard.

Methods: A simulation study was undertaken to assess the impact of clinically plausible non-constant baseline hazard rates on the power, magnitude and coverage of commonly utilized regression-based measures of treatment effect and tests of survival curve difference for these two forms of non-PH used in RCTs with time-to-event outcomes.

Results: In the presence of even mild departures from PH, the power, average treatment effect size and coverage were adversely affected. Depending on the nature of the nonproportionality, non-constant event rates could further exacerbate or somewhat ameliorate the losses in power, treatment effect magnitude and coverage observed. No single summary measure of treatment effect was able to adequately describe the full extent of a potentially time-limited treatment benefit whilst maintaining power at nominal levels.

Conclusions: Our results show the increased importance of considering plausible potentially non-constant event rates when non-proportionality of treatment effects could be anticipated. In planning clinical trials with the potential for non-PH, even modest departures from an assumed constant baseline hazard could appreciably impact the power to detect treatment effects depending on the nature of the non-PH. Comprehensive analysis plans may be required to accommodate the description of time-dependent treatment effects.

Keywords: non-proportionality; non-constant hazards; flexible parametric models; weighted logrank tests; restricted mean survival time

Background

Randomised clinical trials (RCTs) have an overarching objective to understand if a new treatment is effective compared to existing treatments. RCTs with time-to-event outcomes can examine when, and for how long, the treatment exhibits an effect. Nevertheless, the vast majority of RCTs with time-to-event outcomes are analysed using methods that are maximally powerful under an assumption of proportional hazards, implying time-independent or ‘fixed’ magnitude treatment effects. The main analytical approaches currently reported in major medical journals can be broadly categorised as tests of equal survival functions which provide a p-value for inference only, or modelling approaches which provide an estimate of treatment effect along with a p-value for inference (1-5). When designing trials, as well as the assumption of time-independent treatment effects, there is often an explicit or implicit assumption of constant event rates – constant baseline hazards - used to determine the number of events required and hence the number of patients that need to be recruited for the trial to have the desired power in the sample size calculations methods employed (4, 6).

Paradigm shifts in oncology treatments over the past two decades provides motivation for assessing the effect of non-proportionality on analytical methods for time-to-event outcomes (7). Two broad classes of time-dependent treatment effects, early effect that attenuates and lag to effect, have emerged as there has been a shift to biomolecular-targeted and immunotherapy-based treatments implemented either alone or as an adjunct to surgical and chemotherapy-based approaches. Many of the first wave of biomolecular-based anticancer agents were observed to improve patient survival initially but have limited long-term survival benefit due to acquired biological resistance to, or accumulated toxicities from the treatment. This is an example of an early treatment effectiveness which attenuates or becomes harmful over time. A subsequent wave of immunotherapy-based treatments act to stimulate the patient's own immune system to kill cancerous cells. This circumvents the problems observed with toxicity and resistance to the biological-based agents. However, this mechanism of action via immune system activation is typically associated with a delay of varying months' duration until any treatment effect may be observed, an example of a lag until treatment effectiveness. Recent reappraisals using reconstructed data of published phase III oncology trials have highlighted how prevalent time-dependent treatment effects may be, and that the use of standard analytical approaches assuming time-fixed treatment effects may underestimate the magnitude of, or miss completely. treatment effects that provide substantial survival benefits (5, 8)

The two most popular analysis approaches for comparing survival curves in different treatment groups are the logrank (LR) test used to evaluate the null hypothesis of identical survival functions, and the Cox PH model to obtain an estimate of the treatment effect as a summary hazard ratio (HR). Under PH, these two approaches are known to be maximally powerful and provide an asymptotically equivalent test of significance. When non-proportionality exists, the LR test can lose power to detect survival curve differences with the magnitude of the loss dependent on the configuration of the non-proportionality. Extensions to the LR test have been proposed which maintain power under different anticipated scenarios of non-proportionality. These include the Fleming-Harrington (FH) family of weighted LR test statistics which can be differentially weighted to emphasise events that occur earlier, in the middle, or later over the survival time horizon of interest (9). Other weighting approaches exist that use more flexible data-driven procedures to specify weight functions that maintain power, such as Yang and Prentice's adaptive model (10) or Magirr and Burman's modestly weighted LR test for delayed-onset non-proportionality (11). Weighted LR tests can be criticised because they treat some events as more important than others and that there is not necessarily an accompanying estimate of treatment effect available for clinical interpretation. An alternative approach to testing for a generalised treatment effect is to use the combined results of multiple significance tests appropriately standardised to maintain the null distribution. Examples of these combined tests include using the minimum of the Cox PH model p-value and a permutation test based on the restricted mean survival time (12) or selecting the minimum of the three p-values from the FH family weighted LR tests under equal, early effect and lag to effect weighting scenarios (13).

When the assumption of proportionality of the treatment effect is met, the summary HR from a Cox PH model is a suitable parameter to provide a clinically meaningful measure of the relative difference between two survival curves. When not met, the clinical interpretation of a single summary measure such as the HR is not clear. When the underlying HR varies over time, assuming that there are a series of periods in which the PH assumption holds, then the magnitude of the summary HR can be interpreted as a weighted average of the sum of the proportion of events and estimated HR in each of the periods. These weights depend on the event rates, accrual distribution and the dropout pattern, and these dependencies could result in different parameter estimates in different trials, even with identical survival curves, thus removing the integrity of the summary HR as a meaningful measure of overall treatment effect.

An alternative estimand of treatment effect for time-to-event outcomes that does not rely on the PH assumption is the restricted mean survival time (RMST) (14). The RMST is the mean duration of survival for the trial population up to a given time point (often designated t^*). Recent research on the use of the RMST to estimate treatment effects as an adjunct estimand to the HR has shown agreement in terms of statistical significance of the treatment effect under PH (14-16). Since the choice of estimand and analytical method needs to be pre-specified in a clinical trial, to avoid any bias from selective reporting, a summary HR from a Cox model is often stipulated as the primary analysis because at that point in time there may be an absence of meaningful data from which to justify the treatment effect as a time-varying quantity. However, it has been recommended that the difference in RMST, or the ratio of RMST, be reported complementary to, or as the primary outcome measure in trials whether or not non-proportionality of the treatment effect could be anticipated (17, 18). As well as not relying on a PH assumption, the RMST also has desirable properties for (i) interpretability in that it can be expressed in both relative and absolute measures and the chosen metric is time, not risk, and (ii) performance since it is a summary measure that captures the temporal profile of all events up to the cut off time t^*.

When conducting clinical trials, in order for a single test of RMST difference to be valid, the selected time point of interest t^* must be pre-specified at the design stage. Choices of t^* relatively late in the follow up confer power similar to that observed with the Cox PH model. Depending on the patterns of non-PH, other choices of t^* may considerably increase the power to detect a difference. Royston and Parmar have also developed a generalised test of treatment effect, which tests the RMST difference at several prespecified values of t^* during the follow-up, taking the smallest p-value as the basis for the test after adjusting for multiple testing (12). By combining this p-value and the p-value from the Cox PH model, an overall p-value for the combined test (designated pCT) can be derived and has the correct distribution under the null hypothesis of equal survival curves.

Accelerated failure time (AFT) models (19-21) also model the treatment effect on a time-based rather than a hazard-based metric, enabling potentially more intuitive clinical understanding. These models include a survival model based on the Weibull distribution which has both PH and AFT interpretations depending on the parameterisation selected, thus acting as a conduit model for investigating treatment effects in both risk-based and time-based metrics.

A further consideration, as yet unexamined in the comparisons of the performance of analysis methods, is the shape of the hazard in the baseline treatment group. Reviews of adequacy of the event rate parameters used in sample size calculations compared to that observed in the trial have found that event rates were often underestimated (22) or that there were large discrepancies between the assumed parameters and the estimated ones from observed data (23). Sample size calculations assuming constant, or at the most, piecewise constant event rates were applied even when prior information on the shape of the underlying event rate was available (6).

The Cox model makes no assumption about this shape whereas parametric modelling approaches, including fractional polynomials (24) or splines (25) model the underlying shape of the baseline hazard function. If the PH assumption holds, the time when the events occur does not influence the magnitude, coverage, power or type I error rate of the HR estimate. However, in the presence of a time-dependent effect of treatment, the summary HR provides an 'average' effect with the averaging being weighted by the number of events and the timing of their occurrence. While it is reasonably intuitive (14) to infer that the shape of the hazard function in the control group will impact on the extent to which a HR from a Cox PH model is a misleading summary of time-dependent effects of treatment, there is limited work that has quantified this phenomenon nor explored general properties of the Cox PH model HR estimand when the model is mis-specified in this way. The properties of other analytical approaches that estimate effects of treatment have also not been examined in this context.

This paper evaluates the impact of a non-constant event rate on the suitability of three measures of treatment effect - the HR, the difference in RMST (*Δ*RMST), and an acceleration factor expressed as a time ratio (TR) under scenarios where PH do not hold. Suitability of the treatment effect estimates will be assessed in terms of their estimated magnitude, coverage and power benchmarked to that assumed at the design phase of the trial. The properties of three modelling approaches will be examined, the semiparametric Cox PH model, the Royston-Parmar (RP) models utilising flexible restricted cubic splines and parametric models assuming the exponential or Weibull distributions. A landmark (LM) approach to the parametric modelling that allow for multiple estimates of time period-specific or conditional treatment effects will also be undertaken. Additionally, the impact of non-constant event rates on the power of commonly pre-specified analytical approaches that provide a test of equal survival curve significance but not an estimate of treatment effect will be assessed. These approaches include using the p-values obtained from the Cox PH model, the LR test, weighted LR tests and omnibus extensions to the weighted LR test and the combination test based on the RMST.

The structure of the article is as follows. In the Methods section we describe the aims of the simulation study, the data-generating models used for the different non-PH scenarios, the estimands of treatment effect and tests of equal survival functions to be compared and the measures used to assess the performance of the analysis methods. In the Results section, we report the results of the findings of the simulations. We end with a Discussion and some recommendations and conclusions.

Methods

We aimed to assess the effect of non-constant event rates on the suitability of the estimates from three measures of treatment effect, the HR, the time ratio (TR) and the *Δ*RMST, and on the performance of tests of equal survival function under PH and two non-PH scenarios. Our motivation came from phase II and III clinical trials of immunotherapies for late stage cancers (5, 8). In the absence of treatment, most participants were likely to experience the event of interest within the study's proposed follow-up time of 50 months. We based the simulation on a generic two-group trial to detect a 33% reduction in the hazard rate underlying progression-free survival with 80% power and a significance level 0.05. Assuming a constant – or equivalently proportional - event rate and PH, a sample size calculation based on the LR test with HR=0.67, (log(HR)=-0.4) would require 202 events to be observed (26). Characteristics of the Design model used in the simulations are detailed in Table 1, along with the Data-Generating models (DGMs) for the simulation and Analysis models that could be chosen for pre-specification in a trial protocol.

Table 1: Characteristics of the Design model, the Data-Generating models and the Analysis models

Design Model:	Weibull baseline hazard (constant event rate), proportional hazards (PH), treatment effect HR=0.67, maximum time $t = 50$ $h(t) = \lambda\gamma t^{\gamma-1}\exp(\beta X_{TRT})$ where $\lambda = 0.10$, $\gamma = 1.0$, $\beta = -0.4$ and $X_{TRT} = 0,1$ for control and treatment groups		
Data Generating Models (DGMs):	Weibull baseline hazard (decreasing, constant and increasing event rates), non-proportional hazards		
	Event rate scenario	Baseline hazard values	Non-proportional hazard change times
	Lag until effect, HR=1 if $t \le t_{lag}$, HR=0.67 if $t > t_{lag}$; $h(t) = \lambda\gamma t^{\gamma-1}\exp(\beta X_{TRT} \times I(t > t_{lag}))$		
	Decreasing	$\lambda_d = 0.15$, $\gamma_d = 0.9$	
	Constant	$\lambda_c = 0.10$, $\gamma_c = 1.0$	$t_{lag} = 0, 1, 3$ or 10; $t_{lag} = 0$ are PH DGMs
	Increasing	$\lambda_i = 0.07$, $\gamma_i = 1.1$	
	Early effect ceasing, HR=0.67 if $t \le t_{early}$, HR=1 if $t > t_{early}$; $h(t) = \lambda\gamma t^{\gamma-1}\exp(\beta X_{TRT} \times I(t \le t_{early}))$		
	Decreasing	$\lambda_d = 0.15$, $\gamma_d = 0.9$	
	Constant	$\lambda_c = 0.10$, $\gamma_c = 1.0$	$t_{early} = 3,10,20,50$; $t_{early} = 50$ are PH DGMs
	Increasing	$\lambda_i = 0.07$, $\gamma_i = 1.1$	
Analysis Models:	Cox PH (Cox)	$h_i(t) = h_0(t)\exp(\beta X_{TRT})$	Average HR from all events in t
	Landmark (LM)	$h_i(t) = h_0(t)\exp(\beta X_{TRT} \times I(t > t_{LM}))$	Average HR from events *after* t_{LM}[1]
	Piecewise exponential (PE1)	$h_i(t) = \lambda_j \exp(\beta X_{TRT})$	Average HR from all events in t
	Piecewise exponential (PE2)	$h_i(t) = \lambda_j \exp(\beta X_{TRT} \times I(t > t_{PE}))$	Average HR from events *after* t_{PE}[2]
	Royston Parmar PH (RP(PH))	$\ln(H_i(t)) = s(\ln(t)\|\boldsymbol{\gamma}_s, \mathbf{k}_0) + \beta X_{TRT}$	Average HR from all events in t ΔRMST from all events in t
	RP time-dependent (RP(TD))	$\ln(H_i(t)) = s(\ln(t)\|\boldsymbol{\gamma}_s, \mathbf{k}_0) + s(\ln(t))X_{TRT} + \beta X_{TRT}$	ΔRMST from all events in t
	Accelerated Failure Time (AFT)	$\ln(t_i) = \beta X_{TRT} + \varepsilon_i$	Average TR from all events in t

1. Pre-specified $t_{LM} = 3$ for lag until effect non-PH, $t_{LM} = 10$ for early effect ceasing non-PH
2. Pre-specified $t_{PE} = 3$ for lag until effect non-PH, not reported for early effect ceasing non-PH

Data-generating processes for simulation scenarios

Using a Weibull data-generation model, three different event rate scenarios were considered by selecting a scale parameter λ and a shape parameter γ such that there was a near zero probability of survival by the end of an administratively imposed time in each scenario. For the constant event rate scenario, we determined the value for the scale factor (λ_c) that would result in less than 0.7% chance of survival in the absence of treatment effect under a constant event rate ($\gamma_c = 1$; ie the exponential distribution) within the specified trial time frame ($t = 50$ months). In the second and third scenarios, clinically plausible values of the shape parameter were selected to provide modest decreasing ($\gamma_d = 0.9$) and increasing ($\gamma_i = 1.1$) event rate scenarios. For these latter scenarios, we determined the scale parameter that would result in the same survival probability by the end of follow up ($t = 50$), and hence observation of the same number of events in the absence of treatment, as under the constant event rate (see Table 1). This enabled us to assess the effects of non-constant event rates on the different analytical approaches with the same total number of events in each scenario with only the timing of the events differing due to the selected shape of the baseline hazards. We selected modest values of the shape parameter to assess the impact of non-constant event rates in circumstances where an assumption of constant event rates at the design stage of the trial would have been considered appropriate. Use of more extreme values of the shape parameter may have resulted in far more impactful effects on simulation performance measures, but would not have been reflective of typical experiences with clinical trials. The baseline hazard, cumulative hazard and survival functions for the three event rate scenarios for the control and treatment groups are shown in Figure 1.

Figure 1 title: Three event rate scenarios depicted on the hazard scale, cumulative hazard and survival curves

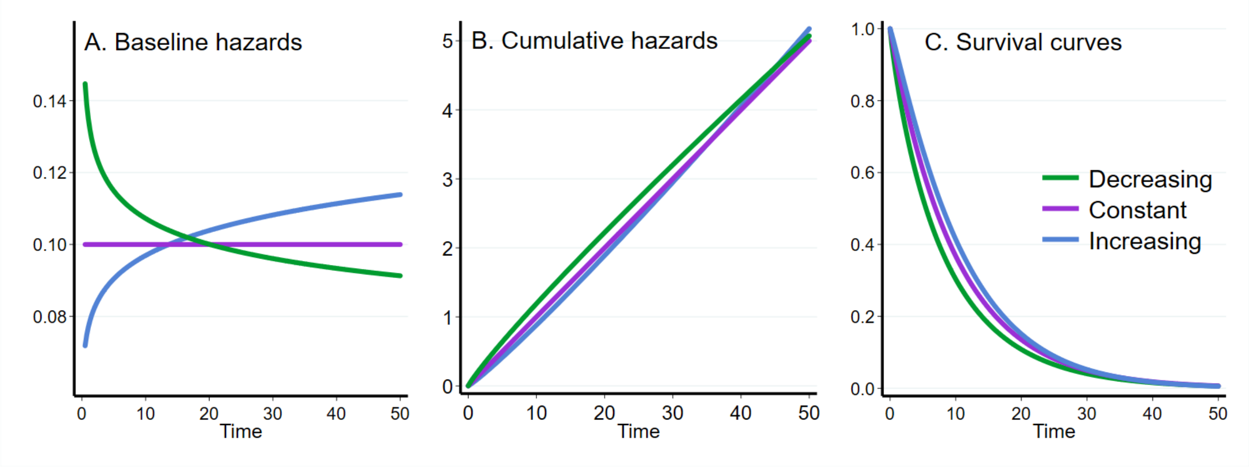
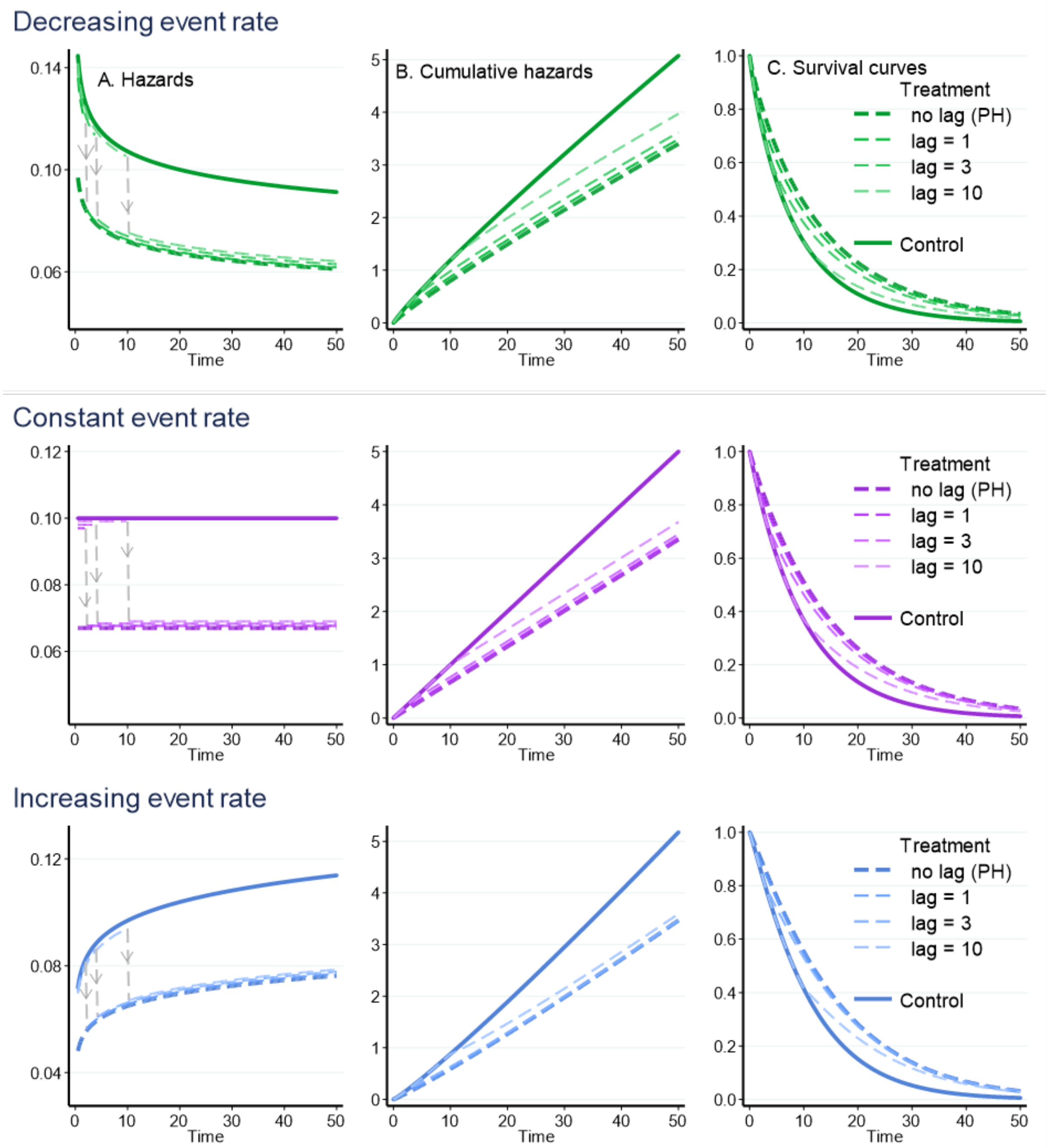


Figure 1 legend: Lines depict baseline hazards – or instantaneous risk of event occurrence in the control group over time – under the three scenarios used for data generation. Decreasing, constant and increasing event rate scenarios are indicated by the green, purple and blue lines respectively. By design, the survival proportion will be the same at t=50 under all three event rates.

Event times were simulated using the survsim command in Stata (27). A binary covariate for treatment group status (X_{trt}) was simulated from a Bernoulli random variable with probability $p = 0.5$ to mimic 1:1 randomisation. Non-proportional hazards were introduced by dividing the analysis time into two periods with a change point at t_{lag} or t_{early} depending on the non-PH scenario. The baseline hazard in the control group was either a decreasing, constant or increasing continuous event rate the same as depicted in Figure 1A. For simulations investigating a lag until treatment effect, the hazard in the treatment group during the first period prior to t_{lag} was the same as in the control group, ie there was no effect of treatment ($\beta = 0$). After t_{lag} the hazard in the treatment group had the anticipated beneficial design effect ($\beta = -0.4$). The lag period lengths investigated were $t_{lag} = 0, 1, 3$ and 10 months within the maximum follow-up time $t = 50$, with the setting $t_{lag} = 0$ representing PH. The three lag durations were selected to enable us to investigate a range of power values and treatment effect magnitudes from the stipulated design values to nearly null values, with the maximum delayed effect of 20% of study duration the longest lag time likely to be encountered in practice. The hazard, cumulative hazard and survival functions for the PH and increasing lag until effect times for the control and treatment groups under the decreasing, constant and increasing event rate scenarios are shown in Figure 2.

Figure 2 title: Hazard functions, cumulative hazard curves and survival curves for lag until effect non-PH scenario.



245

246 Figure 2 legend: Lag period lengths investigated were $t_{lag} = 0, 1, 3$ and 10 months within the

247 maximum follow-up time $t = 50$, with the setting $t_{lag} = 0$ representing PH. The lag period

248 instantaneous change point times from control group hazard to treatment group hazard are

249 indicated by the vertical gray lines. Decreasing times for treatment effectiveness as a result of

increasing lag times are indicated by the decreased shading of the dashed lines used for the treatment group. Decreasing, constant and increasing event rate scenarios are indicated by the green, purple and blue lines respectively.

Simulations were also performed for the scenario of a treatment that is effective for an initial period then ceases. The period prior to t_{early} was the period in which the treatment had the anticipated design effect ($\beta = -0.4$), and the period after t_{early} was when there was no effect of treatment ($\beta = 0$). The early effect period lengths investigated were $t_{early} = 3, 10, 20$ and 50 months, with the setting $t_{early} = 50$ representing PH. Again, these early effect durations were selected to cover power values and treatment effect magnitudes from nearly null to the nominal design values. The DGM section of Table 1 details the simulation characteristics for survival data for three different baseline hazard functions under PH and two different non-PH scenarios. Supplementary Figure S1 presents the hazard, cumulative hazard and survival functions for the PH and early effect that ceases non-PH scenarios for the decreasing, constant and increasing event rates in Additional File 1.

Estimands of treatment effect

The estimands of treatment effect in the simulation study were the hazard ratio, the time ratio and the difference in restricted mean survival time.

Hazard Ratio (HR)

The HR is obtained by comparing the instantaneous event rates in the treatment group ($\mathrm{X}_{trt} = 1$) to the control group ($X_{trt} = 0$). For the Weibull data generation model, the effect of treatment is measured as

$$\mathrm{HR} = \frac{\exp(\beta_0 \ + \beta_1)\gamma t^{\gamma-1}}{\exp(\beta_0)\gamma t^{\gamma-1}} = \exp(\beta_1)$$

where β_1 is the co-efficient of the covariate for treatment group status. In the simulation study comparing different modelling approaches, summary estimates of HR were obtained by fitting a Cox PH model, a piecewise exponential (PE) regression model and a Royston-Parmar model (28) under the assumption of PH (time-fixed treatment effects). Time-period specific estimates of HR, either conditional on being event-free at a pre-specified landmark time point, or from allowing an interaction with a discrete-period time point indicator in the PE model were also measured.

Difference in Restricted Mean Survival Time (*ΔRMST)*

The RMST μ of a time-to-event random variable T is the mean of $\min(T, t^*)$ where the cut off time t^* is greater than zero. RMST can be derived as the area under the survival curve $S(t) = P(T > t)$ from $t = 0$ to $t = t^*$. In a randomised two-group trial with survival functions $S_{X_T}(t)$ and $S_{X_C}(t)$ for the treatment group and the control group respectively, the difference in RMST between groups can be calculated as

$$\Delta\mathrm{RMST} = \int_0^{t^*} [\, S_{X_T}(t) - S_{X_C}(t) \,] \; dt$$

In the simulation study an estimate of the *Δ*RMST was obtained by fitting a RP model under the assumption of PH (RP(PH): time-fixed treatment effects) or allowing for non-PH (RP(TD): time-dependent treatment effects). The *Δ*RMST with t^* taken to be the last uncensored observed event time was obtained by predicting the log cumulative hazard functions for the treatment and the control groups over a grid of time values, transforming into the survival functions and integrating over $(0, t^*)$. Standard errors were estimated using the delta method (29). By using the last uncensored observed event time, the same events were used for the estimation of ΔRMST as were used for the estimates of HR and TR.

Time Ratio (TR)

The TR is an estimand of treatment effect that arises from direct comparison of the time that elapses until experiencing the outcome event, and for the Weibull data generation model used

$$\mathrm{TR} = \left(\frac{-\ln(S(t))^{\frac{1}{\gamma}} \exp(\beta_0 + \beta_1)}{-\ln(S(t))^{\frac{1}{\gamma}} \exp(\beta_0)} \right) = \exp(\beta_1)$$

In the PH parameterisation of a Weibull regression model, the effect of a covariate is multiplicative by a factor of $\exp(\beta)$. In an AFT parameterisation, the effect of a covariate is to accelerate time by a factor of $\exp(\beta)$ where the relationship between the coefficients in the two parameterisations is $\beta_{\mathrm{PH}} = -\beta_{\mathrm{AFT}} \times \gamma$.

Methods to assess treatment effect

Cox Proportional Hazards (PH) Model

In the Cox PH model the hazard rate for the i^{th} individual is $h_i(t) = h_0(t)\exp(X_i\beta)$ with regression coefficients β to be estimated and $h_0(t)$ denoting the baseline hazard function or event rate (30). The estimate of treatment effect from the Cox model is obtained by comparing the hazard in the treatment group to the hazard in the control group to obtain the HR. If non-proportional hazards are anticipated, landmark analyses can be obtained by undertaking a Cox analysis conditional on individuals being event free at the pre-specified LM time point t_{LM}. Events prior to t_{LM} do not contribute to the estimation of the LM HR.

Piecewise exponential (PE) regression

The simplest parametric proportional hazards model is the exponential survival model which assumes that the hazard rate is constant over the entire analysis time. To accommodate a non-constant hazard, a useful extension is the piecewise exponential model which allows the time scale to be split into an arbitrary number of intervals each of differing lengths, with a constant hazard rate assumed within each interval. The PE model can be written as $h_i(t) = \lambda_j \exp(X_i\beta)$ where $h_i(t)$ is the hazard rate for the i^{th} individual, λ_j is the baseline hazard rate for the j^{th} follow up interval, X_i is the vector of covariates for the i^{th} individual and β are log hazard-ratios to be estimated. The PE model provides a summary estimate of the HR for the treatment effect for the entire analysis time, or can be extended to provide period-specific estimates of the (HR_j) for the treatment effect by including an indicator variable for each period with an interaction with treatment effect.

Weibull Accelerated failure time (AFT) model

An alternative parameterisation of the Weibull model is the accelerated failure-time model which has the parameterisation $\ln(t_i) = X_i\beta + \epsilon_i$ where ϵ_i has an extreme value distribution. Under this parameterisation for the Weibull distribution, the treatment effect is estimated as a summary fixed effect TR in an equivalent manner to the summary HR estimated under the PH assumption.

Royston Parmar (RP) models

Royston-Parmar parametric models utilise restricted cubic splines to estimate complex shape functions. The models describe the baseline log cumulative hazard function on the log timescale as a series of cubic spline subfunctions joined at knots with a ‘restriction’ that the first and last subfunctions beyond the boundary knots are linear functions instead of cubic.

The RP PH model can be written as $\ln(H(t)) = s(\ln(t)|\boldsymbol{\gamma_s}, \mathbf{k}_0) + X_i\beta$ where $s(\ln(t)|\boldsymbol{\gamma_s}, \mathbf{k}_0)$ is the restricted cubic spline that is the function of the coefficients of the spline-derived variables $(\boldsymbol{\gamma_s})$ and the number of knots $\mathbf{k_0}$. In the PH context, the RP model is a generalisation of the Weibull distribution where the restricted cubic spline function models the Weibull log cumulative hazard function $\ln[H_0(t)] = \ln(\lambda) + \gamma\ln(t) + X_i\beta$ on the log timescale. The HR and *Δ*RMST for treatment effect can be estimated from this PH model. We assigned 5 degrees of freedom (df) to the baseline distribution which should provide for an adequately flexible fit to a wide variety of survival curves (31). The *Δ*RMST allowing for TD treatment effects was estimated by including interactions between the treatment variable and additional spline function in the RP model. We assigned 5 df to the baseline distribution as in the PH model, and 2 df to the TD treatment effect to account for possible non-PH.

Tests of equal survival functions

Many tests of difference between two survival curves have been proposed that aim to achieve acceptable power under PH and under anticipated non-PH patterns whilst maintaining type I error rates close to the nominal level. Few have become widely accepted as analytical approaches for analysing trials. In this simulation we included tests from two broad categories of test statistics - weighted variants of the LR test designed to improve power under particular non-PH patterns, and omnibus global tests that combine results of several individual tests of significance in an attempt to improve power across a wider range of non-PH patterns. Tests from these two broad categories were identified as the most utilised in recent reviews of analysis methods used in clinical trials with time-to-event outcomes (4, 5).

The classical LR test assesses the null hypothesis that there is no difference between the survival curves of two groups in the probability of an event at any time point over the total survival time period under consideration. The analysis is based on the sum of differences of the estimated hazard function at each observed event time with an implicit equal weighting of one for all event times. Fleming and Harrington proposed a family of weighted tests, the extended $\mathrm{FH}(\rho, \gamma)$ tests with weighting $[\hat{S}(t-)]^{\rho}[1-\hat{S}(t-)]^{\gamma}, \rho, \gamma \geq 0$ where $\hat{S}(t-)$ is the Kaplan-Meier estimate of the survival rate based on the pooled data from the two treatment groups. When $\rho = 0, \gamma = 0$, the $\mathrm{FH}(0,0)$ corresponds to the LR test with equal weights (32). When $\rho > \gamma$, the test gives more weight to earlier events than to later ones, and when $\rho < \gamma$ more weight is given to later events than to earlier ones. In this simulation, the power of the FH tests $\mathrm{FH}(1,0)$, $\mathrm{FH}(1,1)$ and $\mathrm{FH}(0,1)$ weighting early, middle and latter events respectively will be assessed.

The performance of two omnibus tests will be compared in this simulation. The performance of the default form of the versatile test proposed by Karrison (13) considers $Z_m = \max(|Z_1|, |Z_2|, |Z_3|)$ where Z_1, Z_2 and Z_3 are Z statistics from the $\mathrm{FH}(0,0)$, $\mathrm{FH}(1,0)$ and $\mathrm{FH}(0,1)$ extended family respectively, and $Z_m \sim \mathrm{N}_3(\mu, \Sigma)$ an asymptotic, trivariate normal distribution with $\boldsymbol{\mu}$ the vector of means and Σ the variance-covariance matrix. This combination of Z statistics was selected to provide relatively good coverage across the range of likely scenarios encompassing PH, early and late treatment effect scenarios. The second omnibus test which will be assessed in this simulation, the combined test proposed by Royston (12) utilises information from the Cox test and a permutation test based on the maximal squared standardized ΔRMST between treatment groups. The motivation for the development of the combined test was to capitalise on the optimal power of the Cox test when the assumption of PH is met, and to provide some insurance should non-PH be present.

Performance measures

In this simulation study we are interested in assessing the impact of non-constant event rates under two non-PH scenarios on the estimated treatment effect from a range of analysis models. Under PH, the three data-generating models would all result in the same number of events occurring within the specified follow up time. We compared the performance of estimators from an analysis model against the design model knowing that the design model would not necessarily accord with the data-generating model. Discussion of performance measures is in relation to design model using the parameters from the design stage of the trial. This point will be further explained in the context of specific performance measures below.

Power, the first performance measure, was obtained as the proportion of simulations where the p-value was less than the nominal significance level α. The anticipated power specified at the design stage was 80%. The second performance measure was the scaled treatment effect (STE). The mean treatment effect for each simulation scenario was scaled so that a value of 100% corresponded to the full design-stipulated treatment effect, and a value of 0% would be the anticipated magnitude in the absence of any treatment effect. The scaling was calculated as $(1 - \text{mean}[\widehat{HR}])/(1 - HR_{design}) \times 100$ for the HR estimands, as $(\text{mean}[\widehat{TR}] - 1)/(TR_{design} - 1) \times 100$ for the TR estimand, and as $(\text{mean}[\Delta\widehat{\text{RMST}}])/\Delta\text{RMST}_{design} \times 100$ for the ΔRMST with the $\Delta\text{RMST}_{design}$ value obtained empirically from a large N=250,000 simulation of the design setting. This scaling of treatment effect utilizing the exponentiated measures as reported was designed to allow direct intuitive comparison of the impact of the different simulation scenarios on the magnitude of the three different estimands even though they are a mix of relative and absolute measures, and the beneficial treatment effect can be a value less than 1 (HR) or a value greater than 1 (TR and ΔRMST). The final measure, coverage was calculated as the proportion of simulations in which the $100 \times (1 - \alpha)\%$ confidence interval around analysis model $\hat{\beta}$ included the anticipated β from the design model. This allowed assessment of whether the empirical coverage rate approached the desired rate. The anticipated coverage specified at the design stage was 95%.

Number of simulations

We generated 2000 simulated datasets for each scenario. The Monte Carlo standard errors (MCSEs) for coverage and power are maximized when either 50% power or 50% coverage is observed. In this worst-case scenario, the MCSE for the simulation would be 1.1%. Should coverage and power be optimal at 95% and 80% respectively as implemented under the design scenario, the expected MCSEs would be correspondingly less than 0.5% and 0.9% which we deemed to be acceptable.

Results

Type I error

Prior to comparing performance measures such as power for scenarios with a known treatment effect, it is important to assess that analytical approaches are controlling the Type I error level at the same or similar nominal value when there is truly no effect. We compared that empirical Type I errors were maintained reasonably well and similar to other simulation studies (33, 34). Additional detail of the Type I error assessment is presented in Additional File 1.

Lag until treatment effect

Power of regression model approaches

Figure 3 presents the simulation results investigating the effect of lag times for eight different modelling approaches to estimating the HR, *Δ*RMST and TR. For an indication of data maturity, the average number of events for the constant event rate during the no effect period was 10%, 26% and 65% of the total number of events observed for the lag times of one, three and ten months respectively. For the decreasing hazard event rate, the average number of events during the no-effect period were 14%, 34% and 71%, and for the increasing hazard event rate, the average number of events during the no-effect period were 7%, 21% and 60% of the total number of events observed for the lag times of one, three and ten months respectively. A summary of event numbers during the inactive and active phases of treatment effect under this non-PH scenario is presented in Supplementary Table S2 in Additional File 1.

Figure 3 title: Performance measures of regression-based approaches for treatment effect estimation under increasing lag until effect DGM.

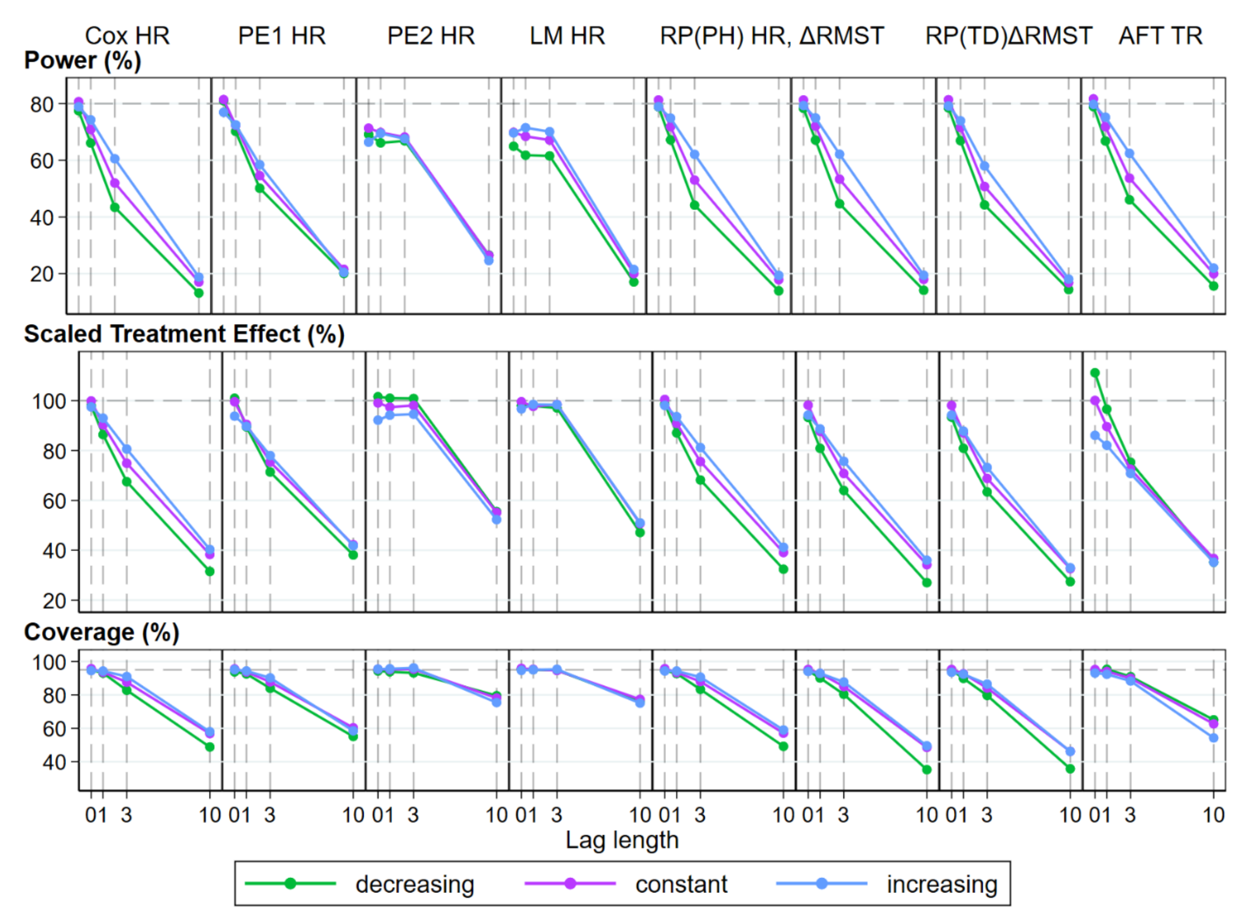


Figure 3 legend: The power (%), scaled treatment effect magnitude (%) and coverage (%) are presented as relative to that anticipated at the design stage of the trial assuming PH. Lag period lengths investigated were $t_{lag} = 0, 1, 3$ and 10 months within the maximum follow-up time $t = 50$, with the setting $t_{lag} = 0$ representing PH.

In the top panel of Figure 3 for the first scenario with no lag to effect ($t_{lag} = 0$, the PH scenario), we observed power very close to the design model value of 80% for all estimates of treatment effect. There was lower power for the two period-specific power estimates (PE2 and LM) resulting from the smaller number of events used in the estimation of HR after the prespecified cut points of t_{PE} and t_{LM} were applied. For all methods, there was an appreciable loss of power in these non-PH scenarios. This loss of power was present even when $t_{lag} = 1$ with greater loss of power observed with increasing lag times.

The impact of non-constant event rates in the presence of non-PH can also be clearly observed, with the difference in power most differentiated when $t_{lag} = 3$.In general, an increasing event rate slightly attenuated the loss of power as a result of fewer events occurring during the lag period, relative to the number of events observed under a constant event rate. Conversely, the losses in power observed under a decreasing event rate in the presence of a lag until effect were magnified as a result of more events occurring during the period where the treatment had no effect. This pattern of relative power loss with non-constant event rates was observed for the HR, TR and *Δ*RMST.

Scaled Treatment Effects (STE) estimates of regression model approaches

The middle panel of Figure 3 presents the STE results. In the scenario of no lag until treatment effect ($t_{lag} = 0$) estimates close to the design model values are observed except for the HR from the PE2 model and the TR from the AFT model. For these two estimators, an increasing event rate resulted in a lower STE under PH whilst a decreasing event rate resulted in a higher STE. The presence of any lag period resulted in STE of decreased average magnitude as there were less events occurring during the period where the treatment was effective. Compared to a constant event rate, an increasing event rate was able to partially ameliorate this decrease in STE whilst a decreasing event rate compounded the decrease.

Coverage of regression model approaches

In the bottom panel of Figure 3, coverage of the estimators for the treatment effect used in the design model is presented. Under PH, we observed coverage at, or very close to, the design model value of 95%. In the presence of a lag until treatment effect, there was a consistent decrease in the observed coverage with increasing lag for all methods. The presence of non-constant event rates has less impact on this performance measure. The summary estimates for bias, coverage and power with the Monte Carlo standard errors (MCSEs) for simulations in the presence of a lag until treatment for the decreasing, constant and increasing baseline hazards are presented in Supplementary Tables S3, S4 and S5 respectively in Additional File 1.

Power of the tests of equal survival curves

Figure 4 presents the results for seven tests of equal survival functions compared in the simulation. The power of the z-test for the treatment effect from the Cox model is included in the panel as a comparator. Results are broadly similar to that observed for the modelling approaches. In the scenario equivalent to PH, the LR, Cox, versatile and combination tests achieved power values close to the design model value of 80%. The power dropped swiftly with increasing lag times. The decreased or increased loss of power observed could be substantial for some tests exceeding ±10% of the power observed under a constant event rate depending on the length of the lag effect under consideration.

Figure 4 title: Power of tests of equal survival function under increasing lag until effect DGM.

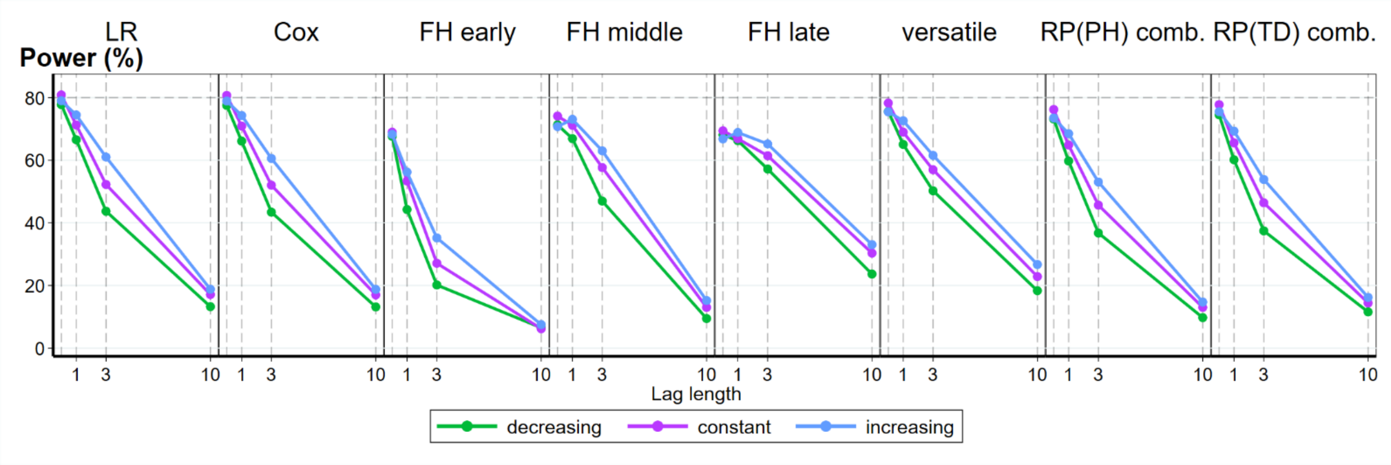


Figure 4 legend: Effect of non-constant event rates on the power of seven tests of equal survival function. The power of the *z*-test for the HR treatment effect from the Cox PH model is included in the panel as a comparator. Lag period lengths investigated were $t_{lag} = 0, 1, 3$ and 10 months within the maximum follow-up time $t = 50$, with the setting $t_{lag} = 0$ representing PH.

Early effect that ceases

The early effect that ceases non-PH scenario is the inverse in treatment effect timing to the lag until treatment effect. The performance measures for the early effect that ceases non-PH scenario were similarly the converse to that observed in the lag until treatment effect non-PH simulations. In summary, increasing losses of power and decreased magnitude of the treatment effects and coverage were observed as the length of the treatment effect period decreased. Relative to a constant event rate, more events occurred during the early effective period under a decreasing baseline hazard resulting in some offset of the losses in performance measures observed. Under an increasing event rate, some reduction of the losses observed under the constant event rate were observed. This pattern of relative loss was observed for all three estimands and similar losses in power were observed in the tests of equal survival curves as were observed for the regression-based approaches. Results are described in more detail in the Supplementary Results section in Additional File 1

Discussion

We have shown that when time-dependent treatment effects are anticipated, then non-PH and non-constant event rates should both be considered at the time of designing a trial. The adverse impact of non-PH on power can be further exacerbated or potentially ameliorated by the shape of the baseline hazard. Non-proportionality of treatment effects has been increasingly observed in clinical trials (16, 35). New treatments being assessed are often more complex, involving comparison of new oncology treatments with different biological time courses of action, or comparing treatments with different mechanisms of action such as surgical versus chemotherapeutic approaches, or involving the use of composite outcomes - multiple endpoints jointly assessed as a primary outcome - all increasing the chance of encountering non-PH (36). Due to increased oversight and increased awareness of the importance of personalised medicine, trials are often longer in planned follow up, with larger numbers of participants included to allow for greater assessment of differently responsive sub-populations within them. Trials of longer duration allow a greater opportunity for non-PH to arise over time, and larger numbers of events enable assessment of the presence of any non-PH to be more conclusive. The potential impact of non-PH has been brought into focus due to these longer, larger trials being conducted (33, 37, 38). For these trials, non-constant event rates will also be more likely to be observed, yet the interplay between non-PH and the shape of the baseline hazard rates has received little attention before now, despite the reasonable anticipation that it could also to have important design implications for clinical trials.

Comparison of power of tests of survival curve difference

In our results, when there was a lag until treatment effect, the best performing test of survival curve difference in terms of maintaining power under PH and shorter and longer lengths of effective treatment time was the versatile test. The FH late test was more powerful when there was longer lags until effect, but was less powerful under shorter lags and PH scenarios more likely to be encountered in trials compared to the versatile test. When there is an early effect that ceases, the versatile test closely followed by the RP(TD) combined test would be the recommended option. Increasing and decreasing event rates affected the power of the tests compared to a constant event rate, in accordance with the timing of when events were likely to be observed with respect to the periods of effective treatment. Power was increased when relatively more events occurred during effective treatment times and decreased when relatively fewer events occurred during effective treatment times. At the time of designing a trial, if assumptions about the presence and form of non-PH are not made, then our results suggest that the versatile test covering PH, early and late forms of non-PH is recommended as a pre-specified analysis method. This test will retain power under more modest levels of non-PH whilst maintaining near nominal power under PH and will be less adversely affected by non-constant event rates.

Our results accord with similar comparative studies published recently that focus on tests of survival curve difference (33, 34, 38). As part of Cross-Pharma Non-Proportional Hazards (NPH) working group, Lin et al (2020) compared nine tests of survival curve difference in the presence of non-PH covering the LR and weighted LR tests, weighted Kaplan-Meier based tests (incorporating the RMST) and combination tests (38). Royston and Parmar also included a similar range of tests covering weighted LR tests and composite tests based on their own (39) and Karrison's work (13). Jimenez et al (2019) investigated the properties of the weighted LR tests in the presence of trials with delayed effects (34). There is substantial overlap between the tests included in this simulation study and the three other studies, with similar focus on early (treatment effects that cease) and late (lag until treatment effect) forms of non-PH. For the tests of survival curve difference in the presence of any non-PH, broadly similar conclusions were reached by all four studies: that what might have been regarded as minimal amounts of non-PH - whether expressed in terms of information fraction or percent of study duration - can noticeably affect the power to detect survival curve differences, and for the trials assessing different forms of non-PH, there is no consistently powerful test across all non-PH scenarios. Forms of a versatile test combining information from multiple weighted LRs were the recommended form of pre-specified test when considering early and late non-PH scenarios (33, 38). When late non-PH is the only consideration, LR tests weighted to emphasize late differences are recommended to maintain higher power albeit at the expense of slight Type I error rate inflation (34).

Treatment effect estimands - HR v RMST v AFT

We compared three different estimands for treatment effect - the HR, the TR and *Δ*RMST. There have been many studies comparing these estimands and variants of them for their use in research with TTE outcomes (14, 16, 20, 21, 40-43). There are strengths and limitations in their usage - relative measures such as the HR and TR do not contain any information about the absolute effect and can be challenging to interpret and communicate the survival benefit observed. Estimates provided in a time-based metric such as the TR and the RMST expressed either as a ratio or a difference, can be considered more interpretable for a wider audience. The ΔRMST has an additional advantage of being a summary measure of survival time distribution that does not rely on the PH assumption although it does require specification of the cutoff timepoint. In this work, we estimated ΔRMST using both the last uncensored event occurrence as the cutoff time following recommended practice (25) as well as the maximum follow up time ($t = 50$). By design, the last uncensored event would have been expected to occur at a time very close to the maximum follow up time. As a consequence of these design choices, we observed essentially no differences within simulation error in any of the performance measures of ΔRMST using either the last uncensored event cut off or the maximum follow up time, and hence presented the results for the last uncensored event time cutoff only in the interests of clarity.

For this work, the three estimands we compared were broadly similar across the non-PH scenarios in terms of the power, magnitude of treatment effect estimate and coverage values benchmarked to the values specified by the design model. Judicious selection of designated cutpoints for no effect (PE2) or landmark timepoints (LM) could result in improved estimates of treatment effect magnitude using the period-specific analysis methods in the presence of a lag until effect non-PH, but also resulted in decreased power if there was PH. Similarly, the ΔRMST could be assessed at a number of prespecifed clinically relevant time points in order to provide insight into how treatment effects may change with follow up time. The potential for increased Type I error that may arise from multiple comparisons would need to be monitored, and empirical measures to correct for any inflation would have to be incorporated into the trial design (34).

The impact of non-constant event rates in the presence of non-PH was to partially diminish or further exacerbate losses in power and treatment effect magnitude. When time-dependent treatment effects are present, there is no single summary measure that can adequately describe the treatment benefit. Analysis methods such as the RP models which allow for the shape of the baseline hazard make it possible to more fully explore the timing and magnitude of any treatment effect either graphically or in a series of time period-based estimates.

Designing trials with non-constant event rates in the presence of non-PH

Simulation studies can only ever include a limited range of scenarios. It is critical that selections are made so as to provide insight on the wider and varied spectrum of scenarios involving non-PH and non-constant event rates that are likely to be encountered in real RCTs. We restricted attention to simplified forms of non-PH - piecewise constant HRs with a single change point - comparing PH with early and late forms of non-PH. Change points were placed at times that enabled us to observe effects over a large proportion of calculated power values with magnitudes of treatment effect ranging from the design-stipulated to nearly null estimates. Hence our results may not generalize to more complex forms of non-PH. When choosing non-constant event rates, we aimed to cover clinically plausible values of the shape parameter in our data-generating Weibull model that are modest and hence might be assumed to be ‘close enough’ to constant at the design stage of a trial. More extreme settings could have been chosen and the impacts on power and effect estimation would have been exaggerated to the point of being quite drastic; however, we felt that this would represent uncommon scenarios in practice. Our simulations also featured almost complete follow up of all events before undertaking analysis which, whilst unrealistic in some applications, resulted in almost identical numbers of total events being observed in each scenario, and hence provided a fair basis for comparison. We did not cover the effects of censoring and enrolment rates, nor did we investigate the effect of adjusting sample size and follow up times all of which impact on the interplay of non-PH and event rates and may need to be considered in practice. Sample size calculation options are available for specific forms of non-PH (44), parametric event rates (45, 46), piecewise models that allow for different treatment effects within multiple 'stages' of a planned trial (47, 48). However, the most flexible approach to take is to base the sample size on simulation (49, 50). These approaches have been employed in multi-arm multi-stage and other forms of adaptive trial design. The additional complexity includes the need for prior specification of additional parameters and a higher degree of programming skill to explore scenarios covering anticipated event rates and the direction and timing of non-proportionality.

Conclusions

The mechanisms of action of treatments on time to event outcomes may require nuanced definitions of treatment effectiveness that go beyond simple single summary estimates assuming proportional hazards. Our simulations found that even small deviations from proportionality can result in substantial observed loss of power using standard analysis methods that are maximally powerful under a PH assumption, and this loss can be exacerbated in the presence of non-constant event rates. It is a desirable strategy to design trials to use analysis methods that can accommodate delayed treatment effects, or early treatment effects that cease if these are to be anticipated with the treatment under study. This however requires decisions on what test to employ and what estimand(s) will be the target. Our simulations provide some guidance on this choice. In practice, new trials may require the use of bespoke simulation studies to guarantee that power is maintained under a range of plausible scenarios consistent with expected mechanisms of treatment action and allowing for departures from non-constant underlying event rates.

List of abbreviations

RCT: Randomised controlled trial; PH: Proportional hazards; LR: logrank; HR: Hazard ratio; FH: Fleming-Harrington; RMST: restricted mean survival time; AFT: accelerated failure time; RP: Royston -Parmar; LM: landmark; TR; time ratio; DGM: data-generating model; PE: piecewise exponential; TD: time-dependent; STE: scaled treatment effect; MCSE: Monte Carlo standard error

Declarations

Ethics approval, accordance and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and material

All data generated or analysed during this study are included in this published article

[Additional_file_2.pdf].

Competing interests

The Authors declare that they have no competing interests.

Funding

KJ was supported in part by an Australian Government Research Training Program (RTP) Stipend and RTP Fee-Offset Scholarship through Federation University Australia and a National Health and Medical Research Council of Australia grant (APP1128222). The funding bodies had no role in the design of the study, the collection, analysis and interpretation of data or in the writing of the manuscript.

Authors' contributions

KJ conceived the simulation, and drafted the manuscript. RW helped with the drafting of the manuscript. SH revised draft versions of the manuscript. All authors read and approved the final manuscript.

Acknowledgements

Not applicable

References

1. Altman DG, De Stavola BL, Love SB, Stepniewska KA. Review of survival analyses published in cancer journals. Br J Cancer. 1995;72(2):511-8.
2. Mathoulin-Pelissier S, Gourgou-Bourgade S, Bonnetain F, Kramar A. Survival End Point Reporting in Randomized Cancer Clinical Trials: A Review of Major Journals. J Clin Oncol. 2008;26(22):3721-6.
3. 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.
4. Jachno K, Heritier S, Wolfe R. Are non-constant rates and non-proportional treatment effects accounted for in the design and analysis of randomised controlled trials? A review of current practice. BMC Med Res Methodol. 2019;19(1):103.
5. Rahman RM, Fell G, Ventz S, Arfe A, Vanderbeek AM, Trippa L, et al. Deviation from the Proportional Hazards Assumption in Randomized Phase 3 Clinical Trials in Oncology: Prevalence, Associated Factors and Implications. Clin Cancer Res. 2019:clincanres.3999.2018.
6. Zhang X, Long Q. Modeling and prediction of subject accrual and event times in clinical trials: a systematic review. Clinical Trials. 2012;9(6):681-8.
7. Ferrara R, Pilotto S, Caccese M, Grizzi G, Sperduti I, Giannarelli D, et al. Do immune checkpoint inhibitors need new studies methodology? J Thorac Dis. 2018:S1564-S80.
8. Castañon E, Sanchez-Arraez A, Alvarez-Manceñido F, Jimenez-Fonseca P, Carmona-Bayonas A. Critical reappraisal of phase III trials with immune checkpoint inhibitors in non-proportional hazards settings. Eur J Cancer. 2020;136:159-68.
9. Fleming TR, Harrington DP. Weighted Logrank Statistics. Counting Processes and Survival Analysis: Wiley Series in Probability and Statistics; 2005. p. 255-85.
10. Yang S, Prentice RL. Assessing potentially time-dependent treatment effect from clinical trials and observational studies for survival data, with applications to the Women's Health Initiative combined hormone therapy trial. Stat Med. 2015;34(11):1801-17.
11. Magirr D, Burman C-F. Modestly weighted logrank tests. Stat Med. 2019;38(20):3782-90.
12. Royston P, Parmar MK. Augmenting the logrank test in the design of clinical trials in which non-proportional hazards of the treatment effect may be anticipated. BMC Med Res Methodol. 2016;16(1):16.
13. Karrison TG. Versatile Tests for Comparing Survival Curves Based on Weighted Log-rank Statistics. Stata Journal. 2016;16(3):678-90.
14. Royston P, Parmar MK. Restricted mean survival time: an alternative to the hazard ratio for the design and analysis of randomized trials with a time-to-event outcome. BMC Med Res Methodol. 2013;13(1):152.

15. Uno H, Claggett B, Tian L, Inoue E, Gallo P, Miyata T, et al. Moving Beyond the Hazard Ratio in Quantifying the Between-Group Difference in Survival Analysis. J Clin Oncol. 2014;32(22):2380-5.

16. Trinquart L, Jacot J, Conner SC, Porcher R. Comparison of Treatment Effects Measured by the Hazard Ratio and by the Ratio of Restricted Mean Survival Times in Oncology Randomized Controlled Trials. J Clin Oncol. 2016;34(15):1813-9.

17. Royston P. Estimating the treatment effect in a clinical trial using difference in restricted mean survival time. Stata Journal. 2015;15(4):1098-117.

18. Stensrud MJ, Hernán MA. Why Test for Proportional Hazards? JAMA. 2020;323(14):1401-2.

19. Wei LJ. The accelerated failure time model: A useful alternative to the cox regression model in survival analysis. Stat Med. 1992;11(14-15):1871-9.

20. Kay R, Kinnersley N. On the Use of the Accelerated Failure Time Model as an Alternative to the Proportional Hazards Model in the Treatment of Time to Event Data: A Case Study in Influenza. Drug Inf J. 2002;36(3):571-9.

21. Swindell WR. Accelerated Failure Time Models Provide a Useful Statistical Framework for Aging Research Exp Gerontol. 2009;44(3):190-200.

22. Mahmoud KD, Lennon RJ, Holmes DR. Event Rates in Randomized Clinical Trials Evaluating Cardiovascular Interventions and Devices. The American Journal of Cardiology. 2015;116(3):355-63.

23. Charles P, Giraudeau B, Dechartres A, Baron G, Ravaud P. Reporting of sample size calculation in randomised controlled trials: review. BMJ. 2009;338.

24. Royston P, Sauerbrei W. Multivariable Model-Building. A Pragmatic Approach To Regression Analysis Based On Fractional Polynomials For Modelling Continuous Variables: John Wiley & Sons, Ltd; 2008.

25. Royston P, Lambert PC. Flexible Parametric Survival Analysis Using Stata: Beyond the Cox Model: Stata Press; 2011.

26. Schoenfeld DA, Richter JR. Nomograms for Calculating the Number of Patients Needed for a Clinical Trial with Survival as an Endpoint. Biometrics. 1982;38(1):163-70.

27. Crowther MJ, Lambert PC. Simulating complex survival data. Stata Journal. 2012;12(4):674-87.

28. Royston P, Parmar MKB. Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. Stat Med. 2002;21(15):2175-97.

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

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

31. Royston P, Parmar MKB. An approach to trial design and analysis in the era of non-proportional hazards of the treatment effect. Trials. 2014;15:314.

32. Harrington DP, Fleming TR. A Class of Rank Test Procedures for Censored Survival Data. Biometrika. 1982;69(3):553-66.

33. Royston PB, Parmar MK. A simulation study comparing the power of nine tests of the treatment effect in randomized controlled trials with a time-to-event outcome. Trials. 2020;21(1):315.

34. Jiménez JL, Stalbovskaya V, Jones B. Properties of the weighted log-rank test in the design of confirmatory studies with delayed effects. Pharm Stat. 2019;18(3):287-303.

35. Rahman R, Fell G, Trippa L, Alexander BM. Violations of the proportional hazards assumption in randomized phase III oncology clinical trials. J Clin Oncol. 2018;36(15_suppl):2543-.

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

37. Eaton A, Therneau T, Le-Rademacher J. Designing clinical trials with (restricted) mean survival time endpoint: Practical considerations. Clinical Trials. 2020;17(3):285-94.

38. Lin RS, Lin J, Roychoudhury S, Anderson KM, Hu T, Huang B, et al. Alternative Analysis Methods for Time to Event Endpoints Under Nonproportional Hazards: A Comparative Analysis. Statistics in Biopharmaceutical Research. 2020;12(2):187-98.

39. Royston P. A combined test for a generalized treatment effect in clinical trials with a time-to-event outcome. Stata Journal. 2017;17(2):405-21.

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

41. Coory M, Lamb KE, Sorich M. Risk-difference curves can be used to communicate time-dependent effects of adjuvant therapies for early stage cancer. J Clin Epidemiol. 2014;67(9):966-72.

42. Zhao L, Claggett B, Tian L, Uno H, Pfeffer MA, Solomon SD, et al. On the restricted mean survival time curve in survival analysis. Biometrics. 2016;72(1):215-21.

43. Dehbi H-M, Royston P, Hackshaw A. Life expectancy difference and life expectancy ratio: two measures of treatment effects in randomised trials with non-proportional hazards. BMJ. 2017;357.

44. Sit T, Liu M, Shnaidman M, Ying Z. Design and analysis of clinical trials in the presence of delayed treatment effect. Stat Med. 2016;35(11):1774-9.

45. Wu J. Power and Sample Size for Randomized Phase III Survival Trials Under the Weibull Model. J Biopharm Stat. 2015;25(1):16-28.

46. Phadnis MA, Wetmore JB, Mayo MS. A clinical trial design using the concept of proportional time using the generalized gamma ratio distribution. Stat Med. 2017;36:4121-40.

47. Barthel FMS, Babiker A, Royston P, Parmar MK. Evaluation of sample size and power for multi-arm survival trials allowing for non-uniform accrual, non-proportional hazards, loss to follow-up and cross-over. Stat Med. 2006;25(15):2521-42.

48. Bratton DJ, Choodari-Oskooei B, Royston P. A Menu-driven Facility for Sample-size Calculation in Multiarm, Multistage Randomized Controlled Trials with Time-to-event Outcomes: Update. The Stata Journal. 2015;15(2):350-68.

49. Hooper R. Versatile sample-size calculation using simulation. Stata Journal. 2013;13(1):21-38.

50. Wittes J. Sample size calculations for randomized controlled trials. Epidemiol Rev. 2002;24(1):39-53.

Chapter 4

4.1 Manuscript introduction: Examining evidence for time-dependent treatment effects using alternative regression-based methods in clinical trials

Although the last two decades have seen interest in alternative regression approaches to modelling time-dependent treatment effects, the uptake of these methodologies has been limited as described in the review undertaken for this thesis and reported in Chapter 2. As an illustration of the potential application of these methods, an applied project in presented in this chapter. The project consisted of examining the evidence for time-dependent treatment effects in selected endpoints from a large, long-running community-based clinical trial. The ASPREE trial aimed to determine if aspirin improved healthy ageing with a primary composite endpoint of death, dementia or persistent physical disability and a range of secondary endpoints. Data collection in ASPREE was comprehensive. The 19,114 participants had regular assessments multiple times per year through face to face visits, phone call contacts and medical records review and linkage. Retention was high with follow up for a median of 4.7 years (IQR 3.6-5.7 years) for the trial.

The motivations for the study relate to illustrating the potential for new insights or increased clinical understanding into the magnitude and persistence of treatment effects for selected endpoints. Such insights could be obtained even in the absence of any compelling evidence of nonproportionality. We investigated potential time-dependent treatment effects of aspirin directly for each of the endpoints, and also the existing evidence for time-dependent interaction effects of aspirin usage by age and gender subgroups. Relative and absolute estimands of treatment effect provided complementary information about the evolution of treatment impact over time.

Four modelling approaches for the estimation of the summary treatment effect estimated as either a HR or a ΔRMST were used in the study. The HRs were obtained under the assumption of PH from

(i) the semi-parametric Cox model,

(ii) the parametric Weibull model, and

(iii) the flexible parametric models using restricted cubic splines to model the baseline hazard.

The ΔRMSTs were estimated using

(iv) the spline-based FPMs assuming PH, ie the same model as in (iii),

(v) FPMs allowing for time-dependence and

(vi) generalised linear modelling of transformed datasets of pseudo-observations which allow for non-parametric estimation of treatment effect equivalent to Kaplan-Meier estimation of survival probability.

From the simulation study presented in Chapter 3, we focused on regression-based methods that allow for multiple measures of treatment effect estimation and graphical presentations that are suitable for facilitating communication, clinical evaluation and understanding.

The main content of this chapter is presented in the next section in the form of an applied research paper written with input from clinicians and ASPREE trial investigators that has been submitted to the journal *Pharmaceutical Statistics* and is currently under review. The supplementary material for the paper is provided in Appendix E of this thesis.

Examining evidence for time-dependent treatment effects using alternative regression-based methods in clinical trials

Corresponding Author:
Kim Jachno
Address: School of Public Health and Preventive Medicine, Monash University, Melbourne, Victoria, Australia
Email: kim.jachno@monash.edu

Stephane Heritier
Address: School of Public Health and Preventive Medicine, Monash University, Melbourne, Victoria, Australia
Email: stephane.heritier@monash.edu

Robyn L.Woods
Address: School of Public Health and Preventive Medicine, Monash University, Melbourne, Victoria, Australia
Email: robyn.woods@monash.edu

Suzanne Mahady
Address: School of Public Health and Preventive Medicine, Monash University, Melbourne, Victoria, Australia and Gastroenterology, Melbourne Health, Parkville, Victoria, Australia
Email: suzanne.mahady@monash.edu

Andrew Chan
Address: Clinical and Translational Epidemiology Unit, Department of Medicine, Massachusetts General Hospital, Boston, MA, USA
Email: achan@mgh.harvard.edu

Andrew Tonkin
Address: School of Public Health and Preventive Medicine, Monash University, Melbourne, Victoria, Australia
Email: andrew.tonkin@monash.edu

Anne Murray
Address: Berman Center for Outcomes and Clinical Research, Hennepin Health Research Institute, Hennepin, Minneapolis, MN, USA and Division of Geriatrics, Department of Medicine, Hennepin County Medical Center and University of Minnesota, Minneapolis, MN, USA
Email: amurray@bermancenter.org

John J. McNeil
Address: School of Public Health and Preventive Medicine, Monash University, Melbourne, Victoria, Australia
Email: john.mcneil@monash.edu

Rory Wolfe
Address: School of Public Health and Preventive Medicine, Monash University, Melbourne, Victoria, Australia
Email: rory.wolfe@monash.edu

Drafted for submission to Pharmaceutical Statistics

Key words: time-dependent treatment effects, proportional hazards, clinical trials, flexible parametric modelling, treatment effect heterogeneity

Abstract

For the design and analysis of clinical trials with time-to-event outcomes, the Cox proportional hazards model and the logrank test have been the cornerstone methods for many decades. Increasingly, the key assumption of proportionality – or time-fixed effects - that underpins these methods has been called into question, and with it the presentation of fixed-magnitude treatment effects as the key inferential findings of a trial. The availability of novel therapies with new mechanisms of action and clinical trials of longer duration mean that non-proportional hazards are now more frequently encountered.

We compared several regression-based methods to model time-dependent treatment effects. For illustration purposes we used selected endpoints from a large, community-based clinical trial of low dose daily aspirin in older persons. Relative and absolute estimands were defined and analyses were conducted in all participants. Additional exploratory analyses were undertaken by selected subgroups of interest using interaction terms in the regression models.

In the trial with median 4.7 years follow-up, we found evidence for non-proportionality and a time-dependent treatment effect of aspirin on cancer mortality not previously reported in trial findings. We also found some evidence of time-dependence to an aspirin by age interaction for major adverse cardiovascular events. For other endpoints time-fixed treatment effect estimates were confirmed as appropriate. The consideration of treatment effects using both absolute and relative estimands enhanced clinical insights into potential dynamic treatment effects. We recommend these analytical approaches as an adjunct to primary analyses to fully explore findings from clinical trials.

Section 1: Introduction

The most commonly utilised approach for analysis of time-to-event data in clinical trials is the Cox proportional hazards (PH) model [1]. The advantage of this model is its lack of assumptions about the shape of the underlying hazard functions and presentation of treatment effects on a relative scale as hazard ratios (HRs). Increasingly, trials are being conducted in which the key assumption of PH that underpins this approach, and presentation of the treatment effect summarised as being of single fixed magnitude is questionable [2, 3]. Trials of longer duration and larger trials enable investigation of the natural history of the disease and interplay of mechanistic processes over time. They offer compelling rationale for consideration of alternate measures of treatment effect that allow for the examination of non-PH treatment effects over time. Examples of time-dependent (TD) effects include delays until treatment effectiveness as observed in immunotherapy-based oncology trials with minimal benefit in the first few months of treatment followed by a period of effectiveness after the immune system has been activated. In contrast, vaccinations for influenza and whooping cough provide examples of a treatment that is beneficial early after administration but whose effectiveness diminishes over time. Despite the potential importance of TD treatment effects, detailed assessment and reporting of the PH assumption required to assess the appropriateness of presented time-fixed trial results has been less than optimal [4-6].

Parametric models that make assumptions about the shape of the underlying hazard function can be used as an alternative to the Cox model. Models based on the Weibull and gamma distributions can specify increasing, decreasing and inverted hazard functions. However, these models may fail to capture more complex hazard function. A flexible parametric model (FPM) uses spline functions to model the underlying hazard function of any shape or complexity with the advantages of modelling within a regression-based framework [7]. Specifying the baseline hazard allows for the direct estimation of relative and absolute effects of treatment in addition to other useful measures such as differences between survival and hazard functions to be estimated. In particular, the use of the restricted mean survival time (RMST) difference between groups as a distribution-free measure of treatment effect has been gaining attention as a valid measure of treatment effect even when nonproportionality is present [8, 9].

In addition to capturing complex hazard functions under PH [10], flexible parametric survival models can be easily extended to assess for TD treatment effects on the cumulative hazard or hazard scales [11, 12]. A second regression-based method to assess for evidence of TD treatment effects involves pseudo-observations - or jackknife estimates - based on the non-parametric Kaplan-Meier (KM) curves. These pseudo-observations are used to create estimates constructed in such a way that their sample mean estimates the parameter of interest at pre-determined times of interest. The effect of covariates may then be modelled with the pseudo-observations as the response variable in generalised linear models (GLMs) with a suitable link function [13, 14].

Heterogeneity of treatment effects is another form of non-PH that can arise in clinical trials. Treatment effect heterogeneity is when different subgroups of a trial population respond differently to treatment. Prior clinical knowledge of potentially strong predictive factors can - and should - be incorporated into the study design and prespecified analysis plans through selection of sufficiently homogeneous populations that can be expected to benefit from the treatment [15, 16]. Subgroup heterogeneity may in itself also be time-dependent hence reported averaged treatment effects, even in subgroup analysis, can obscure interesting insights available from the trial [17].

The goal of this paper is to examine whether regression-based methods allowing for TD treatment effects can provide additional or new insights. For illustration we apply the methods to the effects of daily low-dose aspirin in initially healthy older persons using the large community-based ASPirin in Reducing Events in the Elderly (ASPREE) clinical trial. The ASPREE trial aimed to determine if aspirin improved healthy ageing with a primary composite endpoint of death, dementia or persistent physical disability. Secondary efficacy and safety endpoints were also collected. For some endpoints event rates were anticipated to substantially increase with ageing. The large number of participants and long duration of the treatment phase of the trial provide an opportunity to assess the evidence for potential TD treatment effects of clinical interest and to investigate any potential interplay between underlying event rates and non-PH. Editorials accompanying the trial findings support the need for ongoing follow up of the ASPREE participants to more robustly address hypotheses regarding benefits or harms of aspirin on endpoints in this older population, with additional mechanistic studies particularly for cancer incidence and mortality being critical [18-20].

The rest of the paper is structured as follows: in Section 2 we give a brief introduction to the different methods used. In Section 3 we provide further detail of the ASPREE trial and a selection of endpoints chosen to best illustrate the functionality and interpretability of modelling time dependence of treatment effects. In Section 4 we present the ASPREE results using the methods described. Finally, we provide discussions and recommendations in Section 5.

Section 2 METHODS

We compare four regression-based approaches for the estimation of the summary treatment effect estimated as either a hazard ratio (HR) or a difference in restricted mean survival time (*Δ*RMST). The HR estimates were obtained from the Cox model, the Weibull model and the spline-based flexible parametric model (FPM) all under an assumption of PH. The *Δ*RMST was estimated using the FPM PH model, the FPM allowing for time-dependence of treatment effects and from generalised linear modelling of transformed datasets consisting of pseudo-observations, being jackknife estimates of time-to-event observations for a specific pre-designated time interval had there not been censoring present.

2.1 Semi-parametric Cox PH model

Under a Cox proportional hazards model [1], the hazard function for the i^{th} patient can be written as

$$h_i(t) = h_0(t)\exp(x_i\beta)$$

where x_i represents covariates with regression coefficients β (log hazard ratios) to be estimated from the data and $h_0(t)$ denotes the baseline hazard function or event rate when all of the covariates are equal to zero or at their specified baseline levels.

The Cox PH model treats the baseline hazard function as a nuisance parameter by maximising the partial likelihood function which permits estimation of the regression parameters but not the baseline hazard function. A key assumption of the Cox PH model is that of PH, in that the effect of a covariate remains constant or fixed in magnitude over the entire follow up. The Cox model can be extended to incorporate non-proportional effects by including an interaction of the covariate(s) of interest with some function of time. Various diagnostics have been proposed to assess the PH assumption including graphical approaches and analysis based on residuals or by including an interaction of a covariate of interest with a function of time [21, 22]. These tests of PH assumption require correct specification of the function of time and often lack power to detect non-proportionality [23].

2.2 Parametric Weibull model

When non-constant event rates are anticipated, parametric models are an alternative to the Cox model [6]. Undertaking a parametric approach to the analysis of survival data has a number of benefits. By directly modelling the baseline hazard function, measures of absolute risk, as well as relative risk, can be directly quantified with an associated estimate of uncertainty. There are efficiency gains if the baseline hazard is correctly specified in a parametric approach compared to the equivalent semi-parametric approach. The modelling of TD effects in continuous time can be conducted more easily within a parametric framework. In the ASPREE trial, monotonically increasing event rates were anticipated - and observed - for the majority of the endpoints which motivated the use of a Weibull hazard function to model the baseline hazard rate for this work. The estimates of treatment effect from this fixed distributional parametric approach act as a comparator to both the semi-parametric Cox model and the more flexible parametric models described below.

2.3 Royston-Parmar flexible parametric models (FPMs)

Royston and Parmar introduced FPMs that use restricted cubic splines to model transformations of the survival function, most commonly using the log cumulative-hazard function [7, 24] and later extended to the log hazard function [25] as a tool to capture simple and more complex hazard functions under both PH and non-PH scenarios. In this way, the attraction of the Cox model - allowing the shape of the baseline hazard to be free of any distributional assumptions - is still achieved by allowing the basis function of cubic splines to flexibly fit the baseline hazard. Additionally, FPMs attain the efficiency of parametric models for estimation and interpretability, providing both relative and absolute estimates of treatment effect.

FPMs use restricted cubic spline functions to model the transformation of the survival function. Restricted cubic splines are piecewise cubic polynomials joined together at ‘knots’ with smoothing constraints placed on knot joins, and a restriction that the spline function is linear beyond the first and last knots to ensure an overall smooth function that is not unduly affected by sparse data. In the general approach, FPMs are implemented on the log cumulative hazard scale using one set of spline variables with predefined knot positions based on evenly spaced centiles of uncensored log survival times, with boundary knots at the minimum and maximum uncensored log survival times. The number of knots used to model the baseline hazard can be guided by clinical input and model selection criteria.

Time-dependent effects were modelled using a different set of spline variables for each covariate of interest, possibly using a different number of knots in potentially different locations than the spline variables used to model the baseline hazard. Defining $\mathbf{k_0}$ to denote the number of knots for the baseline hazard function, $\mathbf{k}_j$ to denote the knots for the jth TD effect with associated parameters $\boldsymbol{\delta}_j$ when there are D covariates with TD effects, the log cumulative hazard model is

$$\ln[H_{\boldsymbol{i}}(t|x)] \; = \; s(ln(t)|\boldsymbol{\gamma}, \boldsymbol{k}_0) \; + \int_{j=1}^{D} s(ln(t)|\boldsymbol{\delta}_j\,, \boldsymbol{k}_j)\mathbf{x}_{\boldsymbol{i}j} + \; \mathbf{x}_{\boldsymbol{i}}\boldsymbol{\beta}$$

In order to assess the complexity required for the baseline hazard for each endpoint of the ASPREE trial, a series of preliminary models were fit with varying numbers of knots considering possible degrees of freedom (df) ranging from one df to five df for the baseline spline function. Comparisons were then made between the models visually and through using the Akaike information criterion and Bayesian information criterion statistics with smaller values preferred. For all endpoints assessed, allowing for one (corresponding to the Weibull distribution) to three df for the baseline hazard resulted in suitably smooth curves without evidence of overfitting. Time dependence of the treatment effect could be captured with either one or two df for the five different endpoints. We utilised a model with three df for the baseline hazard and allowed for two df for any TD treatment effect [10, 26]. This was a compromise between the most parsimonious model for any given endpoint and the clinical utility of fitting the same model to each of the endpoints.

Figure 1 is a graphical presentation of a hypothetical example where non-proportionality of the treatment effect was present. The true hazard functions (dashed lines), modelled hazards (solid lines – panels b-d) and treatment effects (arrows) in the form of HRs that would arise from application of the Cox PH, the Weibull and the PH and TD flexible modelling approaches are depicted. The arrows in the Cox PH approach (panel a) represent the constant HR with the absence of solid lines underlining that the hazard function need not be estimated. The solid lines in the Weibull and PH flexible modelling approaches (panels b, c) illustrate the constant HR estimated in these approaches. Finally, the varying arrow sizes in the TD flexible modelling approach (panel d) indicate that the estimated treatment effect varies over time, unlike the models represented in panels a-c.

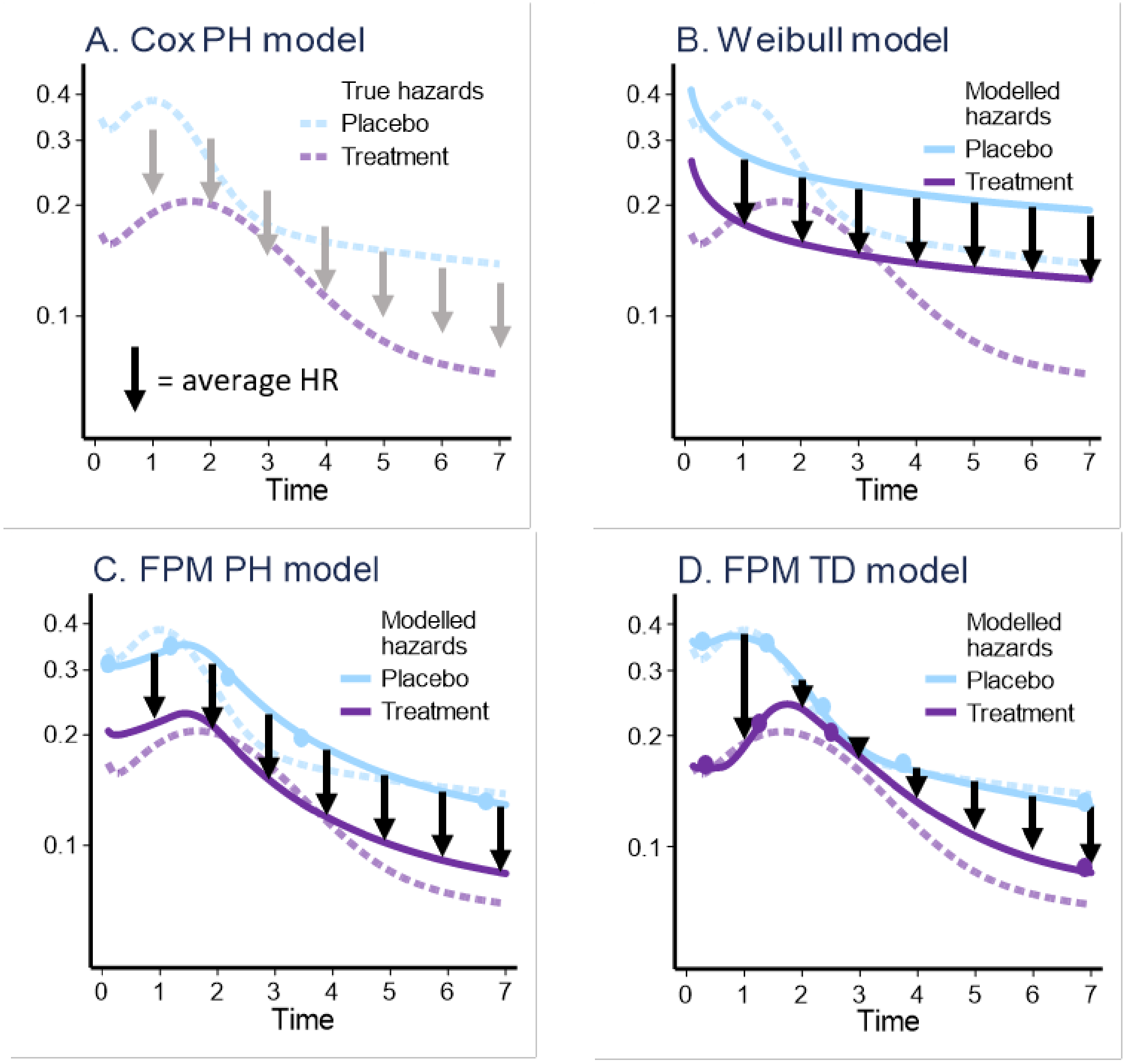


Figure 1: Estimated hazards (y-axes) and treatment effects from the Cox PH, the Weibull, the FPM PH and TD models when non-proportionality of the true hazards (dashed lines) was present. The arrows indicate the magnitude and direction of treatment effect as measured from the modelled baseline hazard (solid light blue line) to the modelled treatment line (solid purple line).

2.4 Pseudo-observations approach

Pseudo-observations provide non-parametric estimates of a parameter of interest at the individual participant level [13]. Pseudo-observations are jackknife estimates constructed in such a way that their sample mean estimates the parameter of interest, here the RMST. The pseudo-observations are a transformation of the original data that provides a dataset without censoring. The effect of covariates such as treatment group on the RMST may then be modeled with the pseudo-observations as the outcome variable in GLMs with an appropriate link function. Standard errors of parameter estimates use the robust "sandwich" estimator. The treatment effect estimates of ΔRMST obtained through the pseudo-observations approach are distribution-free since they are based on the KM survival curve estimates and can be used to compare the magnitude of the ΔRMST estimates from the TD FPM. To maintain comparability of the HRs and ΔRMST estimates obtained by the comparator methods, the pseudo-observations approach used the last uncensored event time in the dataset for each endpoint as the time point chosen at which to estimate the mean survival. For analyses of the yearly incremental estimates of treatment effect included as a guide to assessing for non-PH of the main treatment effect, the indicated duration of time was used to estimate the ΔRMST.

Section 3 The ASPirin in Reducing Events in the Elderly (ASPREE) Trial

The ASPREE trial was a community-based randomised trial comparing daily low-dose aspirin versus placebo with the aim of extending the duration of disability-free survival in healthy older adults and was conducted in the US and Australia. Inclusion criteria included ages 70 years or above, except for African-American and Hispanic participants in the US who were included from age 65 years. Reporting of the ASPREE trial on the primary endpoint and other clinical endpoints utilised a Cox PH modelling approach. This analysis was carried out because the PH assumption was deemed plausible for the primary endpoint components [27-29].

Our analyses were facilitated by the comprehensiveness of data collection in ASPREE, with recruitment of 19,114 participants who attended regular face to face annual study visits for a median of 4.7 years (IQR 3.6-5.7 years). In addition, all major endpoints were adjudicated by Endpoint Committees whose members were blinded to treatment allocation. This enabled us to examine evidence for TD effects of aspirin as well as investigate treatment-covariate interactions of interest. These analyses are to be viewed as supplementary subsidiary analyses to the pre-specified primary analyses already published. Our aim is to illustrate the methods for investigating the magnitude and duration of any treatment effect over time, overall and in specific subgroups of participants even when there was no statistical evidence against the assumption of proportionality.

In this paper, we reexamine the analysis of the primary endpoint of disability-free survival and four other selected endpoints, clinically significant bleeding, major adverse cardiovascular events (MACE), solid tumour cancer incidence and solid tumour cancer mortality. For each endpoint, we estimate the summary HR treatment effect measure presented previously utilizing three different regression-based approaches. Additionally, we provide the summary ΔRMST treatment effect measure estimated using the same events as for estimation of the summary HR, and graphically display the HR and ΔRMST endpoint measures over time.

3.1 Disability-free survival

Disability-free survival was the primary endpoint of the ASPREE trial. It was a composite endpoint defined as survival free from dementia or persistent physical disability and was derived from the time to first occurrence of any one of the three components of death, dementia or persistent physical disability in an individual. The endpoint aimed to capture the qualitative and quantitative components of an ongoing healthy life span in an older population considered sufficiently healthy to be enrolled in a primary prevention trial.

Details regarding the health measures and definitions used in the trial and the primary conclusion that aspirin use in healthy older adults did not prolong disability-free survival (HR 1.01, 95% confidence interval (CI) 0.92 to 1.11, p-value=0.79) have been reported elsewhere [27].

3.2 Clinically significant bleeding

An increased risk of a clinically significant bleeding event is an adverse effect of aspirin usage [30]. The clinically significant bleeding endpoint of the ASPREE trial included haemorrhagic stroke, symptomatic intracranial bleeding and clinically significant extracranial bleeding, which were defined as bleeding that led to hospitalisation, prolongation of hospitalisation, surgery or death. The trial showed the risk of bleeding was significantly higher with aspirin than with placebo (HR 1.38, 95% CI 1.18 to 1.62, p<0.001). The observation of a constantly increasing separation of cumulative incidence curves suggested that the rate of participants newly experiencing bleeding was constant over time [28]. Our analyses further assess and quantify the evidence for persistence of a constant elevated bleeding risk associated with aspirin over the duration of the trial.

3.3 Major adverse cardiovascular events (MACE)

MACE was a non-prespecified composite endpoint which included fatal coronary heart disease (excluding death from heart failure), nonfatal myocardial infarction, and fatal or nonfatal ischaemic stroke. These events were adjudicated as part of the broader cardiovascular disease endpoints, and included the conditions related to ischaemia and atherothrombosis that were anticipated to be affected favourably by low-dose aspirin. The effect of aspirin on MACE events in the trial has been reported previously as a HR of 0.89, 95% CI 0.77, 1.03 [28].

3.4 Solid tumour cancer mortality and incidence

Cancer incidence was a prespecified endpoint in the trial. At the time of the trial's conception, there was emerging evidence to suggest that low dose regular aspirin usage may be a potential cancer preventative [31]. As participants with a history of cancer were able to enter the trial, incident cancer events included in analysis required diagnosis of new site-specific cancers post randomisation. For the present analysis, only solid tumour cancers were considered in order to be consistent with previous analyses [31]. The effect of aspirin on solid tumour cancer incidence was reported as a HR of 1.05, 95%CI 0.95 to 1.14; the effect of aspirin on cancer mortality was reported as a HR of 1.35, 95%CI 1.13 to 1.61 [32]. Possible time-dependence of these cancer endpoints was acknowledged with additional mechanistic studies and further follow up called for [19]. We aim to further explore possible time-dependence of treatment effect for the solid tumour cancer endpoints as suggested by progressive separation of the cumulative incidence curves in previous reports [29, 32].

SECTION 4 Results

Table 1 presents results for the two estimands of treatment effect (HR and ΔRMST) for the selected five endpoints. HR estimates were obtained from the Cox PH model, the Weibull model and the FPM PH model. ΔRMST estimates were obtained from the FPM PH model, the FPM TD model and the pseudo-observations (p-obs) dataset. The duration of time at which the final summary estimates of HR were assessed extended from time of randomisation to the time of last endpoint in the trial dataset. The same time period was used for the estimation of the ΔRMST.

Table 1: Summary of the ASPREE trial results for five endpoints using regression-based modelling approaches assuming PH or allowing for TD treatment effects.

Endpoint	Estimation model	HR (95% CI), p-value	Estimation model	ΔRMST (95% CI), p-value
Primary	Cox PH	1.01 (0.92,1.11), 0.79	FPM PH	-0.006 (-0.047, 0.035), 0.79
	Weibull	1.01 (0.92,1.11), 0.79	FPM TD	-0.005 (-0.046, 0.036), 0.81
	FPM PH	1.01 (0.92,1.11), 0.79	GLM p-obs	-0.007 (-0.049, 0.035), 0.75
MACE	Cox PH	0.89 (0.77,1.03), 0.12	FPM PH	0.021 (-0.006, 0.049), 0.13
	Weibull	0.89 (0.77,1.03) 0.12	FPM TD	0.021 (-0.006, 0.048), 0.12
	FPM PH	0.89 (0.77,1.03), 0.12	GLM p-obs	0.021 (-0.008, 0.050), 0.16
Clinically significant bleeding	Cox PH	1.38 (1.18,1.62), <0.001	FPM PH	-0.050 (-0.075, -0.026), <0.001
	Weibull	1.38 (1.18,1.62), <0.001	FPM TD	-0.052 (-0.077, -0.027), <0.001
	FPM PH	1.38 (1.18,1.62), <0.001	GLM p-obs	-0.057 (-0.084, -0.029), <0.001
Cancer incidence	Cox PH	1.05 (0.95,1.15), 0.32	FPM PH	-0.020 (-0.059, 0.019), 0.32
	Weibull	1.05 (0.95,1.15), 0.32	FPM TD	-0.018 (-0.058, 0.021), 0.36
	FPM PH	1.05 (0.95,1.15), 0.32	GLM p-obs	-0.024 (-0.068, 0.020), 0.29
Cancer mortality	Cox PH	1.36 (1.13,1.63), 0.001	FPM PH	-0.032 (-0.047, -0.013), 0.001
	Weibull	1.36 (1.13,1.63), 0.001	FPM TD	-0.029 (-0.048, -0.010), 0.003
	FPM PH	1.36 (1.13,1.63), 0.001	GLM p-obs	-0.033 (-0.055, -0.012), 0.003

For all five endpoints, the summary results presented here for the Cox PH model agree with the previously reported results in the main and follow up trial publications [27-29, 32, 33]. The three modelling approaches with the underlying PH assumption gave almost identical estimates of the HR. P-values from the three PH modelling approaches and across the HR and ΔRMST estimates from the FPM PH model were also similar. There were some differences between the estimates of ΔRMST from the flexible TD and pseudo-observation modelling approaches, however these were small and unlikely to have any substantive impact on the clinical interpretation of the results. The FPM PH modelling approach provides a link between the HRs and ΔRMSTs, giving a means to relate the magnitude of treatment effect of a relative hazard reduction to an absolute decreased mean survival time on average. As an illustration, for the clinically significant bleeding endpoint, a 38% increased relative risk of bleeding expressed in terms of the ΔRMST could be equivalently expressed as on average during the trial, a participant on low-dose aspirin would have experienced a bleeding event 0.050 years – or approximately 18 days – sooner than a participant on placebo.

For each endpoint, the HR and ΔRMST at yearly incremental durations of time after randomisation are additionally presented in Supplementary Tables S1-S5. These yearly estimates are a tabular subset of the PH and TD analyses of treatment effect presented in panels C and D of Figures 2 and 3 (and Supplementary Figures S1-S3). Qualitative assessment of TD treatment effects comes from comparing the HRs from yearly incremental durations of follow up, and by comparing the overall HRs with the duration-specific HRs. This is undertaken here regardless of statistical evidence to indicate non-proportionality of treatment effect so caution is warranted with these exploratory analyses to avoid over-interpretation.

Concerning solid tumour cancer mortality, there was an overall increased risk (HR 1.36, 95% CI 1.13, 1.63) found at the end of the trial using a Cox model. However, for this endpoint there was statistical evidence to indicate non-proportionality of treatment effect (PH test p=0.01 [22]) with the incremental assessments providing some insight into the evolution of this treatment effect. The estimated hazard ratio gradually changed from 0.90 for the first year of the trial (95% CI 0.47,1.73) to 1.20 (95% CI 0.96, 1.50) suggestive of a possible adverse effect of treatment emerging at four years from randomization (Supplementary Tables S1-S5).

For the major haemorrhage endpoint there was no statistical evidence to indicate non-proportionality of treatment effect, and although an initial higher treatment-related adverse effect was seen during the first year of follow up this stabilised to a lower - but still adverse - effect for the remaining years.

For the primary endpoint, MACE and cancer incidence endpoints, the similarity of the duration-specific HRs over time suggest that a summary estimate of treatment effect was appropriate with little to suggest any time-dependence of effect.

4.1 Exploring time-dependence of treatment effect for the solid tumour cancer mortality endpoint

Figure 2 shows a four-panel graphical presentation of the treatment effect over time for the cancer mortality endpoint. Figure 2, panel A (top left) shows KM survival curves for aspirin and placebo arms, an FPM analysis assuming PH and an FPM analysis allowing for TD of the treatment effect. The KM curves shown in black for the aspirin (solid lines) and placebo (dashed lines) arms in the top left panel (A) show little difference in the first 2-3 years with an apparent separation of the two curves beginning from year 3 onwards. The survival curves from a conventional analysis assuming PH (blue curves) appear to capture the pattern reasonably well. However, even with the greatly expanded y-axis used here, differences in the survival proportions can be difficult to discern graphically. The summary HR from the conventional FPM PH model estimates the treatment effect as 1.36 (95% CI 1.13, 1.63; p=0.001) and the ΔRMST to be -0.032 (-0.052, -0.013; p=0.001) indicating worse outcomes in the aspirin arm. The survival curves from the analysis allowing for a TD treatment effect (green curves) are able to capture the lack of separation of the non-parametric KM curves in the first few years of the trial and the increasing separation in the latter years.

The hazard rates by treatment group are presented in Figure 2, panel B. On this scale, the initial lack of separation of the two groups, followed by a clear separation can be clearly discerned in the curves generated from the FPM allowing for a TD treatment effect. An indication of uncertainty is provided with a shaded 95% CI around the estimated curves. Figure 2, panel C is the difference in RMST (ΔRMST) between the two curves assessed at incremental durations of time since randomisation over the time period 0.25-6.75 years. The emergence of a treatment effect in later years of follow-up is apparent and it is evident, on the ΔRMST scale, regardless of whether a PH model or a TD model is used. The timing of the emergence of the delayed adverse treatment effect appears to differ between the chosen models. The PH analysis resulted in a larger estimate of treatment effect at all follow up times considered.

In Figure 2, panel D, the HR estimates as a function of time since randomisation from the PH and TD analyses of treatment effect are presented. Compared to the summary HR from the PH analysis presented as the constant horizontal line, the HR estimates in the TD analysis varied from an initial small non-significant benefit during the first year of the trial to a gradually increasing harmful effect of aspirin. From a likelihood ratio test of model fit, there is evidence to suggest that the TD model better fits the data compared to the PH model (p=0.03).

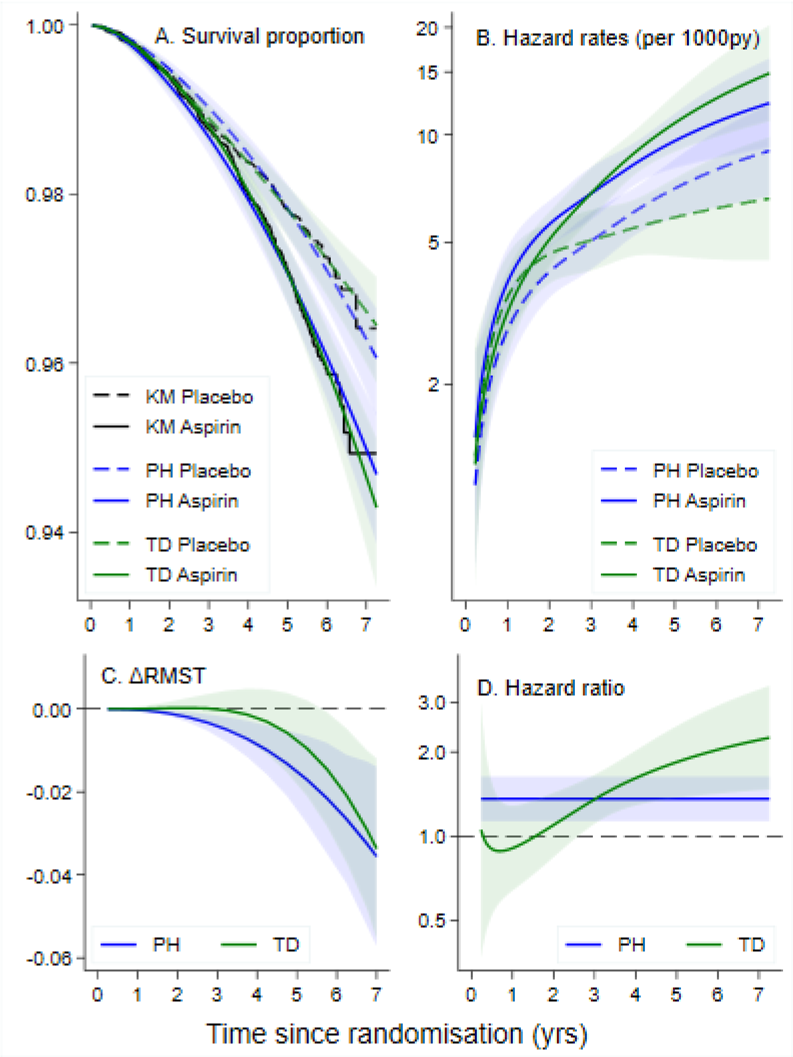


Figure 2: Survival curves (panel A) and hazard rates (panel B) by treatment arm, and difference in RMST (ΔRMST; panel C) and HR (panel D) over time from PH (blue curves) and TD (green curve) analysis models for the cancer mortality endpoint. Y-axes scales are chosen to emphasise any model or treatment differences.

4.2 Absence of any time dependence of treatment effect for the primary and other ASPREE endpoints

An exploratory analysis of treatment effect on disability-free survival, the ASPREE primary endpoint, presented in Figure 3, shows the survival curves for the aspirin and placebo arms of the trial are almost identical for the entire duration of the trial (panel A). There was no evidence of a treatment effect and the summary HR estimate of 1.01 (95% CI 0.92, 1.11; p=0.79) or the ΔRMST of -0.006 (-0.047, -0.035; p=0.79) provide an adequate description of the lack of effect of aspirin on this composite outcome over the duration of the trial. Even with an expanded survival proportion axis, the survival proportion curves for the aspirin and placebo arms are almost identical for the entire duration of the trial. The duration of follow up captured by these analyses is from randomisation until the last uncensored event time in the dataset occurring at 7.01 years.

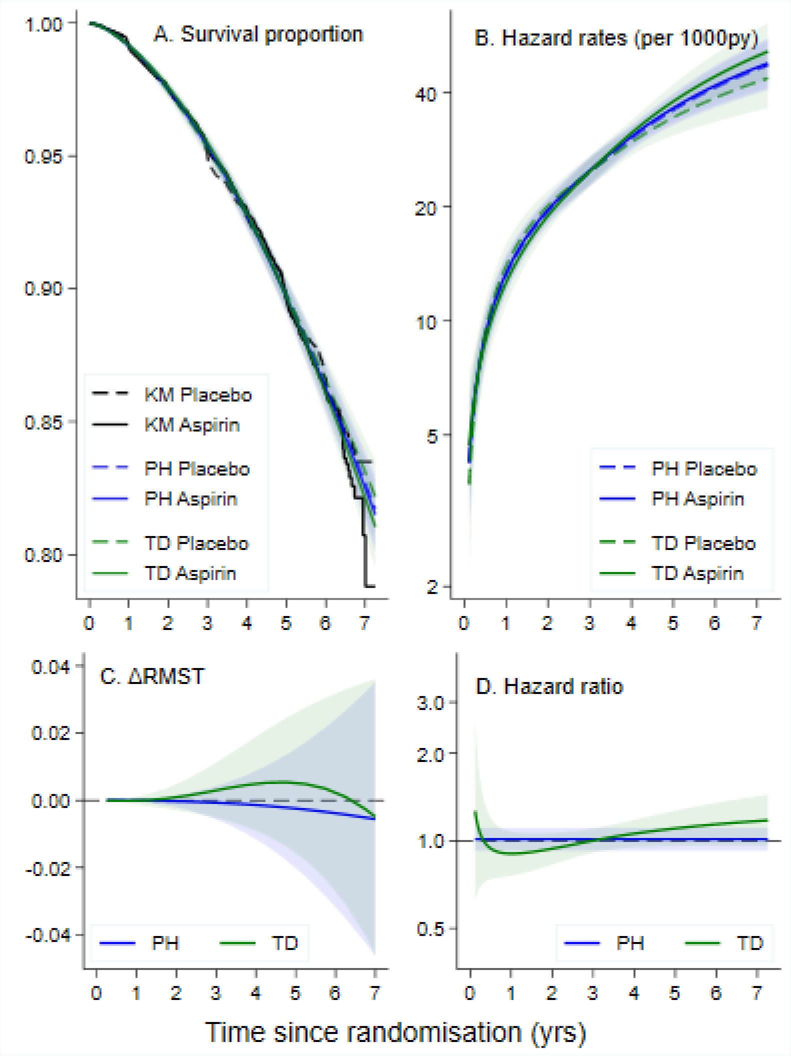


Figure 3: Survival curves (panel A) and hazard rates (panel B) by treatment arm, and difference in RMST (ΔRMST; panel C) and HR (panel D) over time from PH and TD analysis models for the composite primary endpoint.

Similar four panel presentations for the MACE, clinically significant bleeding and cancer incidence endpoints are in Supplementary Figures S1, S2 and S3. For the MACE and cancer incidence endpoints, there is little to differentiate visually between the PH and TD analysis models, confirming the appropriateness of applying single summary estimates of treatment effect for these three endpoints. There is an overall increased risk of clinically significant bleeding due to aspirin with some suggestion that this risk is highest for the first six months after commencement of daily usage. This transitory treatment effect is explored further as part of assessing for time-dependent treatment effects by sex (section 4.3). For all three endpoints, there is no suggestion of improvement of the overall model fit from the likelihood ratio tests comparing the PH and TD approaches.

4.3 Time-dependent treatment effects by subgroup: clinically significant bleeding in males and females

The flexible modelling approaches being examined here can also be used to provide additional insight into interactions between time-dependent treatment effects and subgroups of interest. Here, this is conducted as a post-hoc exploratory analysis although it could form part of a pre-specified analysis plan.

For the clinically significant bleeding endpoint, from a comparison of the HR from PH and TD models (see Supplementary Figure S2 panel D) there is some evidence for an elevated risk in the first year of taking low dose aspirin daily (HR 1.84 95% CI 1.25, 2.70, p=0.002), which then plateaued after the first year to a lower, but still elevated risk (HR 1.30 95% CI 1.08, 1.55, p=0.003) similar to the reported overall HR 1.38 95%CI 1.18, 1.62, p<0.001 for the overall treatment effect from the PH model. Published subgroup analysis by sex did not show strong evidence of different treatment effects in males and females (males HR=1.21 95% CI 0.97, 1.51; females HR=1.58, 95% CI 1.26, 1.99; interaction p-value = 0.1) [27]. The potential time-dependence of this interaction is explored visually in Figure 4.

For males, the increased risk of a major bleeding event due to aspirin was at its highest during the first few months although a still-elevated risk persisted throughout the follow-up and was estimated to be approximately constant after the first year of treatment. Compared to males, females had a higher increased risk of bleeding due to daily aspirin usage throughout follow-up. For females, the acute increased risk persisted for most of the first year, and this risk decreased more slowly over the duration of the trial than males. The shaded area in Figure 4 indicates the uncertainty band around the estimated time-dependent HR for all participants enrolled in the trial and highlights the increasing uncertainty at later timepoints. Supplementary Figure S4 contains graphs for the difference by sex in the HR(t) from the TD analysis for the other four endpoints under consideration.

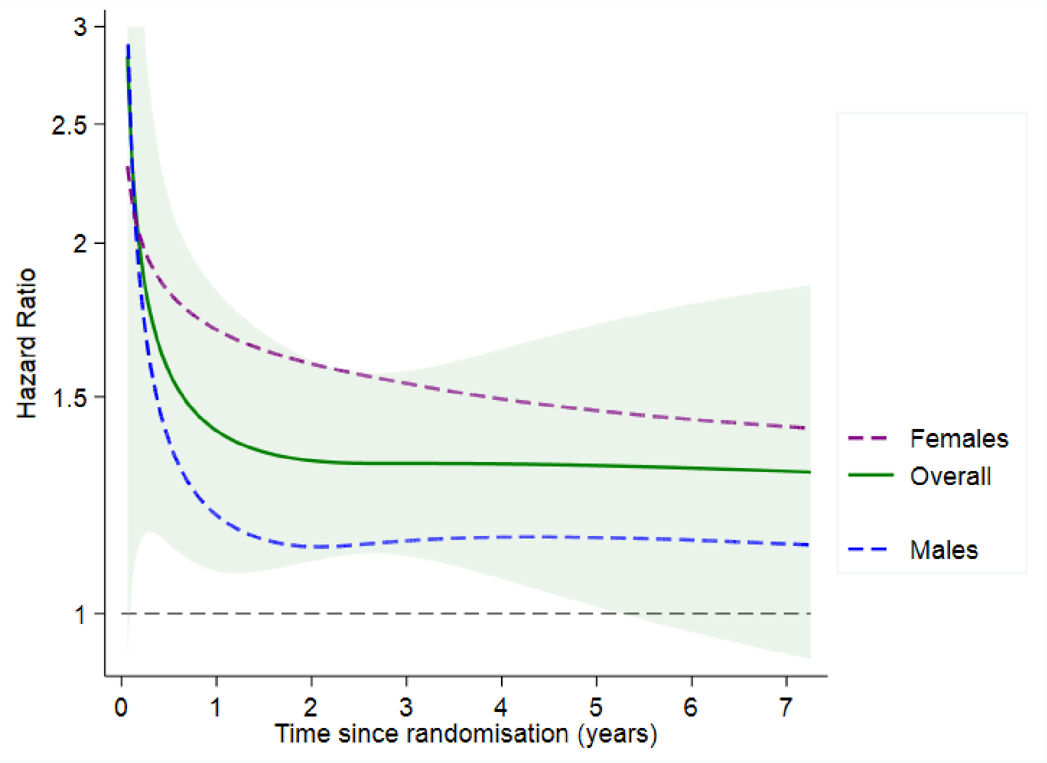


Figure 4: Assessing time-dependence of effect of aspirin for males and females on risk of clinically significant bleeding. The overall estimated HR(t) for treatment effect is the solid green line with the shaded green area indicating the 95% CI width. The HR(t) for treatment effect estimated from females only is indicated by a purple dashed line, and the HR(t) for treatment effect estimated from males only indicated by the blue dashed line.

4.4 Time-dependent treatment effects by subgroup: MACE by age as a continuous covariate

Insight into potential treatment effects and continuous predictor covariates can also be obtained using the FPM approaches. For the ASPREE primary analysis, subgroup effects by age at randomisation were examined categorised as younger than the median age (<74 years) vs older (74+ years) as specified in the statistical analysis plan. For illustration purposes here in order to maximise power to detect any treatment effect interactions, age was analysed on a continuous scale.

For the MACE endpoint a tendency towards a greater beneficial treatment effect for the <74 yrs age group (HR=0.76, 95% CI 0.59, 0.97) compared to the 74+ age group (HR=0.97, 95%CI 0.81, 1.17) has been reported although this interaction was not statistically significant (p-value$_{int}$ = 0.11) [27]. To illustrate application of the method, age at baseline was included in the PH FPM model as a continuous covariate with an assumed linear association with the endpoint. The evidence of an interaction effect between aspirin and (continuous) age at randomization was summarized by p$_{int}$ = 0.06. When allowing for TD of the effect of aspirin and age on MACE the evidence of an interaction effect between aspirin and age was similar (p$_{int}$ = 0.04). Figure 5 presents these PH and TD FPM analyses assessing treatment effect of aspirin according to age for the MACE endpoint. When a linear relationship between age and MACE was assumed – and one accepts the hypothesis that there is an interaction - the FPM PH analysis showed a protective effect of aspirin at younger baseline ages, increasing towards an absence of any benefit at older ages (blue line with 95% CI shaded area). From the exploratory analysis of the time-dependence of this effect depicted in the green lines in Figure 5, there is some evidence to suggest that the possible beneficial effect of aspirin for ASPREE participants younger than the median was greatest during the earlier years following randomization and reduced with time. For participants older than the median, there was no evidence of any benefit of aspirin during the trial. Supplementary Figure S5 contains graphs of the effect of age with treatment for the other ASPREE endpoints examined in this report. There was no evidence of any interaction effect between aspirin and age in either the PH or TD FPM analyses for these other endpoints.

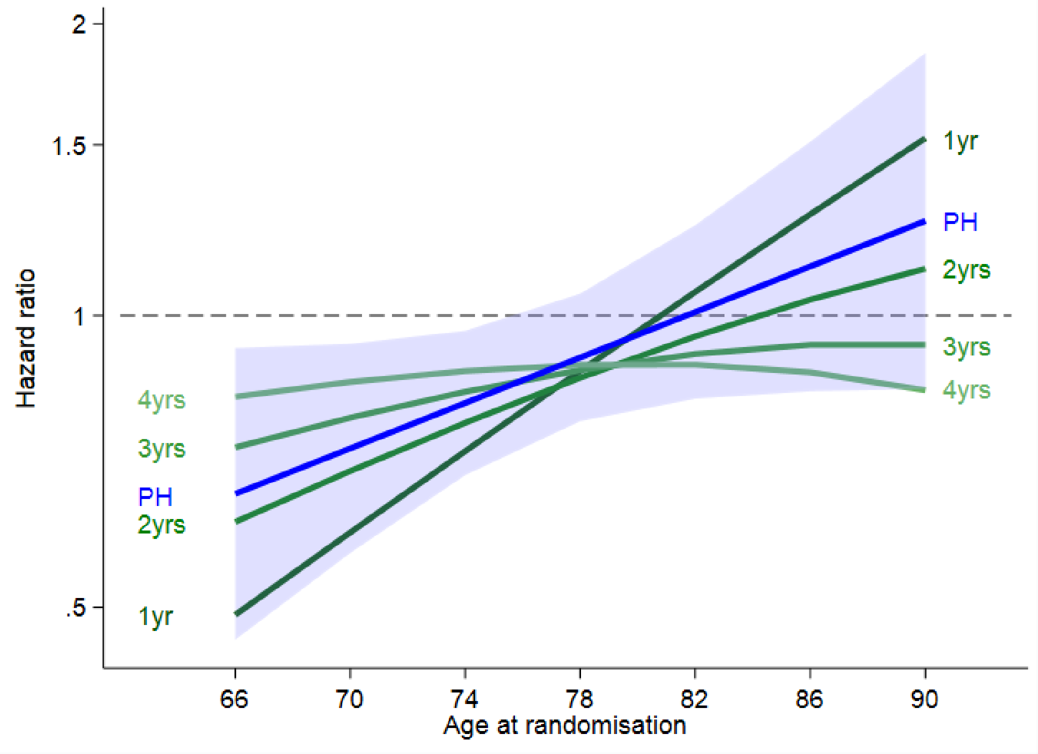


Figure 5: The effect of aspirin on age at randomisation in PH and TD analysis for the MACE endpoint. The estimated age by treatment interaction effect from the PH model is the solid blue line with the shaded area indicating the 95% CI width. The interaction treatment effect from the TD model at yearly intervals is indicated by the green lines with color intensity decreasing over time.

SECTION 5 Discussion

In this paper we demonstrated the potential for increased clinical insight using regression-based analysis methods to model the time-dependence of treatment effects compared to methods that assume proportionality of the treatment effect. For five endpoints of the ASPREE trial, we compared the results obtained using the Cox and Weibull PH models to alternative flexible modelling methods utilising splines that are suitable in the context of non-PH and which describe time-dependent treatment effects. We have shown enhanced interpretability by flexibly modelling the baseline hazard or by using the approach of pseudo-observation jackknife estimates in a generalised linear modelling approach. We have further demonstrated the potential of the flexible modelling approaches to explore time-dependent treatment effect heterogeneity in subgroups.

There has been a proliferation of research into analysis methods when non-PH is anticipated or detected with much focus on weighted adaptations to the standard logrank (LR) test in the presence of specific forms of non-pH such as delayed effects [34-39]. Combination tests have also been proposed that combine multiple weighted LR tests and/or weighted LR tests with tests for non-PH designed to provide robust power to detect survival curve differences under a range of non-PH scenarios [40-45]. These hypothesis testing approaches have been aimed at maintaining power to detect statistical significance in clinical trials in the primary analysis. We have focused instead on regression-based approaches and graphical exploratory analyses to examine the evidence for TD treatment effects. In particular, we have utilised the flexible parametric modelling approach as, unlike test-based approaches, it provides estimation of treatment effects under PH and non-PH.

From a clinical perspective, there is utility in being able to present any treatment effects with estimates in both risk-based and time-based metrics which provide complementary information. They provide equivalent information albeit on different metrics when a one-summary treatment effect is sufficient to describe the findings from a trial. When treatment effects vary over time, the different metrics may provide insight into the timing and duration of period specific effects reflective of clinician and patient interest. For three endpoints in the ASPREE trial: disability-free survival, MACE and cancer incidence, a single HR or ΔRMST provided an appropriate and clinically meaningful summary of the effect of aspirin in healthy older adults, similar in magnitude and direction of treatment effect for the entire duration of the trial. In contrast, for solid tumour cancer mortality and clinically significant bleeding, there was some evidence of time-dependent treatment effects that we now discuss in further detail.

The possible time-dependence of the effect of aspirin on solid tumour cancer mortality suggested adverse effects of treatment emerging by the third year of the trial. We provided evidence that the time-dependent model was a more appropriate fit to the trial data than the proportional hazards model used in the original trial analyses. The findings contrast with the longer-term beneficial effects of aspirin observed in other RCTs. Previously postulated hypotheses to account for this unexpected increase in cancer mortality suggest that the effect of aspirin may have biological effects that vary according to the timing of the exposure, or vary according to age or other participant-specific characteristics. It is conceivable that aspirin may have short-term actions on pathways specific to ageing or tumour cell types in older hosts that could explain the worsened survival among participants in ASPREE in the absence of any apparent effect on cancer incidence [20]. Continued follow up of ASPREE participants is currently underway to examine legacy effects of the intervention.

For clinically significant bleeding, plausible observations of clinical interest from an analysis of time-dependent treatment effects were seen. An increased risk with aspirin was durable to five years of exposure and beyond. There appeared to be a particularly elevated risk of bleeding events with aspirin in the first few months after beginning treatment, which by the end of the first year of follow-up had plateaued to a lower but still increased harmful effect which was then sustained for the remainder of follow up. However, care is required not to over-interpret this conclusion as the existence of this time-dependence of treatment effect was not confirmed by a statistical test. Hence clinical and mechanistic plausibility should be considered carefully, and additional studies would be necessary to confirm the working hypotheses regarding any time-dependent aspirin treatment effects.

Further insights into the potential benefits and harms of treatment effects can be demonstrated using flexible modelling approaches by incorporating categorical covariates for subgroups, and by allowing continuous covariates to be investigated assuming linear and more flexible spline functional forms. These analyses can provide a more nuanced understanding of potential treatment subgroup heterogeneity and time-dependent treatment effects. Clinical trials are rarely adequately powered to detect interaction effects so any findings need to be considered with the requisite understanding of the exploratory nature of these investigations.

For the clinically significant bleeding endpoint of ASPREE, by allowing for the treatment effect to differ in males and females and allowing that difference to be time-dependent, we were able to demonstrate an acute period of higher risk upon starting daily aspirin usage for both males and females. Our analyses also suggest that females had a relatively higher increased risk of clinically significant bleeding at all times compared to males.

Previous assessments for possible treatment-age interactions for the MACE endpoint had been performed using pre-specified categorical groupings of the continuous age at randomisation covariate. Based on the selected categorisations, there had been little evidence to suggest any treatment-age interaction effect (see Supplement S7, S8 in [27]). Our detailed exploratory analysis suggested a beneficial effect of aspirin for ASPREE participants younger than the median age (<74 years) particularly in the early years of follow up, but for older participants (74+ years), there was no indication of aspirin benefit during the trial.

To more fully report the information in a trial, tabulation of both relative and absolute measures of treatment effect at key times of clinical interest, and graphical presentation of complementary measures of treatment effect over time for subgroups should be encouraged. In this way, readers can ascertain any time-dependence of treatment effects and subgroup heterogeneity. We note that apparent time-dependent treatment effects can arise if underlying event susceptibility varies between participants, a flaw of using relative measures such as the hazard ratio for casual inference [46]. Effect measures directly estimable from absolute risks such as the ΔRMST and difference in survival proportion retain their causal interpretability regardless of the proportionality of the treatment effect and should be used to supplement reports of relative effect measures [47].

Conclusion

We have compared a range of regression-based approaches allowing for assessment of time-dependent treatment effects and illustrated their potential using a range of endpoints from the ASPREE trial. We recommend these analyses as exploratory and supplementary to the pre-specified primary analyses, aiming to provide enhanced insight and understanding to the mechanisms of any treatment effect, over time and in subgroups of interest. In order to facilitate interpretation, results should be presented using relative and absolute measures of treatment effect in a range of graphical and tabular presentations to provide complementary insights into the timing, magnitude and duration of any treatment effects in a trial.

Funding

The ASPREE trial was supported by a grant (U01AG029824) and the extension study (ASPREE-XT) is supported by a grant (U19AG062682) from the National Institute on Aging and the National Cancer Institute at the National Institutes of Health, by grants (334047 and 1127060) from the National Health and Medical Research Council of Australia, and by Monash University and the Victorian Cancer Agency. KJ was supported in part by an Australian Government Research Training Program (RTP) Stipend and RTP Fee-Offset Scholarship through Federation University Australia and a National Health and Medical Research Council of Australia grant (APP1128222). The contributions of RW and SH was supported by a Centre of Research Excellence grant (1171422) from the National Health and Medical Research Council of Australia to the Australian Trials Methodology Research Network (AusTriM). Dr. Chan is also supported by NCI R35 CA253185 and is a Stuart and Suzanne Steele MGH Research Scholar. The funding bodies had no role in the design of the study, the collection, analysis and interpretation of data or in the writing of the manuscript.

Data sharing statement

The datasets used and/or analysed for this publication are available via the ASPREE Principal Investigators. Requests for data access can be directed to aspree.ams@monash.edu.

Author contributions

KJ was responsible for the concept and design of work, data analysis, visualisation and interpretation and preparation of the first draft of the manuscript. SH and RW contributed to conceptualisation, drafting and critical revision of the manuscript. RLW, SM, AC, AT, AM and JMcN critically reviewed and revised the manuscript. All authors approved the final version of this manuscript.

REFERENCES

1. Cox, D.R., *Regression Models and Life-Tables.* Journal of the Royal Statistical Society. Series B (Methodological), 1972. **34**(2): p. 187-220.
2. Trinquart, L., 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): p. 1813-1819.
3. Rahman, R., et al., *Violations of the proportional hazards assumption in randomized phase III oncology clinical trials.* Journal of Clinical Oncology, 2018. **36**(15_suppl): p. 2543-2543.
4. Altman, D.G., et al., *Review of survival analyses published in cancer journals.* British Journal of Cancer, 1995. **72**(2): p. 511-518.
5. Mathoulin-Pelissier, S., et al., *Survival End Point Reporting in Randomized Cancer Clinical Trials: A Review of Major Journals.* Journal of Clinical Oncology, 2008. **26**(22): p. 3721-3726.
6. Jachno, K., S. Heritier, and R. Wolfe, *Are non-constant rates and non-proportional treatment effects accounted for in the design and analysis of randomised controlled trials? A review of current practice.* BMC Medical Research Methodology, 2019. **19**(1): p. 103.
7. Royston, P. and M.K.B. Parmar, *Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects.* Statistics in Medicine, 2002. **21**(15): p. 2175-2197.
8. Royston, P. and M.K. Parmar, *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): p. 152.
9. Zhao, L., et al., *On the restricted mean survival time curve in survival analysis.* Biometrics, 2016. **72**(1): p. 215-21.
10. Rutherford, M.J., M.J. Crowther, and P.C. Lambert, *The use of restricted cubic splines to approximate complex hazard functions in the analysis of time-to-event data: a simulation study.* Journal of Statistical Computation and Simulation, 2015. **85**(4): p. 777-793.
11. Royston, P. and P.C. Lambert, *Flexible Parametric Survival Analysis Using Stata: Beyond the Cox Model*. 2011: Stata Press.
12. Crowther, M.J. and P.C. Lambert, *Simulating biologically plausible complex survival data.* Statistics in Medicine, 2013. **32**(23): p. 4118-4134.
13. Andersen, P.K. and M. Pohar Perme, *Pseudo-observations in survival analysis.* Statistical Methods in Medical Research, 2010. **19**(1): p. 71-99.
14. Overgaard, M., P.K. Andersen, and E.T. Parner, *Regression Analysis of Censored Data Using Pseudo-observations: An Update.* Stata Journal, 2015. **15**(3): p. 809-821.
15. *International Conference on Harmonisation of Technical Requirements for Pharmaceuticals for Human Use. ICH Harmonised Tripartite Guidelines: Statistical*

Principles for Clinical Trials E9. 1998, London, England: European Medicines Agency 1998.

16. *ICH E9 (R1) addendum on estimands and sensitivity analysis in clinical trials to the guideline on statistical principles for clinical trials*. 2020, Amsterdam, The Netherlands: European Medicines Agency.
17. Kent, D.M., et al., *Risk and treatment effect heterogeneity: re-analysis of individual participant data from 32 large clinical trials.* International Journal of Epidemiology, 2016. **45**(6): p. 2075-2088.
18. Ridker, P.M., *Should Aspirin Be Used for Primary Prevention in the Post-Statin Era?* New England Journal of Medicine, 2018. **379**(16): p. 1572-1574.
19. Hawk, E.T. and K.C. Maresso, *The ASPREE Trial: An Unanticipated Stimulus for Greater Precision in Prevention?* JNCI: Journal of the National Cancer Institute, 2020.
20. Chan, A.T. and J. McNeil, *Aspirin and Cancer Prevention in the Elderly: Where Do We Go From Here?* Gastroenterology, 2019. **156**(3): p. 534-538.
21. Schoenfeld, D., *Partial Residuals for The Proportional Hazards Regression Model.* Biometrika, 1982. **69**(1): p. 239-241.
22. Grambsch, P.M. and T.M. Therneau, *Proportional Hazards Tests and Diagnostics Based on Weighted Residuals.* Biometrika, 1994. **81**(3): p. 515-526.
23. Austin, P.C., *Statistical power to detect violation of the proportional hazards assumption when using the Cox regression model.* Journal of Statistical Computation and Simulation, 2018. **88**(3): p. 533-552.
24. Lambert, P.C. and P. Royston, *Further Development of Flexible Parametric Models for Survival Analysis.* Stata Journal, 2009. **9**: p. 265-290.
25. Bower, H., M.J. Crowther, and P.C. Lambert, *strcs: A Command for Fitting Flexible Parametric Survival Models on the Log-hazard Scale.* Stata Journal, 2016. **16**(4): p. 989-1012.
26. Royston, P., *Estimating the treatment effect in a clinical trial using difference in restricted mean survival time.* Stata Journal, 2015. **15**(4): p. 1098-1117.
27. McNeil, J.J., et al., *Effect of Aspirin on Disability-free Survival in the Healthy Elderly.* New England Journal of Medicine, 2018. **379**(16): p. 1499-1508.
28. McNeil, J.J., et al., *Effect of Aspirin on Cardiovascular Events and Bleeding in the Healthy Elderly.* New England Journal of Medicine, 2018. **379**(16): p. 1509-1518.
29. McNeil, J.J., et al., *Effect of Aspirin on All-Cause Mortality in the Healthy Elderly.* New England Journal of Medicine, 2018. **379**(16): p. 1519-1528.
30. Zheng, S.L. and A.J. Roddick, *Association of Aspirin Use for Primary Prevention With Cardiovascular Events and Bleeding Events: A Systematic Review and Meta-analysis.* JAMA, 2019. **321**(3): p. 277-287.
31. Rothwell, P.M., et al., *Short-term effects of daily aspirin on cancer incidence, mortality, and non-vascular death: analysis of the time course of risks and benefits in 51 randomised controlled trials.* The Lancet, 2012. **379**(9826): p. 1602-1612.
32. McNeil, J.J., et al., *Effect of aspirin on cancer incidence and mortality in older adults.* Journal of the National Cancer Institute, 2020. **113**(3): p. 258-265.
33. Mahady, S.E., et al., *Major GI bleeding in older persons using aspirin: incidence and risk factors in the ASPREE randomised controlled trial.* Gut, 2020: p. gutjnl-2020-321585.
34. Sit, T., et al., *Design and analysis of clinical trials in the presence of delayed treatment effect.* Statistics in Medicine, 2016. **35**(11): p. 1774-1779.

35. Xu, Z., et al., *Designing cancer immunotherapy trials with random treatment time-lag effect.* Statistics in Medicine, 2018. **37**(30): p. 4589-4609.

36. Ye, T. and M. Yu, *A robust approach to sample size calculation in cancer immunotherapy trials with delayed treatment effect.* Biometrics, 2018. **74**(4): p. 1292-1300.

37. Wu, J. and J. Wei, *Cancer immunotherapy trial design with delayed treatment effect.* Pharmaceutical Statistics, 2019. **1**(12).

38. Jiménez, J.L., V. Stalbovskaya, and B. Jones, *Properties of the weighted log-rank test in the design of confirmatory studies with delayed effects.* Pharmaceutical Statistics, 2019. **18**(3): p. 287-303.

39. Ristl, R., et al., *Delayed treatment effects, treatment switching and heterogeneous patient populations: How to design and analyze RCTs in oncology.* Pharmaceutical Statistics, 2020. **20**: p. 129-145.

40. Lee, S.-H., *On the versatility of the combination of the weighted log-rank statistics.* Computational Statistics & Data Analysis, 2007. **51**(12): p. 6557-6564.

41. Yang, S. and R.L. Prentice, *Assessing potentially time-dependent treatment effect from clinical trials and observational studies for survival data, with applications to the Women's Health Initiative combined hormone therapy trial.* Statistics in Medicine, 2015. **34**(11): p. 1801-1817.

42. Royston, P. and M.K. Parmar, *Augmenting the logrank test in the design of clinical trials in which non-proportional hazards of the treatment effect may be anticipated.* BMC Medical Research Methodology, 2016. **16**(1): p. 16.

43. Karrison, T.G., *Versatile Tests for Comparing Survival Curves Based on Weighted Log-rank Statistics.* Stata Journal, 2016. **16**(3): p. 678-690.

44. Royston, P., *A combined test for a generalized treatment effect in clinical trials with a time-to-event outcome.* Stata Journal, 2017. **17**(2): p. 405-421.

45. Magirr, D. and C.-F. Burman, *Modestly weighted logrank tests.* Statistics in Medicine, 2019. **38**(20): p. 3782-3790.

46. Hernán, M.A., *The Hazards of Hazard Ratios.* Epidemiology (Cambridge, Mass.), 2010. **21**(1): p. 13-15.

47. Bartlett, J.W., et al., *The Hazards of Period Specific and Weighted Hazard Ratios.* Statistics in Biopharmaceutical Research, 2020. **12**(4): p. 518-519.

Chapter 5

5.1 Manuscript introduction: Complementing the Kaplan-Meier plot to enable assessment of treatment effect consistency with proportional hazards

In Chapter 2 it was found that Kaplan-Meier plots have been used almost exclusively to visually present the survival experience of different treatment groups over time and earlier in the thesis it was noted that these plots do not provide for an assessment of the treatment effect which is of primary interest to trialists. In this chapter a complementary plot of treatment effect measure is proposed to accompany Kaplan-Meier plots to provide for direct assessment of treatment effect consistency with proportional hazards.

Previous reviews and guidelines for the presentation of survival curve estimates have provided a series of recommendations based on graphical principles that are applicable to any plots, and recommendations that are specific to survival curve plots. These recommendations were collated and harmonised and used to assess the plots from trials in the review from Chapter 2 for adherence. Through presentation of a variety of reconstructed individual patient datasets from previously published trials, we illustrate the utility of our recommended composite presentation of a Kaplan-Meier survival curve and a treatment effect plot.

In the next section is presented a manuscript submitted to the journal *Trials*. Three supplementary files for the manuscript are available as Appendices F, G and H of this thesis. These provide the citations references for the trials used in the review (Appendix F), supplementary figures for presentation (Appendix G) and example code to create the complementary plots (Appendix H).

RESEARCH

Complementing the Kaplan-Meier plot to enable assessment of treatment effect consistency with proportional hazards

Kim M. Jachno*, Stephane Heritier and Rory Wolfe

*Correspondence:
kim.jachno@monash.edu
Public Health and Preventive Medicine, Monash University, Melbourne, Australia
Full list of author information is available at the end of the article

Abstract

Background: Kaplan-Meier plots are typically used to present the results from clinical trials with time-to-event outcomes. They display the survival experience over time in different treatment arms. However, when used to assess for treatment effect there can be a disconnect between visual impression and the statistical evidence. The hazard ratio from a Cox proportional hazards model provides a summary treatment effect measure. Increasingly, the key assumption of proportionality – or time-fixed effect - that underpins this model has been called into question, potentially casting doubt on the presentation of a fixed-magnitude treatment effect as the key inferential finding of a trial.

Methods: We investigated how clinical trials with time-to-event outcomes present results graphically utilizing our review of all original reports from four medical journals during the first half of 2017. We assessed the published Kaplan-Meier plots against a series of general graphical and survival curve-specific recommendations based on reviews and researcher guidelines. We used reconstructed individual patient datasets from published trials exhibiting nonproportionality to illustrate our recommended complementary treatment effect plots.

Discussion: We reviewed 65 trials that presented a Kaplan-Meier plot to present primary outcome results. Adherence to all general graphical recommendations and most survival curve-specific recommendations was excellent with the depiction of the level of uncertainty around survival curves the main area for improvement identified. We illustrated our recommendations for presenting combinations of survival curves and treatment effect measures over time using selected trials showing different levels of proportionality and baseline event rates.

Conclusions: There is still scope to improve the presentation of Kaplan-Meier plots, especially for depicting the uncertainty associated with survival curve estimates over time. Further, we present a complementary plot to the Kaplan-Meier survival curves that enables more intuitive insight into the dynamic nature of any treatment group differences over time. Visual presentation is effective in conveying the information of primary interest on the treatment effect – be it a point difference in time, a ratio or cumulative summary of change over time – and in this respect the proposed treatment effect plot complements and enhances the value of the Kaplan-Meier plot.

Keywords: survival analysis; Kaplan-Meier plots; nonproportionality; clinical trial; treatment effects

Introduction

Clinical trials with time-to-event outcomes almost invariably present Kaplan-Meier (KM) estimated survival probabilities over time as the graphical means to present results. The evidence for any difference in these survival curves between treatment groups is typically provided by an accompanying logrank test or estimation of a hazard ratio (HR) typically from a Cox proportional hazard (PH) model [1, 2, 3] A major strength of KM plots is that they appear relatively intuitive to read, easily providing information about the survival experience of the groups presented, and a visual indication of the difference between the survival proportions and quantiles of survival time over time. However, the information to detect survival curve differences comes from the number of events occurring in each group relative to the number of participants available. This can create a disconnect between the visual impression and the statistical evidence. The survival curves are often visually closest together at earlier times and if this is when more events occurred, then small differences between curves may be estimated from a large number of events and thus maybe determined to be statistically significant. Conversely, at latter times survival may be estimated from fewer events and it is possible that a perceived large difference between survival proportions could be based on relatively few event occurrences, and hence be determined to be not statistically significant.

Pointwise estimates of uncertainty around each survival curve are useful additions to survival plots. Clear separation of the 95% confidence intervals (CIs) for the series of point-wise estimates of each survival curve provides a visual confirmation of treatment differences. However, overlapping 95% CIs are possible in the presence of a significant treatment difference, so can't be used as a visual confirmation of lack of treatment effect. The accumulation of censored observations - patients dropping out of the trial without having experienced the event of interest, or having not yet been followed up for their planned observation period - is an added complication which ensures that visual assessments of any treatment differences from a plot of survival curves may be suggestive but usually cannot be confirmatory.

The Cox PH model is the most widely used regression-based analysis method for estimation of a relative treatment effect in the form of a HR although alternative regression-based approaches can be utilised including simple parametric models such as the Weibull [4, 5, 6]. There has been a trend away from the sole reliance on tests of statistical significance such as the logrank test towards regression-based methods that enable estimation of the magnitude of treatment effects with an accompanying estimate of uncertainty around a point estimate [7]. More flexible modelling approaches than semiparametric Cox model or simple parametric models can also be used and have the advantage that the baseline hazard is modelled explicitly – not so the case with the Cox model - and that this modelling is free of any strict distributional assumptions – unlike the simple parametric models [8]. Flexible models enable estimation of multiple measures of treatment effect from the same modelling framework. As well as relative effect measures such as the HR, absolute effect measures that can vary over time such as difference in survival proportion and difference in risks can be obtained from the flexible modelling approach, and it is also possible to estimate distribution-free summary measures of treatment effect such as the difference in mean survival time (MST) restricted to a specified duration of time from the start of treatment [9, 10].

When trial results are summarised in the form of a HR, a key outcome for clinical researchers is to determine whether a single HR captures the effect of treatment with a reasonable degree of consistency across the entire duration of the trial. Using the shape of the survival curves in a KM plot to attempt some assessment of the proportionality or otherwise of the underlying hazards - or instantaneous risks - between treatment arms of a trial is not straight forward and caution has been advised [11].The easiest patterns that can be inferred from KM plots are survival curves showing no apparent difference, and those showing a steady divergence between treatments arms over time. This latter case is expected when the treatment difference is proportional over time on the hazard scale. However, many more complex patterns are observed in reality, such as large divergences between treatment arms early in the survival experience gradually rejoining, or initially similar curves diverging later on. These indicate some form of non-proportionality where the treatment effect is varying over time. Clear indications of non-proportionality can be observed when survival curves cross. To enable assessment of the stability of the HR estimates over time, additional analytical tests or visual presentations should be provided along with KM plots. There have long been concerns about the use of HRs and the testing and reporting of the crucial assumption of non-proportionality in reports of treatment effect in clinical trials [5, 12, 13, 14, 15].

The goals of this paper are (1) to provide a snapshot of adherence to the recommendations of good KM plots, and (2) propose an improvement to the reporting of clinical trials with time-to-event outcomes by visual presentations of treatment effect estimation over time as an accompaniment to the standard KM plots.

Methods

Review of graphical presentation: We assessed the presentation of Kaplan-Meier plots as part of a larger review of practice in designing, analysing and presenting time-to-event outcomes in clinical trials [5]. We reviewed all original reports in four high impact medical journals, the *New England Journal of Medicine*, the *British Medical Journal*, *The Lancet* and the *Journal of Clinical Oncology* during the first six months of 2017. We identified clinical trials for which the primary outcome involved time-to-event analysis. The usage of KM plots as the graphical presentation method was recorded and the quality of the plots against recommendations encompassing general graph components and KM-specific graphing components was assessed. Based on recommendations from previous reviews of the display and interpretation of KM plots, we defined seven recommendations for KM plots to be judged against in this review [11, 12, 16].

General graphing components informed three recommendations assessing whether (1) graphical elements were clear and plots had enough information to be self-explanatory, (2) the use of meaningful time intervals within a time period for which a reasonable proportion of participants had been followed up was used on the horizontal axis, so differences between arms at this time are not “unduly” influenced by chance events, and (3) whether an appropriate vertical axis was chosen so as to convey any treatment difference between arms of the trial but still fill the visual space informatively. This was usually a choice between plotting the KM survival curve displaying the probability of remaining event free over time or the cumulative incidence curve with cumulative incidence estimated using a KM approach as the complement one of survival in the absence of competing events.

Survival curve specific components were assessed as (1) whether step functions were used to join lines acknowledging the event-driven estimation process, (2) an indication of the number of participants at risk at selected times, (3) an indication of event times or censoring events during the trial, and (4) displaying some measure of statistical uncertainty for each of the treatment groups either through the use of 95% CIs at regularly spaced time points or shading to indicate the same. This does not directly display the uncertainty of the treatment effect measure which is usually of primary interest but the absence (or presence) of overlapping 95% CIs can be taken as a rough guide to the significance (or lack thereof) of the treatment effect difference.

Proposal to improve the visual representation of treatment effects:
In order to improve the visual informationregarding treatment effects in clinical trials with time-to-event outcomes, we propose a graphical estimate of treatment effect over time to accompany the KM plot. We demonstrate the utility of this proposal using trials identified from previous reviews exhibiting varying amounts of non-proportionality and separation of the individual survival curves. To this end, we reconstructed individual participant data (IPD) for each treatment arm from published KM curves of these trials using the DigitizeIt graphical digitisation software to create time and survival probability coordinates from the curves. Where possible, we extracted the number of patients at risk at selected times, and total number of events in each arm of the trial. We estimated individual times to event or censoring using the ipdfc Stata command which is based on an algorithm developed originally in R [17, 18].

For each reconstructed dataset, estimates of treatment effect were obtained from the Cox PH model and from parametric modelling of the baseline log hazard using restricted cubic splines [19]. Such flexible parametric models (FPMs) allow for treatment effect estimation under a PH assumption equivalent to a Cox PH model, and can also be extended to allow for time-dependent treatment effects by defining a different set of spline variables used in the estimation of the treatment covariate over time. In order to undertake analysis as might be pre-specified in a statistical analysis plan, here FPM models were fit assigning four degrees of freedom for the baseline hazard function of each reconstructed dataset, with two degrees of freedom used in models allowing for possible time-dependence (TD) of treatment effect. This model specification is flexible enough to fit likely forms of non-constant event rates and non-proportionality encountered in practice whilst maintaining as parsimonious a model specification as possible. The agreement of published curves with the extracted curves was assessed visually and by comparison of the published treatment effect estimates with estimates of treatment effect obtained from the IPD dataset.

The HR and the differences in restricted MST (ΔRMST) were calculated as treatment effect measures and visualization of their evolution over time was constructed. We compared the reported treatment effects obtained from the Cox PH model and an FPM under an assumption of PH. We used the Grambsch-Therneau PH test to assess for evidence of non-proportionality [20]. We also used a likelihood ratio test of model fit comparing the FPM PH model against the FPM TD model as an additional measure of time-dependent treatment effect.

Results

From our review of trials with time-to-event primary outcomes, a KM survival plot or its reciprocal, a cumulative incidence plot, was used in 65 of 66 trials as the graphical presentation of the primary outcome results. Table 1 presents their adherence to recommendations for plotting survival curves.

Table 1 Adherence to graphing recommendations for plots of survival curve analysis

Characteristic	Recommendation	Review of trials n	 (%)
General graphing components	Graphical elements clear, explanation within plot, legend or figure key	65	(100)
	Use of meaningful time intervals, clear indication of time intervals (x-axis)	65	(100)
	Appropriate selection of survival curve or cumulative incidence (y-axis)	64	(98)
Survival curve specific components	Step functions to join survival proportion or cumulative incidence estimates	65	(100)
	Indication of number of patients at risk during trial	62	(95)
	Indication of number of events or censoring during trial	28	(43)
	Depiction of uncertainty through shading or use of confidence intervals	1	(1.5)

Adherence to the recommendations for general graphing components was excellent with all plots showing required graphical elements and use of meaningful time intervals. Similarly, the selection of the appropriate scale and plot type for the y-axis was followed in most of the assessed plots (64/65; 98%). Adherence to two of the survival curve specific recommendations was also high with step functions used to plot the curves in all cases and an indication of the number of patients at risk at selected times during the trial provided in the majority of cases (62/65; 95%). Providing an indication of censoring during the trial occurred in just under half of the plots (28/65; 48%). Censoring events were depicted through the use of markers on the plot. Cumulative event numbers at intervals throughout the trial were usually provided in a risk table. Only one trial provided some depiction of uncertainty around the survival curve plots and this was through the use of 95% CIs at yearly intervals. Full citation details for all trials and review results are available in Additional File 1.

We now present examples of current recommendations for KM plots and our proposed accompanying plot of treatment effect over time plot, the time-dependent hazard ratio, HR(t) plot. We selected five trials to demonstrate the visual information that can be obtained about treatment effect magnitude and direction using an HR(t) plot to accompany a KM plot. The selected trials were previously used by other researchers to demonstrate clear examples of non-proportionality of the treatment effect measure supplemented by one trial from our review. Details of the selected trials are presented in Table 2 – they showed differing event rate patterns and amounts of survival curve separation. In these examples, we have focused on the HR, but acknowledge that this is not always the treatment effect measure used to convey trial outcomes.

Table 2 Information about the trials used for reconstructed IPD results

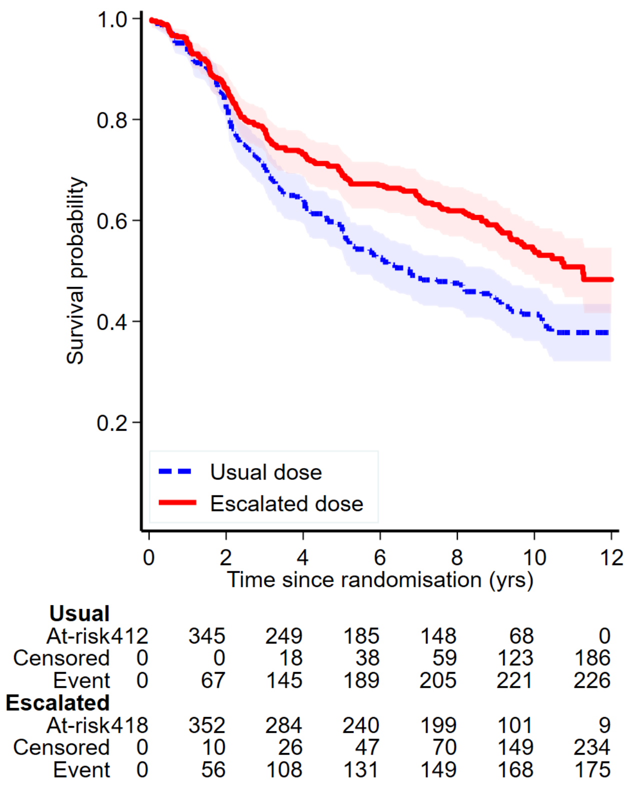
No.	Trials	Treatment effect results
1	**RT01:** compared an escalated dosage of radiotherapy to the usual dosage of radiotherapy in patients with localised prostate cancer over a 12-year time frame [21]. Dataset reconstructed from Figure 2C showing a clear difference in biochemical progression free survival with event numbers to 12 years.	**Reported results:** HR = 0.69, 95% CI 0.56 to 0.84, p=0.0003 **Reconstructed IPD results:** HR = 0.69, 95% CI 0.56 to 0.84, p=0.0002 Assessment for non-PH: p=0.95 Test of model fit (TD versus PH): p=0.41
2	**Head:** compared radiotherapy (RT) to radiotherapy plus a weekly dose of anti-epidermal growth factor receptor monoclonal antibody (RT+mAb) in patients with head and neck cancers over 5 years [22]. Dataset reconstructed from Figure 1 showing good separation of KM estimates of time free from locoregional progression or death to 60 months.	**Reported results:** HR = 0.68, 95% CI 0.52 to 0.89, p=0.005 **Reconstructed IPD results:** HR = 0.72, 95% CI 0.56 to 0.83, p=0.008 Assessment for non-PH: p=0.85 Test of model fit (TD versus PH): p=0.90
3	**RTOG:** prostate cancer therapy trial compared radiotherapy to radiotherapy plus anti-androgen therapy (RT+antiA) with a median of 12 years follow up [23]. Dataset reconstructed from Figure 2A showing some separation of KM curves for overall survival gradually increasing over the duration of the follow up 15 years.	**Reported results:** HR = 0.77, 95% CI 0.59 to 0.99, p=0.04 **Reconstructed IPD results:** HR = 0.76, 95% CI 0.59 to 0.89, p=0.04 Assessment for non-PH: p=0.81 Test of model fit (TD versus PH): p=0.78
4	**ICON7:** ovarian cancer trial compared chemotherapy to chemotherapy plus an anti-growth factor monoclonal antibody (CT+mAb) with a median follow up of 28 months [24]. Reported evidence of non-proportionality (p<0.001) Dataset reconstructed from Figure 2A clearly showing crossing survival curves. Acknowledged lack of meaningful interpretation for a HR in presence of non-proportionality so also provided alternative absolute effect measures including difference in survival proportion at 12mths and ΔRMST at 36mths.	**Reported results:** HR = 0.81, 95% CI 0.70 to 0.94, p=0.004 ΔRMST 36m: 1.5m, 95%CI 0.1m to 2.9m **Reconstructed IPD results:** HR = 0.83, 95% CI 0.72 to 0.94, p=0.006 ΔRMST 30m: 1.3, 95%CI 0.4m to 2.3m Assessment for non-PH: p<0.001 Test of model fit (TD versus PH): p<0.001
5	**EUROPA:** compared rate of cardiovascular events between patients treated with angiotensin-converting enzyme (ACE) inhibitor or placebo with a mean follow up 4.2 years [25]. Dataset reconstructed from Figure 2, cumulative incidence curves showing increasing separation of curves after 1.5 years.	**Reported results:** HR = 0.80, 95% CI 0.71 to 0.91, p=0.0003 **Reconstructed IPD results:** HR = 0.81, 95% CI 0.72 to 0.91, p=0.0006 Assessment for non-PH: p=0.11 Test of model fit (TD versus PH):p=0.13

HR: hazard ratio; ΔRMST: difference in restricted mean survival time
PH: proportional hazards; TD: time-dependent

Example presentation of a current recommended Kaplan-Meier plot with extended risk table and 95% confidence interval shading

Figure 1 shows the results from the RT01 trial with an extended risk table beneath the KM plot and 95% CI shading for each survival curve added. These additions have been recommended to improve (i) the depiction of the state of participants over time, and (ii) uncertainty over time around the survival curve estimates [26]. Figure 1 is thus an example of the application of current recommendations for improving graphical presentation of results from trials with time-to-event outcomes.

From Figure 1, a sustained risk reduction due to the escalated dose over time is apparent from the clear separation observed for most of the follow up period. There



	0	2	4	6	8	10	12
Usual							
At-risk	412	345	249	185	148	68	0
Censored	0	0	18	38	59	123	186
Event	0	67	145	189	205	221	226
Escalated							
At-risk	418	352	284	240	199	101	9
Censored	0	10	26	47	70	149	234
Event	0	56	108	131	149	168	175

Figure 1 Presentation of the RT01 trial following recommendations to improve information conveyed by Kaplan-Meier plots. An area of uncertainty around each Kaplan-Meier survival curve from pointwise 95% CI estimates is shown along with an extended risk table with numbers for the number at risk, censored and having experienced an event at periodic intervals.

appears to be reasonable numbers at risk and event occurrence for most of the trial as shown by the width of the area of uncertainty around the survival curves. For the RT01 trial it was reported that the assumption of proportionality was assessed although no results were provided in the main report. From the reconstructed dataset, we also found no evidence to suggest any non-proportionality (test of PH p=0.95; LR test of TD v PH model fit p=0.41).

Using HR(t) plots to visually assess for time-dependence of treatment effects

The next set of results demonstrate the use of the HR(t) plot to visually assess the magnitude and timing of treatment effect s over the trial duration. This is shown firstly in trials for which the proportional hazards assumption seems reasonable and secondly in trials for which a time-dependence of treatment effect may exist.

The HR(t) plot in trials with no evidence against the proportional hazards assumption subsubsection

The reconstructed primary outcome results for three trials using a KM plot with an accompanying HR(t) plot are depicted in Figure 2. These three trials had similar average reduction in risks of 31%, 28% and 24% risk reduction for the RT01, Head and RTOG trial datasets respectively, with similar 95% CI widths (Table 2). However, the control groups in the trials had different hazard curve shapes. For a visual assessment of model fit, the KM curves are similar to the fitted curves from the FPM TD model as displayed in the top panels of Figure 2. Further evidence that the PH assumption is reasonable is provided by the HR(t) displaying approximately constant point estimates over time.

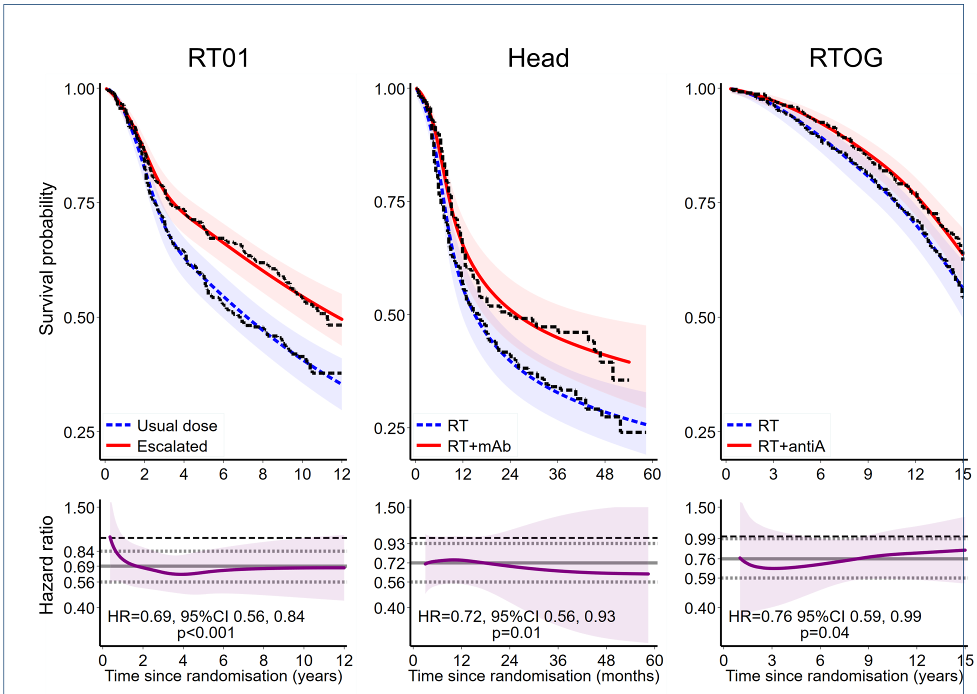


Figure 2 Kaplan-Meier plots with accompanying HR(t) plots for the RT01, Head and RTOG trials. Kaplan-Meier plots (upper panels) showing the estimated Kaplan-Meier curves (dashed black lines) overlaid by the estimated time-dependent curves for each treatment group (dashed blue and solid red lines). Accompanying HR(t) plots (lower panels) showing (a) the average treatment effect estimates (solid grey lines) with flanking 95% CI band (dotted grey lines) estimated from the Cox PH model, (b) the null effect corresponding to HR=1 (dashed black lines) and (c) the estimated time-dependent treatment effect (solid purple lines) and its 95% CI area (purple shading) from the flexible parametric model.

Design elements of the HR(t) plots that we propose are to include the text summarising the treatment effect measure, which is often located in the KM plot, and include the word "average" in the descriptor of this effect measure. These changes are designed to make the HR(t) plot the source of explicit quantification of the treatment effect measure, and to emphasise that an underlying assumption of proportionality is used to obtain a summary, the "average" value, the reasonableness of which can be assessed visually from the HR(t) plot. We construct the HR(t) plot using the summary HR point estimate and 95% CI from the PH model as horizontal lines overlaid with the estimated HRs from the TD model with 95% CI areas. A final design choice is to left-truncate the plots at a time point corresponding to approximately the fifth centile of event times. This avoided undue visual influence of early estimates of the HR based on few events with associated large uncertainty.

From the accompanying HR(t) plots in Figure 2, it can be visually inferred that in each of the three trials the respective average effect was an appropriate summary of the treatment effect over the entire duration of the trial. In each HR(t) panel of Figure 2 the estimated time-dependent HR remains close to the average HR horizontal line and the uncertainty around the HR estimates fit mostly within the average HR's 95% CI band (the dotted lines). This conclusion is confirmed by analytical results suggesting no evidence of non-proportionality in any of the three trials. Additional insight into the precision of the treatment effect estimate over the duration of the trial is conveyed by the shaded area in the lower panel.

The HR(t) plot in trials with possible time-dependence of treatment effect

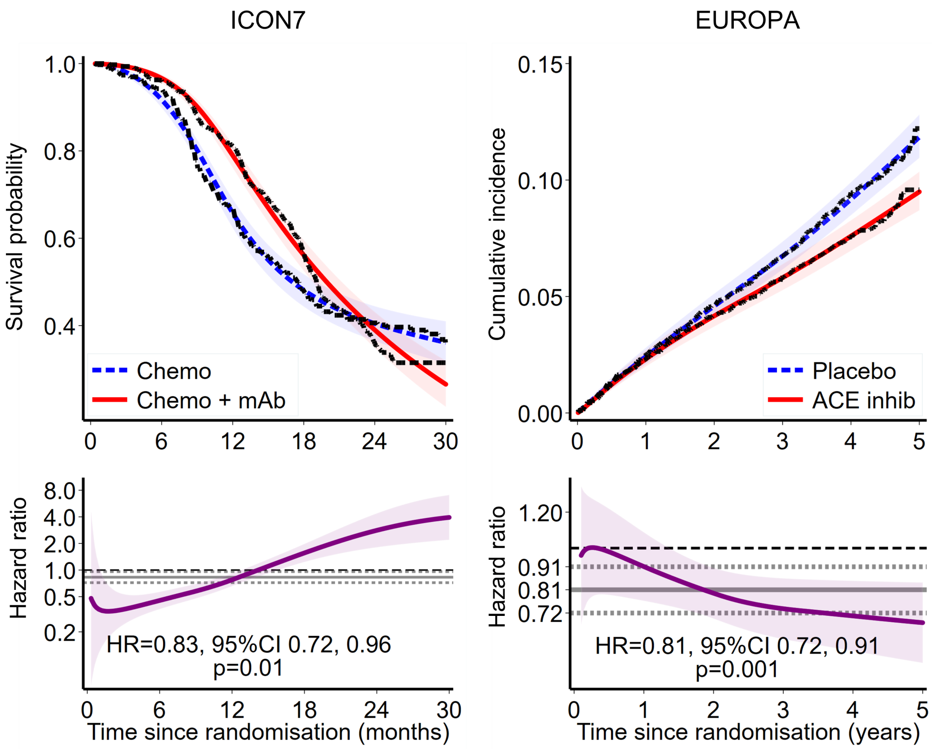


Figure 3 Kaplan-Meier plots with accompanying HR(t) plots for the ICON7 and EUROPA trials. Kaplan-Meier plots (upper panels) showing the estimated Kaplan-Meier curves (dashed black lines) overlaid by the estimated time-dependent curves for each treatment group (dashed blue and solid red lines). Accompanying HR(t) plots (lower panels) showing (a) the average treatment effect estimates (solid grey lines) with flanking 95% CI band (dotted grey lines) estimated from the Cox PH model, (b) the null effect corresponding to HR=1 (dashed black lines) and (c) the estimated time-dependent treatment effect (solid purple lines) and its 95% CI area (purple shading) from the flexible parametric model.

When the treatment effect is time-dependent, assessment of the magnitude and timing of any possible resulting benefit can be obtained from the HR(t) plot. Figure 3 presents the KM and HR(t) plots for two reconstructed primary outcome results from two trials showing different levels of non-proportionality for the treatment effect on the primary outcome. The crossing survival curves from the ICON7 trial showed clear evidence of non-proportionality with a reported test of PH having p<0.001. Despite acknowledging the lack of meaningful interpretation for an average HR in the presence of non-proportionality, the authors reported the treatment effect of the trial as a HR =0.81 95% CI 0,70 to 0.94, p=0.004. From the HR(t) plot for the ICON7 trial in Figure 3, the lack of proportionality is apparent with a beneficial effect of treatment early in the trial and a harmful effect of treatment later in the trial becoming apparent sometime shortly after the first year. This strong time-dependence of the estimated treatment effect is not adequately described by the average HR estimate nor even is it hinted at by the narrow bounds of the 95% CI for the average HR. Sensibly, the results of the trial were published with two alternative measures provided: a maximal improvement in progression-free survival proportion at 12 months of 15.1% (95% CI 10.7% to 19.5%) and the ΔRMST at 36 months being 1.5 months (95% CI 0.1 to 2.9).

A nuanced example of non-proportionality was presented in the EUROPA trial. In our reconstructed dataset we did not find strong evidence of non-proportionality from the test of PH (p=0.11) or the likelihood ratio test of TD versus PH model fit (p=0.13). From the accompanying HR(t) plot for the EUROPA trial in Figure 3, the increased risk reduction after the initial year of minimal treatment benefit is apparent. The trend of increasing risk reduction is sustained such that by the end of the follow up period, the estimated treatment effect was approximately a 30% reduction.

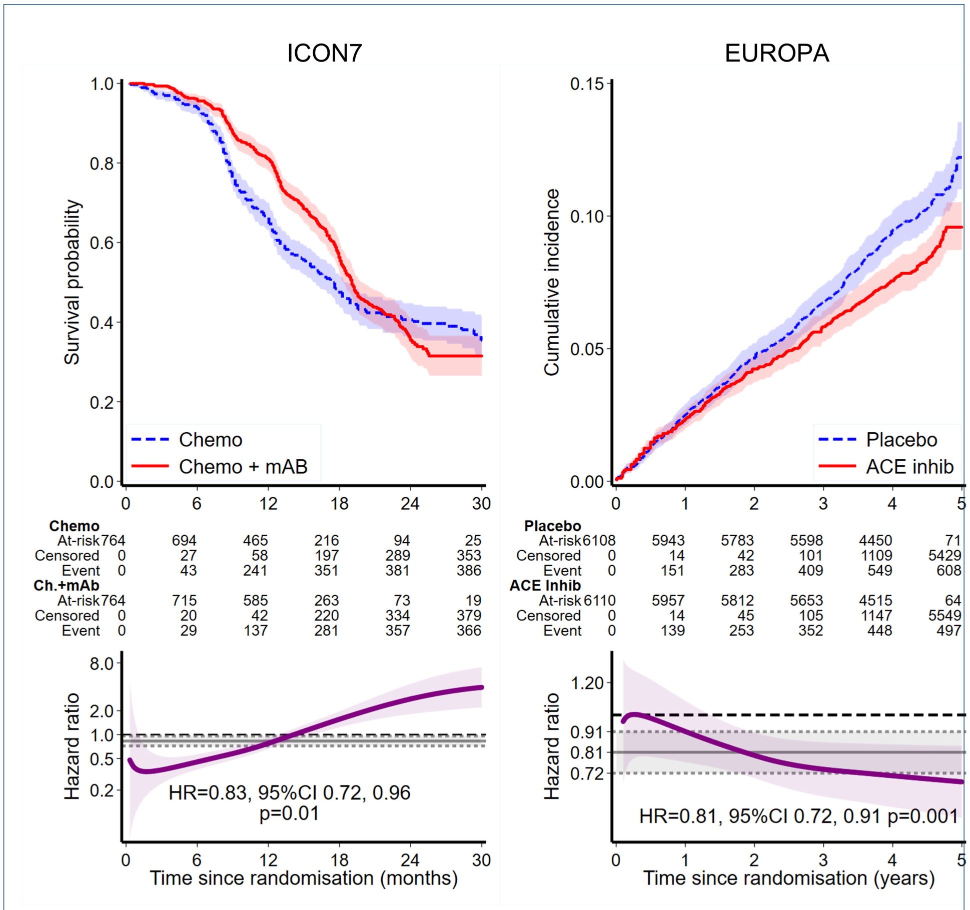
Recommended presentation of Kaplan-Meier plots augmented by time-dependent treatment effect visualisations

In Figure 4, the trials from Figure 3 are re-presented in a format that we propose as an optimal template for trials in general. To facilitate interpretability, we have maintained the focus on the KM plot with the risk table presentation underneath. We include shading to indicate the 95% CI area around the survival curves and the extended risk table information. The HR(t) plot is placed below the risk table with alignment of time on the X-axes of the HR(t) and KM plots.

As a supplementary presentation, an expanded set of plots for all trials are presented in Additional File 2. Supplementary Figures S1 - S5 are four-panel presentations of the survival curves (S(t); panel A) and hazard rates (h(t); panel B) for treatment groups over time. Two estimates of treatment effect, the ΔRMST (panel C) or the HR (panel D) are shown below the by-treatment arm plots. The proportionality of the hazards can be assessed visually in panel B, and also the shape of the underlying baseline hazard which could be otherwise extremely difficult to infer from the shape of the survival curves. Dotted lines in panels C and D indicate the null treatment effect. Supplementary Figure S6 is the ICON7 trial presented with the two alternative treatment effect estimates reported in the original trial findings – the ΔRMST reported as a cumulative estimate to 30 months, and the difference in survival curve proportions, ($\Delta S(t)$) reported as a point estimate at 12 months.

Discussion

We have demonstrated the utility of presenting a plot of treatment effect measure to accompany the Kaplan-Meier plots that are almost universally used to convey



	0	6	12	18	24	30
Chemo						
At-risk	764	694	465	216	94	25
Censored	0	27	58	197	289	353
Event	0	43	241	351	381	386
Ch.+mAb						
At-risk	764	715	585	263	73	19
Censored	0	20	42	220	334	379
Event	0	29	137	281	357	366

	0	1	2	3	4	5
Placebo						
At-risk	6108	5943	5783	5598	4450	71
Censored	0	14	42	101	1109	5429
Event	0	151	283	409	549	608
ACE Inhib						
At-risk	6110	5957	5812	5653	4515	64
Censored	0	14	45	105	1147	5549
Event	0	139	253	352	448	497

Figure 4 Recommended presentation for the ICON7 and EUROPA trials. Kaplan-Meier plots (upper panels) showing the estimated Kaplan-Meier curves for each treatment group (dashed blue and solid red lines). Risk table giving the number of participants at risk, number censored and the cumulative number of events at periodic intervals. Accompanying HR(t) plots (lower panels) showing (a) the average treatment effect estimates (solid grey lines) with flanking 95% CI band (dotted grey lines) estimated from the Cox PH model, (b) the null effect corresponding to HR=1 (dashed black lines) and (c) the estimated time-dependent treatment effect (solid purple lines) and its 95% CI area (purple shading) from the flexible parametric model.

the results of clinical trials with time-to-event outcomes [16]. Plots can be more noticeable than text or tables and convey information about treatment effects and dynamic changes with more immediacy. Measuring the effect of treatment is the primary goal of most clinical trials and potential time-variations in this effect will be of interest to trialists, patients and healthcare providers. Visualisation of the summary treatment effect measure should be part of any trial report. The display of statistical uncertainty and the assessment of assumptions underlying the model are also important aspects of appropriately reporting the findings from a trial, and ideally would be key components of any plots aiming to present trial results visually.

Our review found there has been improvement in the presentation of KM plots with excellent adherence to the recommendations of good graphing practice and some of the KM-specific plot recommendations. However, recommendations to provide for some estimation of uncertainty via point-wise confidence intervals at selected times, or shading of the area depicting the 95% confidence interval area for survival plots were rarely implemented. These findings concur with another review of KM plots [26] which found similar results with recommendations regarding displaying uncertainty rarely implemented whilst the more general graph recommendations and other KM plot recommendations were well implemented. In contrast, earlier review findings found generally poor implementation of most of the recommendations [11, 12, 16].

Large trials, trials with long term follow up and trials with novel mechanisms of action have become commonplace and mean that non-proportionality is being detected more frequently [27, 28, 29]. It has been argued that the default expectation should be that the HR will vary over the follow-up period [30]. The reported summary HR from a Cox PH model should be interpreted as a weighted average of the true time-varying HRs over the entire follow up. Statistical testing of PH provides a quantifiable rationale to support clear evidence of non-proportionality but may miss clinically important deviations from PH in small studies while detecting clinically unimportant deviations from PH in large trials [31]. Visualisation of the degree of non-proportionality in the key treatment effect of a trial gives impetus to explore possible reasons for time-dependence; encouraging further analysis to ascribe, if possible, any non-proportionality to time-dependent treatment effects, subgroup heterogeneity or allowance for unobserved ‘frailty’ factors. Recommendations for the use of more interpretable estimands and clearer reporting of potentially dynamic treatment effects are supported by the causal inference literature and regulatory guidelines [32, 33].

We recommend that the KM curve be accompanied by a plot of treatment effect over time. This treatment effect plot will most likely be in the form of a HR, but could also be a difference in survival proportion, difference in RMST, time ratio or other estimand [27, 34, 35]. In our proposed composite plots, we have incorporated the KMunicate proposals aimed at improving the visual aspect of uncertainty around the within-arm survival comparisons in KM plots by including an extended risk table and the use of shading to indicate the level of uncertainty around individual survival curve estimates [26]. However, the amount of overlap between confidence intervals around two survival curves can only be used as a guide to assess the significance or otherwise of the treatment effect. Direct plotting of the treatment effect measure with its own estimate of uncertainty is a more definitive means of assessing the strength of evidence of any reported finding.

Our recommended plot layout has been informed by the principles of good graphing practice outlined in seminal data visualisation sources [36, 37, 38]. We considered how these principles of data visualisation should be incorporated to enhance readers’ understanding of the alternative displays of the same trial outcomes. Examples include the vertical alignment of the treatment group experience in the KM plots and the difference in treatment effect in the HR(t) plots emphasising that the same information from the trial was being presented, and use of a log scale for the HR(t) plots to provide spatial symmetry to the reference band of treatment effect uncertainty. Code for creating an example graph using Stata is available in Additional File 3. Future research that examines readers’ understanding of the composite displays of potentially different treatment effect measures could help in refining recommendations for presentation.

Conclusions

We believe presentation of treatment effect estimation complements the Kaplan-Meier plot and will improve the reporting of trials with time-to-event outcomes. By visually highlighting the presence of any non-proportionality of treatment effect with a clear display of the associated uncertainty, readers can ascertain whether a summary fixed-magnitude treatment effect adequately captures the treatment effect findings of a trial. Regression-based methods which model the baseline hazard and allow for both relative and absolute time-dependent treatment effect measures to be calculated directly are ideal for this purpose.

Appendix

Acknowledgements

KJ was supported in part by an Australian Government Research Training Program (RTP) Stipend and RTP Fee-Offset Scholarship through Federation University Australia and a National Health and Medical Research Council of Australia grant (APP1128222).

Funding

Not applicable

Abbreviations

KM: Kaplan-Meier; HR: Hazard ratio; PH: proportional hazards; CI: confidence interval; MST: mean survival time; FPM: flexible parametric models; TD: time-dependence; △RMST: difference in restricted mean survival time; HR(t): time-dependent hazard ratio; △S(t): difference in survival curve probability.

Availability of data and materials

All data generated or analysed during this study are included in this published article and its supplementary information files.

Ethics approval and consent to participate

Not applicable

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

KJ conceived the study, reviewed data visualisation resources, extracted the data used in the review and generated the computer code. KJ wrote the first draft of the manuscript. KJ, SH and RW all contributed to revisions of the manuscript and take public responsibility for its content.

Author details

Public Health and Preventive Medicine, Monash University, Melbourne, Australia.

References

1. Kaplan, E.L., Meier, P.: Nonparametric estimation from incomplete observations. Journal of the American Statistical Association **53**(282), 457–481 (1958)
2. Mantel, N.: Evaluation of survival data and two new rank order statistics arising in its consideration. Cancer Chemotherapy Reports **50**, 163–170 (1966)
3. Cox, D.R.: Regression models and life-tables. Journal of the Royal Statistical Society. Series B (Methodological) **34**(2), 187–220 (1972)
4. Royston, P., Parmar, M.K.B.: An approach to trial design and analysis in the era of non-proportional hazards of the treatment effect. Trials **15**, 314 (2014)
5. Jachno, K., Heritier, S., Wolfe, R.: Are non-constant rates and non-proportional treatment effects accounted for in the design and analysis of randomised controlled trials? a review of current practice. BMC Medical Research Methodology **19**(1), 103 (2019)
6. Lin, R.S., Lin, J., Roychoudhury, S., Anderson, K.M., Hu, T., Huang, B., Leon, L.F., Liao, J.J.Z., Liu, R., Luo, X., Mukhopadhyay, P., Qin, R., Tatsuoka, K., Wang, X., Wang, Y., Zhu, J., Chen, T.-T., Iacona, R.: Alternative analysis methods for time to event endpoints under nonproportional hazards: A comparative analysis. Statistics in Biopharmaceutical Research **12**(2), 187–198 (2020)
7. Sato, Y., Gosho, M., Nagashima, K., Takahashi, S., Ware, J.H., Laird, N.M.: Statistical methods in the journal – an update. New England Journal of Medicine **376**(11), 1086–1087 (2017)
8. Royston, P., Lambert, P.C.: Flexible Parametric Survival Analysis Using Stata: Beyond the Cox Model. Stata Press, College Station, TX (2011)

9. Royston, P.: Estimating the treatment effect in a clinical trial using difference in restricted mean survival time. Stata Journal **15**(4), 1098–1117 (2015)
10. Uno, H., Claggett, B., Tian, L., Inoue, E., Gallo, P., Miyata, T., Schrag, D., Takeuchi, M., Uyama, Y., Zhao, L., Skali, H., Solomon, S., Jacobus, S., Hughes, M., Packer, M., Wei, L.-J.: Moving beyond the hazard ratio in quantifying the between-group difference in survival analysis. Journal of Clinical Oncology **32**(22), 2380–2385 (2014)
11. Pocock, S.J., Clayton, T.C., Altman, D.G.: Survival plots of time-to-event outcomes in clinical trials: good practice and pitfalls. The Lancet **359**(9318), 1686–1689 (2002). doi:10.1016/S0140-6736(02)08594-X
12. Altman, D.G., De Stavola, B.L., Love, S.B., Stepniewska, K.A.: Review of survival analyses published in cancer journals. British Journal of Cancer **72**(2), 511–518 (1995). 7640241[pmid]; Br J Cancer
13. Mathoulin-Pelissier, S., Gourgou-Bourgade, S., Bonnetain, F., Kramar, A.: Survival end point reporting in randomized cancer clinical trials: A review of major journals. Journal of Clinical Oncology **26**(22), 3721–3726 (2008). doi:10.1200/jco.2007.14.1192
14. Hernán, M.A.: The hazards of hazard ratios. Epidemiology (Cambridge, Mass.) **21**(1), 13–15 (2010). doi:10.1097/EDE.0b013e3181c1ea43. 20010207[pmid]; Epidemiology
15. 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 **11**(5), 0154870 (2016). doi:10.1371/journal.pone.0154870
16. Pocock, S.J., Travison, T.G., Wruck, L.M.: Figures in clinical trial reports: current practice and scope for improvement. Trials **8**, 36 (2007). doi:10.1186/1745-6215-8-36
17. Wei, Y., Royston, P.: Reconstructing time-to-event data from published kaplan–meier curves. Stata Journal **17**(4), 786–802 (2017). doi:10.1177/1536867X1701700402
18. Guyot, P., Ades, A., Ouwens, M.J., Welton, N.J.: Enhanced secondary analysis of survival data: reconstructing the data from published kaplan-meier survival curves. BMC Medical Research Methodology **12**(1), 9 (2012). doi:10.1186/1471-2288-12-9
19. Royston, P., Parmar, M.K.B.: Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. Statistics in Medicine **21**(15), 2175–2197 (2002). doi:10.1002/sim.1203
20. Grambsch, P.M., Therneau, T.M.: Proportional hazards tests and diagnostics based on weighted residuals. Biometrika **81**(3), 515–526 (1994). doi:10.2307/2337123
21. Dearnaley, D.P., Jovic, G., Syndikus, I., Khoo, V., Cowan, R.A., Graham, J.D., Aird, E.G., Bottomley, D., Huddart, R.A., Jose, C.C., Matthews, J.H.L., Millar, J.L., Murphy, C., Russell, J.M., Scrase, C.D., Parmar, M.K.B., Sydes, M.R.: Escalated-dose versus control-dose conformal radiotherapy for prostate cancer: long-term results from the mrc rt01 randomised controlled trial. The Lancet Oncology **15**(4), 464–473 (2014). doi:10.1016/S1470-2045(14)70040-3
22. Bonner, J.A., Harari, P.M., Giralt, J., Azarnia, N., Shin, D.M., Cohen, R.B., Jones, C.U., Sur, R., Raben, D., Jassem, J., Ove, R., Kies, M.S., Baselga, J., Youssoufian, H., Amellal, N., Rowinsky, E.K., Ang, K.K.: Radiotherapy plus cetuximab for squamous-cell carcinoma of the head and neck. New England Journal of Medicine **354**(6), 567–578 (2006). doi:10.1056/NEJMoa053422
23. Shipley, W.U., Seiferheld, W., Lukka, H.R., Major, P.P., Heney, N.M., Grignon, D.J., Sartor, O., Patel, M.P., Bahary, J.-P., Zietman, A.L., Pisansky, T.M., Zeitzer, K.L., Lawton, C.A.F., Feng, F.Y., Lovett, R.D., Balogh, A.G., Souhami, L., Rosenthal, S.A., Kerlin, K.J., Dignam, J.J., Pugh, S.L., Sandler, H.M.: Radiation with or without antiandrogen therapy in recurrent prostate cancer. New England Journal of Medicine **376**(5), 417–428 (2017). doi:10.1056/NEJMoa1607529
24. Perren, T.J., Swart, A.M., Pfisterer, J., Ledermann, J.A., Pujade-Lauraine, E., Kristensen, G., Carey, M.S., Beale, P., Cervantes, A., Kurzeder, C., Bois, A.d., Sehouli, J., Kimmig, R., Stähle, A., Collinson, F., Essapen, S., Gourley, C., Lortholary, A., Selle, F., Mirza, M.R., Leminen, A., Plante, M., Stark, D., Qian, W., Parmar, M.K.B., Oza, A.M.: A phase 3 trial of bevacizumab in ovarian cancer. New England Journal of Medicine **365**(26), 2484–2496 (2011). doi:10.1056/NEJMoa1103799
25. Fox, K.: 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). The Lancet **362**(9386), 782–788 (2003). doi:10.1016/S0140-6736(03)14286-9
26. Morris, T.P., Jarvis, C.I., Cragg, W., Phillips, P.P.J., Choodari-Oskooei, B., Sydes, M.R.: Proposals on kaplan–meier plots in medical research and a survey of stakeholder views: Kmunicate. BMJ Open **9**(9), 030215 (2019). doi:10.1136/bmjopen-2019-030215
27. Trinquart, L., Jacot, J., Conner, S.C., Porcher, R.: Comparison of treatment effects measured by the hazard ratio and by the ratio of restricted mean survival times in oncology randomized controlled trials. Journal of Clinical Oncology **34**(15), 1813–1819 (2016). doi:10.1200/jco.2015.64.2488
28. Rahman, R.M., Fell, G., Ventz, S., Arfe, A., Vanderbeek, A.M., Trippa, L., Alexander, B.M.: Deviation from the proportional hazards assumption in randomized phase 3 clinical trials in oncology: Prevalence, associated factors and implications. Clinical Cancer Research, 3999–2018 (2019). doi:10.1158/1078-0432.Ccr-18-3999
29. Royston, P., Choodari-Oskooei, B., Parmar, M.K.B., Rogers, J.K.: Combined test versus logrank/cox test in 50 randomised trials. Trials **20**(1), 172 (2019). doi:10.1186/s13063-019-3251-5
30. Stensrud, M.J., Hernán, M.A.: Why test for proportional hazards? JAMA **323**(14), 1401–1402 (2020). doi:10.1001/jama.2020.1267
31. Gregson, J., Sharples, L., Stone, G.W., Burman, C.-F., Öhrn, F., Pocock, S.: Nonproportional hazards for time-to-event outcomes in clinical trials: Jacc review topic of the week. Journal of the American College of Cardiology **74**(16), 2102–2112 (2019). doi:10.1016/j.jacc.2019.08.1034
32. Hernán, M.A., Robins, J.: Causal Inference: What If. Chapman and Hall CRC, Boca Raton (2020)
33. ICH E9 (R1) Addendum on Estimands and Sensitivity Analysis in Clinical Trials to the Guideline on Statistical Principles for Clinical Trials. European Medicines Agency, Amsterdam, The Netherlands (2020)
34. Royston, P., Parmar, M.K.: 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 **13**(1), 152 (2013). doi:10.1186/1471-2288-13-152

35. Zhao, L., Claggett, B., Tian, L., Uno, H., Pfeffer, M.A., Solomon, S.D., Trippa, L., Wei, L.J.: On the restricted mean survival time curve in survival analysis. Biometrics **72**(1), 215–21 (2016). doi:10.1111/biom.12384
36. Tufte, E.: The Visual Display of Quantitative Information 2nd Ed. Graphics Press, Cheshire, Conn. (2001)
37. Cleveland, W.S.: The Elements of Graphing Data. ATT Bell Laboratories, Summit, NJ (1994)
38. Yau, N.: Visualize This: The FlowingData Guide to Design, Visualization, and Statistics. Wiley Pub, Indianapolis, Ind (2011)

Additional Files

Additional File 1 — Citations and dataset of Review trials and results
Additional File 2 — Supplementary Figures
Additional File 3 — Stata code to generate the proposed complementary presentation of Kaplan-Meier and treatment effect plots

Chapter 6 Discussion and Conclusions

The focus of this thesis on time-dependent effects of treatment in randomised trials reflects that clinical trials are the gold standard for examining the impact of these treatments. In clinical trials with time-to-event outcomes, the hazard ratio estimated from a Cox PH model has been used almost exclusively as the measure of treatment effect. However, nonproportional hazards are being detected more frequently with the advent of new treatments with novel mechanisms of action and the use of composite outcomes - multiple endpoints jointly assessed as a single outcome - in clinical trials, calling into question the presentation of a single HR as an adequate summary of a clinical trial findings. Refinements to existing methods and new methods to deal with specific types of nonproportionality such as lag to effect have been proposed although the evaluation and uptake of these methods has not been exhaustive. Guidance on the implementation of the methods and reporting may also require development. The aim of the research in this thesis was to assess how nonproportional hazards and non-constant event rates are allowed for in the design, analysis and reporting of clinical trials with time-to-event outcomes, to examine the relative performance of competing methods, and to identify areas where additional guidance on the implementation could be useful.

In order to examine the approaches to design, analyse and report time-to-event outcomes, a review of current practice was undertaken (Chapter 2). The review assessed the sample size calculation methods to see if the effect of either non-constant hazard rates or anticipated nonproportionality was allowed for during the trial design, and recorded the method to analyse and present the main outcomes of the trials. When an analytical method assuming proportional hazards was employed for the primary outcome, the reporting of assumptions underlying these methods was assessed. A simulation study using the statistical methods identified in the review including tests of survival curve difference and regression-based measures of treatment effect was undertaken to determine the impact of a clinically plausible non-constant baseline hazard on the detection of time-dependent treatment effects (Chapter 3). The findings from the simulation study justified use of alternative regression-based methods to examine in detail the evidence for time-dependent treatment effects from a large long-running community-based clinical trial. This application study provided an opportunity to assess for potential interplay between underlying event rates and nonproportionality (Chapter 4). The review reported in Chapter 2 highlighted that the predominant means to visually present trial findings for time-to-event endpoints was a Kaplan-Meier plot. Of potential concern is that this plot does not directly provide for an assessment of treatment effect consistency over time. Hence in Chapter 5 we presented a complementary plot that enables intuitive assessment of the dynamic nature of any treatment group differences over time.

6.1 Summary of the thesis chapters

6.1.1 Chapter 2 – Are non-constant event rates and non-proportional treatment effects accounted for in the design and analysis of randomised controlled trials? A review of current practice

Chapter 2 presented the results of the review of all original reports published between January and June 2017 in four high impact medical journals involving trials for which the primary outcome involved time-to-event analysis. The aims of the review were to identify whether non-constant event rates and time-dependent treatment effects were allowed for in the sample size calculations of trials, and to assess the methods used for the analysis and reporting of time-to-event outcomes with a focus on the awareness and reporting of testing for nonproportional treatment effects when the main analytical method involved the Cox model.

Key findings from the review included:

- time-to-event outcomes were the predominant primary outcome in phase III trials (66/168; 39%)
- sample size calculations that explicitly assume proportional hazards, or are maximally powerful under an assumption of proportional hazards were used in the majority of trials (48/66; 73%) with
    - calculations based on the logrank test the most common (40/48; 83%)
    - calculations based on a difference between exponential survival distributions (4/48; 8%) or the beta coefficient of the Cox model (4/48:8%) the other approaches used
- simulation-based sample size calculations for predicted non-constant event rates or allowing for non-proportional treatment effects are being employed (7/66; 11%)
- reporting of sample size calculations has improved over time due to more stringent regulatory requirements, however there is still room for improvement
- in an analysis of trials, the HR from a Cox PH model is used most frequently as a means to assess for significance and to quantify treatment effect (64/66; 97%)
- the logrank test of significance of treatment effect was also provided in many trials (58/66; 88%)
- parametric regression-based modelling approaches were planned or used in a minority of trials (7/66; 11%), usually as a supplementary or secondary analysis method to semi-parametric Cox modelling with only one trial using parametric regression for its inferential finding
- graphical presentation of the primary time-to-event outcome was either a Kaplan-Meier survival plot or its reciprocal, a cumulative incidence plot (65/66; 98%)
- when the Cox model was used, awareness and reporting of the importance of the proportional hazards assumption was not optimal

  - half of the trials indicated awareness (34/64; 53%) or included details of planned tests (31/64; 48%)
  - explicit reporting of PH testing results was rare (7/64; 11%)

The review highlights a gradual change in analysis approaches over recent decades with recognition that quantification of treatment effect is of crucial importance in addition to hypothesis testing. Use of the Cox model has increased from 4/113 (4%) trials published during 1991 [37] to 64/66 (97%) of trials in our review. This review was the first to document the level of usage of parametric modelling approaches for analysing trials with time-to-event outcomes.

The review also demonstrated the potential of regulatory guidelines in conjunction with journal editorial boards to impact on the quality of reporting of trials. In September 2004 the International Committee of Medical Journal Editors disseminated a policy requiring pre-trial public registration as a condition of publication of trials with a start date from July 2005 onwards with retrospective registration of trials with pre-July 2005 start dates also strongly encouraged. Following implementation of this policy, all trials after July 2005 were registered prior to, or in a timely manner after the nominated start date of the trial. No trials which began prior to July 2005 were registered prior to their start date, with all bar one trial registered in the ensuing years as they published findings from their trials.

6.1.2 Chapter 3 – Impact of a non-constant baseline hazard on detection of time-dependent treatment effects: a simulation study

Chapter 3 presented a simulation study investigating the impact of a non-constant baseline hazards in the presence of time-dependent treatment effects. The parameter values used in constructing the simulated datasets and the statistical methods evaluated were informed by the findings of the review (Chapter 2).

In our review, many of the trials exhibiting potential nonproportionality of treatment effect were from oncology research. The advent of immunotherapy-based treatments for cancer has resulted in identification of two forms of nonproportionality of particular interest - a time lag until treatment becomes effective and an early effect of treatment that ceases. In sample size calculations for time-to-event outcomes where information is based on the number of events rather than the number of participants, correct specification of the baseline hazard can be crucial when any nonproportionality might be anticipated. From the review, six of the seven trials that employed simulation-based approaches to sample size determinations (in anticipation of changing event rates or changing treatment effects over time) were oncology trials. However, there were over twenty oncology trials involving immunotherapies where standard sample size calculations were employed and these calculations carry an implicit assumption of constant event rates and are maximally powerful under proportional hazards.

The simulation study aimed to assess the impact of clinically plausible non-constant event rates when there was no time-dependent treatment effect ie under a proportional hazards assumption, and also when there exists time-dependent treatment effects in the form of either lag until effect or early effect that ceases. The performance of commonly utilised regression-based measures of treatment effect and tests of survival curve difference was assessed in terms of power.

Key findings from the simulation study included:

- the lack of stability of all commonly utilised methods of analysis in terms of the power to detect treatment effects in the presence of clinically plausible durations of non-proportionality and modest non-constant event rates
- no single summary estimate of treatment effect was able to adequately describe the full extent of a potentially time-limited treatment effect and maintain power at nominal levels
- judicious selection of designated cut points for period-specific estimands could result in improved estimates of treatment effect but may also result in decreased power under proportional hazards and/or increased Type I errors
- depending on the nature of the nonproportionality, non-constant event rates could further exacerbate or somewhat ameliorate losses in power, treatment effect magnitude and coverage
- the novel reporting of the interplay between nonproportionality and the shape of the baseline hazard rates and exploration of the implications for clinical trial designs

This work highlights the importance of analysis methods which allow for the shape of the baseline hazard to enable a richer exploration of the timing, magnitude and persistence of any treatment effects. A range of different effect measures- HRs, piecewise HRs, milestone survival probabilities, RMST difference - presented as a series of time period-based estimates or via graphical formats enables a comprehensive evaluation of the effect based on the whole follow-up time.

6.1.3 Chapter 4 – Examining evidence for time-dependent treatment effects using alternative regression-based methods in the ASPREE clinical trial

Chapter 4 presented the findings for time-dependent treatment effects using selected endpoints from a large long-running community-based clinical trial. Primary analyses of the trial endpoints employed a Cox PH modelling approach and had not identified any compelling evidence of nonproportionality for the primary endpoints. Based on the results of the simulation study (Chapter 3) we focused on regression-based methods and graphical exploratory analyses to examine the evidence for any time-dependent treatment effects in the ASPREE trial. By utilising regression-based methods allowing for time-dependent treatment effects, this work aimed to illustrate potential new insights or increased clinical understanding into the magnitude and persistence of treatment effects that could be gained even when there was no statistical evidence against the assumption of proportionality.

This study estimated treatment effects in the form of a hazard ratio (HR) using
(1) the semi-parametric Cox model,
(2) the parametric Weibull model and
(3) flexible parametric models using splines to model the baseline hazard under an assumption of proportional hazards,

and treatment effects in the form of a difference in restricted mean survival time(ΔRMST) using
(4) flexible parametric models using splines to model the baseline hazard under a proportional hazards assumption or
(5) flexible parametric models using splines to model the baseline hazard allowing for time-dependent treatment effects or
(6) generalised linear modelling of transformed datasets consisting of pseudo-observations which allow for time-dependence of treatment effect equivalent to that estimated non-parametrically by the Kaplan-Meier estimate of survival probability.

Key findings from this research included the following:

- illustrative examples of the use of relative and absolute estimands of treatment effect to obtain complementary information on the emergence, magnitude and balance between benefit and harms over time of estimated treatment effects
- exploration of the evidence for emerging time-dependent treatment effects of aspirin directly, and time-dependent interactions of aspirin in subgroups not previously been reported in trial findings
- visualisation of the modelling approaches and presentation of risk-based and time-based estimates of treatment effect aimed at clinicians enabling comprehensive evaluation of treatment effects

6.1.4 Chapter 5 – Complementing the Kaplan-Meier plot to enable assessment of treatment effect consistency with proportional hazards

In the review publication (Chapter 2) Kaplan-Meier plots were included in almost all reports of trial findings. These curves intuitively display the survival experience in treatment groups over time but do not directly provide for an assessment of the treatment effect measure which is of primary interest to trialists. In the publication presented in Chapter 5 a series of general graphical and survival-curve specific recommendations were collated and harmonised from previous researcher reviews and guidelines for presentation of survival curve estimates. Plots from the trials in the review in Chapter 2 were assessed for adherence to the recommendations and guideline. We proposed a plot of treatment effect over time to be presented as an accompaniment plot to Kaplan-

Meier survival curves. Our proposed arrangement enables intuitive assessment of the consistency of treatment effect over time.

Key results from this research included the following:

- provision of a series of recommendations for general graphing components and survival curve specific components harmonised from previous reviews of Kaplan-Meier plots and further informed by seminal data visualisation resources
- findings of overall excellent adherence to most of the general graphing and survival curve specific components in contrast to the earliest reviews of adherence to recommendations which found adherence to most recommendations was poor
- identification of a remaining area for improvement of the presentation of Kaplan-Meier plots being the depiction of the uncertainty associated with survival curve estimates over time
- our proposal for a complementary plot of treatment effect measure to accompany Kaplan-Meier plots to provide for direct assessment of the treatment effect consistency with proportional hazards
- through presentation of reconstructed individual patient datasets from previously published trials showing different levels of proportionality and baseline event rates, illustration of the utility of the treatment effect plots to enhance intuitive insight into the dynamic nature of any treatment effect measure

6.2 Integrated discussion of overall findings

The overall aim of this thesis was to advance the existing body of knowledge on the design, analysis and reporting of time-to-event analyses in the presence of nonproportionality of treatment effect. Adequately accounting for nonproportionality in trials is an important and active research area as nonproportionality and non-constant event rates are encountered more frequently.

We reviewed the sample size calculations from recently published trials for our review for the adequacy of reporting and for allowances for nonproportionality. Previous reviews found that whilst reporting of sample size calculations has improved over time as a result of more stringent requirements imposed by regulatory bodies and journals [31, 32], there were still inadequacies in the assumption reported. We similarly found that the initial sample size calculation could have been more adequately reported. We found the majority of trials were using calculation methods that explicitly assume proportional hazards, or are maximally powerful under a proportional hazards assumption. No trials used any of the more recently proposed modified sample size calculations to allow for specified forms of nonproportionality [24, 46, 47] but there were encouraging signs that researchers are beginning to anticipate the impacts of nonproportionality and the shape of the baseline event rate with a minority of trials using calculations involving a series of stages within a trial or simulation-based procedures [26, 48, 49].

Lack of awareness of the proportional hazards assumption and concerns about the testing and reporting of this assumption when trial results are based on a Cox model have been evident for several decades. Reviews of the usage of the Cox model over the past three decades have all highlighted the lack of planned testing, or comprehensiveness of any results for assessing for nonproportionality [3, 37, 39, 40]. Over the same time frame, our review and other demonstrates the changes in modelling approaches over recent times as quantification of treatment effects has gained prominence over hypothesis testing approaches. Whilst the logrank test is still employed and reported in many of the trials, it is the hazard ratio from the Cox model that is presented as the primary means to convey the trial findings. Our review also noted that additional landmark analyses and the use of period-specific hazard ratios were used in several of the trials, tacit acknowledgment from the authors that one summary measure of treatment effect did not fully describe trial findings.

As part of our simulation study, we assessed the impact of non-constant event rates on the power of a range of tests of survival curve difference due to treatment in the presence of two forms of nonproportionality, a lag until effect and an early effect that ceases [13–15, 50, 51]. Our conclusions and those of similar comparative studies published recently were that there is no consistently powerful test across all forms of nonproportionality that can be recommended. Forms of a versatile test combining information from multiple weighted logrank tests were the most useful at detecting some level of nonproportionality whilst maintaining adequate Type I control [52–54]. We further highlighted the lack of robustness of these tests to maintain power in the presence of even small deviations from proportionality can be further exacerbated by clinically plausible non-constant event rates.

A range of regression-based approaches and extensions to these estimators enabled us to assess the impact on the magnitude of treatment effect estimate and coverage values benchmarked to the values specified by design assumptions. We compared three different estimates of treatment effect through analytical approaches including a landmark approach to obtain a hazard ratio from the Cox model, piecewise exponential models to obtain period-specific hazard ratios, Royston-Parmar models under a PH assumption to obtain hazard ratios and differences in RMST, Royston-Parmar models allowing for time-dependence of treatment to obtain differences in RMST and used a Weibull accelerated failure time model to estimate a time ratio [2, 16, 55–58]. Despite the multitude of modelling approaches, the three estimands we compared were broadly similar across the nonproportionality scenarios in terms of the adverse impact of increasing amounts of non-PH. Again, the impact of non-constant event rates in the presence of nonproportionality was to partially diminish or further exacerbate losses in power and treatment effect magnitude. Judicious selection of designated cut points or landmark time points could result in improved estimates of treatment effect magnitude but would need to be clearly pre-specified in order to be valid.

The interpretation of changes in period-specific or weighted hazard ratios as changes in treatment effect is concerning from a causal perspective [59]. Period-specific hazard ratios are subject to selection bias due to the existence of 'frailty' factors which affect a patient's survival time. At randomisation, the distribution of these factors in the trial is balanced on average, but at later times during follow-up different treatment groups will have systemically different distributions of these factors [60, 61]. Period-specific HRs estimated from latter periods reflect the effects of treatment and the effect of differences in the distribution of frailty factors between the two groups. These concerns about the lack of comparability between two treatment groups extend to the weighted HR and even the unconditional HR, raising doubts about the interpretability of the HR even under proportional hazards [62, 63]. The use of alternative estimands such as the difference in RMST or differences of survival probability at specified times, or specified quantiles such as the median survival time may be more appropriate than the reporting of hazard ratios when nonproportionality is present, and even when it is not. Reporting of multiple measures of treatment effect enables a more comprehensive assessment of treatment effect over time and facilitates the evaluation of the timing, magnitude and persistence of such effect.

As an illustrative example of this approach to more comprehensively reporting treatment effects for a trial, we compared a range of regression-based approaches allowing for assessment of time-dependent treatment effects using a range of outcomes from a previously reported long-running community trial. For the majority of the selected outcomes, we were able to confirm that summary estimates of treatment effect obtained from models assuming proportional hazards were suitable descriptions of the trial findings. We demonstrated the use of relative and absolute estimands of treatment effect to obtain complementary information on the emergence, magnitude and balance between benefit and harms over time of estimated treatment effects, again useful even when there was no evidence of nonproportionality. We found evidence for a time-dependent treatment effect of aspirin on a cancer outcome that had not previously been reported in trial findings. By investigating possible subgroup interactions with treatment, we found some evidence of differential effects of an adverse side-effect in males and females. We also found a time-dependent interaction effect of treatment and age on the risk of cardiac events.

Recommendations for the use of alternative estimands to the HR [3, 8, 12, 64, 65] and clearer reporting of potentially dynamic treatment effects have been supported by the causal inference literature and regulatory agencies [34, 35, 63, 66, 67]. The main graphical presentation of time-to-event results, survival curves based on the Kaplan-Meier estimates do not show the treatment effect directly and do not enable easy examination for the presence of any nonproportionality. Visual presentations of treatment effect showing time-dependent treatment effects would give impetus to explore possible reasons for the time-dependence, encouraging further analysis to ascribe if possible any non-proportionality to time-dependent treatment effects, subgroup heterogeneity or allowance for unobserved covariate effects. To this end, we present a complementary plot to the Kaplan-Meier survival curves that enables more intuitive insight into the dynamic nature of any treatment effects. We illustrate the utility of our proposed plot using reconstructed datasets from published trials with varying degrees of nonproportionality and different baseline hazards.

6.3 Limitations and future directions

We undertook the scoping review to understand how nonproportionality and non-constant event rates are accounted for in the design and analysis of randomised controlled trials. Although not a systematic review, we aimed to follow the principles of the Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) statement [68] for the planning and documentation of review outcomes, especially for the data management, data itemisation and collection processes. All data extraction was undertaken by one reviewer only (Kim Jachno) so no protocol was required to obtain consensus for any discrepant finding; however, this single data extraction is a limitation of the completed review. Wherever possible, objective automated key word strategies were employed to ensure all relevant sections of the published report and all supplementary information were reviewed. We selected four high impact journals that have emphasised the CONSORT guidelines as part of their submission requirements and might be expected to be home to high quality presentation. Thus, our findings may not reflect the full spectrum of reporting quality possible for clinical trial findings and different insights might have resulted if a more comprehensive range of journals had been included. The restricted range of journals did enable us to discuss our findings alongside previous reviews of clinical trials reports involving time-to-event outcomes from the same - or similar - journals. This enabled trends in modelling approaches and the adequacy of reporting to be identified. Future reviews could continue this assessment of the use of alternative methodologies for trial design, analysis and reporting approaches in the presence of nonproportionality and non-constant baseline hazards.

In the chapters of the thesis exploring alternative analytical approaches to assess for time-dependent treatment effects, we considered only a subset of the possible ways in which time-to-event outcomes could be assessed. There is a growing body of research encompassing tests of significance and regression-based methods that allow for estimation of the effect of treatment over time in the presence of anticipated patterns of nonproportionality. We aimed to include the most widely used approaches implemented in the medical research literature and additionally include examples of the more recently developed proposed approaches that may not yet have been used as pre-specified analytical methods in statistical analysis plans. The tests of significance we included in our simulation study broadly overlapped with two research reports published after our simulation study had been carried out [53, 54]. Their objective was to assess their performance in terms of power to detect a treatment effect under selected non-PH scenarios. Those reports provided similar conclusions to our findings as to the robustness of versatile tests encompassing a range of weightings to allow for multiple non-PH scenarios as a means to establish statistical significance of a treatment difference. Additionally, in our simulation from Chapter 3 and in the application paper presented in Chapter 4, we also assessed regression-based approaches that allow for quantification of the effect of treatment over time for both risk-based and time-based treatment effect measures. Again, not all possible approaches were included but the chosen ones are representative of the methods typically used to account for time-dependent treatment effects and are included in similar illustrative papers [42, 69].

In the simulation study in Chapter 3 we only considered nonproportionality manifesting as time-dependent treatment effects of the form of either a lag until effect or an early effect that ceases. These forms were achieved through use of a piecewise Weibull model with one change point. There are a host of ways in which this simulation work could be extended to further enrich our understanding. The results of our simulation could be generalised by adding more data-generating scenarios such as crossing survival curves or allowing for cure fractions or increasing the number of change points in the piecewise data-generating model. Other parametric baseline hazards, such as the Gompertz or a mixture of Weibull distributions, could be utilised and it is possible to incorporate nonproportionality via continuous covariates to more closely approximate the shape of any baseline hazard and nonproportionality likely to be encountered in realistic settings.

Additionally, we could have investigated multiple treatment effect magnitudes, and explored more complex model formulations with multiple covariates demonstrating treatment effect heterogeneity. We did not cover the effects of censoring and enrolment rates or the effect of adjusting sample size and follow up times all of which impact on the interplay of non-PH and event rates. However, we aimed to undertake a simulation that provided enough scenarios to clearly demonstrate that single summary effect measures are unable to comprehensively describe the magnitude of treatment effect over time when that effect changes over time, that what could be regarded as negligible periods of nonproportionality could have noticeable impacts on the power to detect treatment effects and that the impact of clinically plausible non-constant event rates could further impact on the loss of power depending on the nature of nonproportionality and the shape of the underlying hazard. Nevertheless, there remains further work to expand out knowledge of the circumstances under which nonproportionality must be taken in to account and when it might be safe to ignore it in a simplified analysis.

Whilst we considered how visual display literature [70–72] could be used to inform recommendations for our proposed composite presentation of Kaplan-Meier plots and treatment measure estimations over time, alternative approaches such as consensus-based methods from surveys of end users could also have been undertaken and may have resulted in different recommendations or placed a different emphasis on aspects of the presentation not foreseen by us. The availability of freely available user-friendly software is key to achieving meaningful adoption of reporting recommendations in applied biostatistics. We have provided code to enable other researchers to generate similar graphs using their own preferred means of time-dependent treatment effect estimation. Writing of general-purpose user-friendly software to implement the graphical presentations outlined here utilising Royston-Parmar models [17, 58] would be of value. The extension of the summary plot we proposed to the cumulative incidence curve in the presence of competing events is equally worth pursuing. The use of pseudovalues equivalent to nonparametric estimation [73, 74] of treatment effect incorporating earlier proposals to improve the presentation of Kaplan-Meier curves [75] is another avenue for future work.

For our proposed arrangement of treatment effect plots we undertook preliminary presentations in a seminar context for clinicians involved in the ASPREE trial. We received very positive feedback of the increased clinical insight available by assessing for possible time-dependent treatment effects. Most often these experiences provided reassurance to clinicians that a summary fixed hazard ratio provided by the Cox PH model was an appropriate means to describe the effect of treatment for the entirety of the trial duration. As importantly, the graphs were also able to convey the importance of considering nuanced effects of treatment over time such as gradual increasing or decreasing efficacy or transitory periods of increased risk. These subtle trends might not be detected using formal statistical tests for the PH assumption but can still have important clinical implications, especially when married to a strong biological rationale. Similar trials with a longer follow up, such as those conducted in the field of cardiovascular diseases, would allow for the assessment of long-term effects of treatment that could otherwise be missed. Trials of cardiovascular diseases provide some of the earlier examples of time-dependent treatment effects such as the LIPID trial conducted in the 1990s where the authors established a benefit of statin treatment increasing with time over the seven years of follow up by employing novel tests of time-dependence of effect [76]. A more recent RCT with apparent time-dependent and cross-over treatment benefits was observed in the ISCHAEMIA trial comparing survival outcomes following invasive intervention versus optimal medical therapy in coronary heart disease [77].

Following cessation of the intervention, the majority of the ASPREE trial participants have been enrolled in an extended observational study in order to examine the legacy effects of daily aspirin use. Further work following up these participants for a range of endpoints could involve the analysis approaches presented in this thesis; in particular for examining evidence of long-term aspirin effects on cancer prevention which has been proposed to become apparent only after approximately 5 years and through 10 years and longer follow up [78]. Further clarification and guidance for clinicians and the research community on the interpretability of different estimators of treatment effect and their relevance to an individual patient’s experience is required.

6.4 Conclusions

The Cox PH model with its hazard ratio as a summary measure of treatment effect has been the basis of designing and analysing clinical trials with time-to-event outcomes for many decades. However, nonproportionality is being observed more frequently due to the mechanistic nature of new interventions and because increased regulatory oversight has required the conduct of larger, longer trials. Our review showed that, despite the slow improvement in the design, analysis and reporting of time-to-event outcomes, the presentation of a unique summary measure of treatment effect was not adequate when nonproportionality is present. When the assumption of PH is satisfied, the Cox PH model is the most statistically powerful method and the interpretation of a hazard ratio as the measure of treatment effect is widely understood by clinicians; however, the time has come to rely less systematically on the hazard ratio alone.

Even when the assumption of PH holds, there may be some advantages in presenting treatment effect estimates in both risk-based and time-based metrics which provide complementary information from a clinical perspective. We aimed to illustrate the increased insight and clinical understanding that can be obtained through the application of alternative regression-based methods for time-to-event outcomes and through our proposed presentation of complementary plots of survival probability and treatment effect estimate over time.

When major deviations from PH are anticipated, it may be possible to adapt the design via logistical considerations and/or pre-specify analysis techniques that maintain power to detect treatment effects [13, 15, 53, 54]. When early treatment effects are anticipated for example, recruiting more patients and running a trial of shorter duration may maximise power albeit at the (intentional) cost of no information on the longer-term effects of the treatment. In this setting judicious selection of appropriately weighted tests of survival curve difference or cutoff times for period-based estimands could be employed. Our simulation study demonstrated the need to allow for the additional impact of non-constant event rates should any nonproportionality be anticipated. However, it is not intuitive how to examine the appropriateness of any treatment benefit resulting from the use of differential weight functions from a clinical perspective.

In most circumstances anticipating the existence and correct form of nonproportionality can be hard and ideally trials should continue for sufficient time so that the long-term effects of treatment can be adequately estimated. For this reason, the default pre-specified analysis may still involve methods maximally powerful under PH. However, there should be detailed contingency plans for alternative primary analyses should clear evidence of nonproportionality be detected. Flexible parametric modelling methods allow for a generalised approach by estimating both the magnitude and the shape of treatment effects over time based on the data and should be more widely considered as an analysis approach. Popularising different measures of treatment effect through graphical displays is another way forward. The use of different measures of treatment effect will provide the means to comprehensively assess the evolution of effect over time and facilitate the clinical evaluation of treatments.

References

[1] Sato Y, Gosho M, Nagashima K, et al. Statistical methods in the Journal — an update. *New England Journal of Medicine* 2017; 376: 1086–1087.

[2] Cox DR. Regression models and life-tables. *Journal of the Royal Statistical Society Series B (Methodological)* 1972; 34: 187–220.

[3] 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: 1813–1819.

[4] Rahman RM, Fell G, Ventz S, et al. Deviation from the proportional hazards assumption in randomized Phase 3 clinical trials in oncology: Prevalence, associated factors and implications. *Clinical Cancer Research* 2019; 25: 6339–6345.

[5] Royston P, Choodari-Oskooei B, Parmar MKB, et al. Combined test versus logrank/Cox test in 50 randomised trials. *Trials* 2019; 20: 172.

[6] Therneau TM, Grambsch PM. *Modeling survival data: Extending the Cox model*. Book, New York: Springer, 2000.

[7] Zhang X, Long Q. Modeling and prediction of subject accrual and event times in clinical trials: A systematic review. *Clinical Trials* 2012; 9: 681–688.

[8] Royston P, Parmar MK. 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: 152.

[9] 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: 2409–2421.

[10] Uno H, Claggett B, Tian L, et al. Moving beyond the hazard ratio in quantifying the between-group difference in survival analysis. *Journal of Clinical Oncology* 2014; 32: 2380–2385.

[11] Royston P. Estimating the treatment effect in a clinical trial using difference in restricted mean survival time. *Stata Journal* 2015; 15: 1098–1117.

[12] Zhao L, Claggett B, Tian L, et al. On the restricted mean survival time curve in survival analysis. *Biometrics* 2016; 72: 215–21.

[13] Fleming TR, Harrington DP. Weighted logrank statistics. In: *Counting processes and survival analysis*. Book Section, Wiley Series in Probability; Statistics, pp. 255–285.

[14] Royston P, Parmar MK. Augmenting the logrank test in the design of clinical trials in which non-proportional hazards of the treatment effect may be anticipated. *BMC Medical Research Methodology* 2016; 16: 16.

[15] Karrison TG. Versatile tests for comparing survival curves based on weighted log-rank statistics. *Stata Journal* 2016; 16: 678–690.

[16] Royston P, Parmar MKB. Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. *Statistics in Medicine* 2002; 21: 2175–2197.

[17] Lambert PC, Royston P. Further development of flexible parametric models for survival analysis. *Stata Journal* 2009; 9: 265–290.

[18] Schoenfeld DA, Richter JR. Nomograms for calculating the number of patients needed for a clinical trial with survival as an endpoint. *Biometrics* 1982; 38: 163–170.

[19] Freedman LS. Tables of the number of patients required in clinical trials using the logrank test. *Statistics in Medicine* 1982; 1: 121–129.

[20] Lachin JM. Introduction to sample size determination and power analysis for clinical trials. *Controlled Clinical Trials* 1981; 2: 93–113.

[21] Lachin JM, Foulkes MA. Evaluation of sample size and power for analyses of survival with allowance for nonuniform patient entry, losses to follow-up, noncompliance, and stratification. *Biometrics* 1986; 42: 507–19.

[22] Hsieh FY, Lavori PW. Sample-size calculations for the Cox proportional hazards regression model with nonbinary covariates. *Controlled Clinical Trials* 2000; 21: 552–560.

[23] Hasegawa T. Sample size determination for the weighted log-rank test with the Fleming–Harrington class of weights in cancer vaccine studies. *Pharmaceutical Statistics* 2014; 13: 128–135.

[24] Sit T, Liu M, Shnaidman M, et al. Design and analysis of clinical trials in the presence of delayed treatment effect. *Statistics in Medicine* 2016; 35: 1774–1779.

[25] Barthel FMS, Babiker A, Royston P, et al. Evaluation of sample size and power for multi-arm survival trials allowing for non-uniform accrual, non-proportional hazards, loss to follow-up and cross-over. *Statistics in Medicine* 2006; 25: 2521–2542.

[26] Royston P, Barthel FMS. Projection of power and events in clinical trials with a time-to-event outcome. *Stata Journal* 2010; 10: 386–394.

[27] Heo M, Faith MS, Allison DB. Power and sample size for survival analysis under the Weibull distribution when the whole lifespan is of interest. *Mechanisms of Ageing and Development* 1998; 102: 45–53.

[28] Wu J. Power and sample size for randomized Phase III survival trials under the Weibull model. *Journal of Biopharmaceutical Statistics* 2015; 25: 16–28.

[29] Phadnis MA, Wetmore JB, Mayo MS. A clinical trial design using the concept of proportional time using the generalized gamma ratio distribution. *Statistics in Medicine* 2017; 36: 4121–4140.

[30] Hooper R. Versatile sample-size calculation using simulation. *Stata Journal* 2013; 13: 21–38.

[31] Charles P, Giraudeau B, Dechartres A, et al. Reporting of sample size calculation in randomised controlled trials: Review. *BMJ* 2009; 338: b1732.

[32] Bariani GM, Celis Ferrari ACR de, Precivale M, et al. Sample size calculation in oncology trials: Quality of reporting and implications for clinical cancer research. *American Journal of Clinical Oncology* 2015; 38: 570.

[33] Mahmoud KD, Lennon RJ, Holmes DR. Event rates in randomized clinical trials evaluating cardiovascular interventions and devices. *The American Journal of Cardiology* 2015; 116: 355–363.

[34] Moher D, Schulz KF, Altman D, et al. The CONSORT statement: Revised recommendations for improving the quality of reports of parallel-group randomized trials. *JAMA* 2001; 285: 1987–1991.

[35] 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.

[36] Begg C, Cho M, Eastwood S, et al. Improving the quality of reporting of randomized controlled trials: The CONSORT statement. *JAMA* 1996; 276: 637–639.

[37] Altman DG, De Stavola BL, Love SB, et al. Review of survival analyses published in cancer journals. *British Journal of Cancer* 1995; 72: 511–518.

[38] Pocock SJ, Clayton TC, Altman DG. Survival plots of time-to-event outcomes in clinical trials: Good practice and pitfalls. *The Lancet* 2002; 359: 1686–1689.

[39] 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* 2008; 26: 3721–3726.

[40] Batson S, Greenall G, Hudson P. Review of the reporting of survival analyses within randomised controlled trials and the implications for meta-analysis. *PLoS One* 2016; 11: e0154870.

[41] Rahman R, Fell G, Trippa L, et al. Violations of the proportional hazards assumption in randomized Phase III oncology clinical trials. *Journal of Clinical Oncology* 2018; 36: 2543–2543.

[42] Castañon E, Sanchez-Arraez A, Alvarez-Manceñido F, et al. Critical reappraisal of Phase III trials with immune checkpoint inhibitors in non-proportional hazards settings. *European Journal of Cancer* 2020; 136: 159–168.

[43] Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association* 1958; 53: 457–481.

[44] Pocock SJ, Travison TG, Wruck LM. Figures in clinical trial reports: Current practice & scope for improvement. *Trials* 2007; 8: 36.

[45] Jachno K, Heritier S, Wolfe R. Are non-constant rates and non-proportional treatment effects accounted for in the design and analysis of randomised controlled trials? A review of current practice. *BMC Medical Research Methodology* 2019; 19: 103.

[46] Wu J. Sample size calculation for testing differences between cure rates with the optimal log-rank test. *Journal of Biopharmaceutical Statistics* 2017; 27: 124–134.

[47] Zhang D, Quan H. Power and sample size calculation for log-rank test with a time lag in treatment effect. *Statistics in Medicine* 2009; 28: 864–879.

[48] Barthel FMS, Royston P, Babiker A. A menu-driven facility for complex sample size calculation in randomized controlled trials with a survival or a binary outcome: Update. *Stata Journal* 2005; 5: 123–129.

[49] Blenkinsop A, Choodari-Oskooei B. Multiarm, multistage randomized controlled trials with stopping boundaries for efficacy and lack of benefit: An update to nstage. *Stata Journal* 2019; 19: 782–802.

[50] Yang S, Prentice RL. Assessing potentially time-dependent treatment effect from clinical trials and observational studies for survival data, with applications to the women's health initiative combined hormone therapy trial. *Statistics in Medicine* 2015; 34: 1801–1817.

[51] Magirr D, Burman C-F. Modestly weighted logrank tests. *Statistics in Medicine* 2019; 38: 3782–3790.

[52] Jiménez JL, Stalbovskaya V, Jones B. Properties of the weighted log-rank test in the design of confirmatory studies with delayed effects. *Pharmaceutical Statistics* 2019; 18: 287–303.

[53] Royston PB, Parmar MK. A simulation study comparing the power of nine tests of the treatment effect in randomized controlled trials with a time-to-event outcome. *Trials* 2020; 21: 315.

[54] Lin RS, Lin J, Roychoudhury S, et al. Alternative analysis methods for time to event endpoints under nonproportional hazards: A comparative analysis. *Statistics in Biopharmaceutical Research* 2020; 12: 187–198.

[55] Wei LJ. The accelerated failure time model: A useful alternative to the Cox regression model in survival analysis. *Statistics in Medicine* 1992; 11: 1871–1879.

[56] Kay R, Kinnersley N. On the use of the accelerated failure time model as an alternative to the proportional hazards model in the treatment of time to event data: A case study in influenza. *Drug Information Journal* 2002; 36: 571–579.

[57] Swindell WR. Accelerated failure time models provide a useful statistical framework for aging research. *Experimental gerontology* 2009; 44: 190–200.

[58] Royston P, Lambert PC. *Flexible parametric survival analysis using Stata: Beyond the Cox model*. Book, College Station, TX: Stata Press, 2011.

[59] Bartlett JW, Morris TP, Stensrud MJ, et al. The hazards of period specific and weighted hazard ratios. *Statistics in Biopharmaceutical Research* 2020; 12: 518–519.

[60] Hernán MA. The hazards of hazard ratios. *Epidemiology* 2010; 21: 13–15.

[61] Aalen OO, Cook RJ, Røysland K. Does Cox analysis of a randomized survival study yield a causal treatment effect? *Lifetime Data Analysis* 2015; 21: 579–593.

[62] Stensrud MJ, Hernán MA. Why test for proportional hazards? *JAMA* 2020; 323: 1401–1402.

[63] Hernán MA, Robins J. *Causal inference: What if*. Book, Boca Raton: Chapman & Hall/CRC, 2020.

[64] Andersen PK, Hansen MG, Klein JP. Regression analysis of restricted mean survival time based on pseudo-observations. *Lifetime Data Analysis* 2004; 10: 335–350.

[65] Eaton A, Therneau T, Le-Rademacher J. Designing clinical trials with (restricted) mean survival time endpoint: Practical considerations. *Clinical Trials* 2020; 17: 285–294.

[66] *International Conference on Harmonisation of technical requirements for pharmaceuticals for human use. ICH harmonised tripartite guidelines: Statistical principles for clinical trials e9*. Book, London, England: European Medicines Agency 1998, 1998.

[67] *ICH E9 (R1) addendum on estimands and sensitivity analysis in clinical trials to the guideline on statistical principles for clinical trials*. Book, Amsterdam, The Netherlands: European Medicines Agency, 2020.

[68] Shamseer L, Moher D, Clarke M, et al. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015: Elaboration and explanation. *BMJ* 2015; 349: g7647.

[69] Gregson J, Sharples L, Stone GW, et al. Nonproportional hazards for time-to-event outcomes in clinical trials: JACC review topic of the week. *Journal of the American College of Cardiology* 2019; 74: 2102–2112.

[70] Tufte E. *The visual display of quantitative information 2nd ed*. Book, Cheshire, Conn.: Graphics Press, 2001.

[71] Cleveland WS. *The elements of graphing data*. Book, Summit, NJ: AT&T Bell Laboratories, 1994.

[72] Yau N. *Visualize this: The FlowingData guide to design, visualization, and statistics*. Book, Indianapolis, Ind: Wiley Pub, 2011.

[73] Parner ET, Andersen PK. Regression analysis of censored data using pseudo-observations. *Stata Journal* 2010; 10: 408–422.

[74] Overgaard M, Andersen PK, Parner ET. Regression analysis of censored data using pseudo-observations: An update. *Stata Journal* 2015; 15: 809–821.

[75] Morris TP, Jarvis CI, Cragg W, et al. Proposals on Kaplan–Meier plots in medical research and a survey of stakeholder views: KMunicate. *BMJ Open* 2019; 9: e030215.

[76] Hudson HM, Lo SN, Simes RJ, et al. Semiparametric methods for multistate survival models in randomised trials. *Statistics in Medicine*, 2014; 33: 1621–1645.

[77] Maron DJ, Hochman JS, Reynolds HR, et al. Initial invasive or conservative strategy for stable coronary disease. *New England Journal of Medicine*, 2020; 382: 1395–1407.

[78] Rothwell PM, Price JF, Fowkes FGR, et al. Short-term effects of daily aspirin on cancer incidence, mortality, and non-vascular death: Analysis of the time course of risks and benefits in 51 randomised controlled trials. *The Lancet* 2012; 379: 1602–1612.

SUPPLEMENTARY MATERIAL

ARE NON-CONSTANT AND NON-PROPORTIONAL TREATMENT EFFECTS ACCOUNTED FOR IN THE DESIGN AND ANALYSIS OF RANDOMISED CONTROLLED TRIALS? A REVIEW OF CURRENT PRACTICE

Supplementary Table 1: Citation references for the 66 articles included in the review:

Journal of Clinical Oncology

Powles T, Huddart RA, Elliott T, Sarker SJ, Ackerman C, Jones R, Hussain S, Crabb S, Jagdev S, Chester J, Hilman S, Beresford M, Macdonald G, Santhanam S, Frew JA, Stockdale A, Hughes S, Berney D, Chowdhury S. Phase III, Double-Blind, Randomized Trial That Compared Maintenance Lapatinib Versus Placebo After First-Line Chemotherapy in Patients With Human Epidermal Growth Factor Receptor 1/2-Positive Metastatic Bladder Cancer. Pubmed ID 28034079. J Clin Oncol. 2017 Jan;35(1):48-55. doi: 10.1200/JCO.2015.66.3468.

Beer TM, Kwon ED, Drake CG, Fizazi K, Logothetis C, Gravis G, Ganju V, Polikoff J, Saad F, Humanski P, Piulats JM, Gonzalez Mella P, Ng SS, Jaeger D, Parnis FX, Franke FA, Puente J, Carvajal R, Sengeløv L, McHenry MB, Varma A, van den Eertwegh AJ, et al. Randomized, Double-Blind, Phase III Trial of Ipilimumab Versus Placebo in Asymptomatic or Minimally Symptomatic Patients With Metastatic Chemotherapy-Naive Castration-Resistant Prostate Cancer. Pubmed ID 28034081. J Clin Oncol. 2017 Jan;35(1):40-47. doi: 10.1200/JCO.2016.69.1584.

Perez EA, Barrios C, Eiermann W, Toi M, Im YH, Conte P, Martin M, Pienkowski T, Pivot X, Burris H 3rd, Petersen JA, Stanzel S, Strasak A, Patre M, Ellis P. Trastuzumab Emtansine With or Without Pertuzumab Versus Trastuzumab Plus Taxane for Human Epidermal Growth Factor Receptor 2-Positive, Advanced Breast Cancer: Primary Results From the Phase III MARIANNE Study. Pubmed ID 28056202. J Clin Oncol. 2017 Jan 10;35(2):141-148. doi: 10.1200/JCO.2016.67.4887.

Cloughesy T, Finocchiaro G, Belda-Iniesta C, Recht L, Brandes AA, Pineda E, Mikkelsen T, Chinot OL, Balana C, Macdonald DR, Westphal M, Hopkins K, Weller M, Bais C, Sandmann T, Bruey JM, Koeppen H, Liu B, Verret W, Phan SC, Shames DS. Randomized, Double-Blind, Placebo-Controlled, Multicenter Phase II Study of Onartuzumab Plus Bevacizumab Versus Placebo Plus Bevacizumab in Patients With Recurrent Glioblastoma: Efficacy, Safety, and Hepatocyte Growth Factor and O(6)-Methylguanine-DNA Methyltransferase Biomarker Analyses. Pubmed ID 27918718. J Clin Oncol. 2017 Jan 20;35(3):343-351. doi: 10.1200/JCO.2015.64.7685.

Spigel DR, Edelman MJ, O'Byrne K, Paz-Ares L, Mocci S, Phan S, Shames DS, Smith D, Yu W, Paton VE, Mok T. Results From the Phase III Randomized Trial of Onartuzumab Plus Erlotinib Versus Erlotinib in Previously Treated Stage IIIB or IV Non-Small-Cell Lung Cancer: METLung. Pubmed ID 27937096. J Clin Oncol. 2017 Feb;35(4):412-420. doi: 10.1200/JCO.2016.69.2160.

Pigneux A, Béné MC, Guardiola P, Recher C, Hamel JF, Sauvezie M, Harousseau JL, Tournilhac O, Witz F, Berthou C, Escoffre-Barbe M, Guyotat D, Fegueux N, Himberlin C, Hunault M, Delain M, Lioure B, Jourdan E, Bauduer F, Dreyfus F, Cahn JY, Sotto JJ, et al. Addition of Androgens Improves Survival in Elderly Patients With Acute Myeloid Leukemia: A GOELAMS Study. Pubmed ID 28129526. J Clin Oncol. 2017 Feb;35(4):387-393. doi: 10.1200/JCO.2016.67.6213.

van Imhoff GW, McMillan A, Matasar MJ, Radford J, Ardeshna KM, Kuliczkowski K, Kim W, Hong X, Goerloev JS, Davies A, Barrigón MDC, Ogura M, Leppä S, Fennessy M, Liao Q, van der Holt B, Lisby S, Hagenbeek A. Ofatumumab Versus Rituximab Salvage Chemoimmunotherapy in Relapsed or Refractory Diffuse Large B-Cell Lymphoma: The ORCHARRD Study. Pubmed ID 28029326. J Clin Oncol. 2017 Feb 10;35(5):544-551. doi: 10.1200/JCO.2016.69.0198.

Platzbecker U, Avvisati G, Cicconi L, Thiede C, Paoloni F, Vignetti M, Ferrara F, Divona M, Albano F, Efficace F, Fazi P, Sborgia M, Di Bona E, Breccia M, Borlenghi E, Cairoli R, Rambaldi A, Melillo L, La Nasa G, Fiedler W, Brossart P, Hertenstein B, et al. Improved Outcomes With Retinoic Acid and Arsenic Trioxide Compared With Retinoic Acid and Chemotherapy in Non-High-Risk Acute Promyelocytic Leukemia: Final Results of the Randomized Italian-German APL0406 Trial. Pubmed ID 27400939. J Clin Oncol. 2017 Feb 20;35(6):605-612. doi: 10.1200/JCO.2016.67.1982.

Choueiri TK, Halabi S, Sanford BL, Hahn O, Michaelson MD, Walsh MK, Feldman DR, Olencki T, Picus J, Small EJ, Dakhil S, George DJ, Morris MJ. Cabozantinib Versus Sunitinib As Initial Targeted Therapy for Patients With Metastatic Renal Cell Carcinoma of Poor or Intermediate Risk: The Alliance A031203 CABOSUN Trial. Pubmed ID 28199818. J Clin Oncol. 2017 Feb 20;35(6):591-597. doi: 10.1200/JCO.2016.70.7398.

Agarwala SS, Lee SJ, Yip W, Rao UN, Tarhini AA, Cohen GI, Reintgen DS, Evans TL, Brell JM, Albertini MR, Atkins MB, Dakhil SR, Conry RM, Sosman JA, Flaherty LE, Sondak VK, Carson WE, Smylie MG, Pappo AS, Kefford RF, Kirkwood JM. Phase III Randomized Study of 4 Weeks of High-Dose Interferon-Î±-2b in Stage T2bNO, T3a-bNO, T4a-bNO, and T1-4N1a-2a (microscopic) Melanoma: A Trial of the Eastern Cooperative Oncology Group-American College of Radiology Imaging Network Cancer Research Group (E1697). Pubmed ID 28135150. J Clin Oncol. 2017 Mar 10;35(8):885-892. doi: 10.1200/JCO.2016.70.2951.

Smith I, Yardley D, Burris H, De Boer R, Amadori D, McIntyre K, Ejlertsen B, Gnant M, Jonat W, Pritchard KI, Dowsett M, Hart L, Poggio S, Comarella L, Salomon H, Wamil B, O'Shaughnessy J. Comparative Efficacy and Safety of Adjuvant Letrozole Versus Anastrozole in Postmenopausal Patients With Hormone Receptor-Positive, Node-Positive Early Breast Cancer: Final Results of the Randomized Phase III Femara Versus Anastrozole Clinical Evaluation (FACE) Trial. Pubmed ID 28113032. J Clin Oncol. 2017 Apr 1;35(10):1041-1048. doi: 10.1200/JCO.2016.69.2871.

Thomas X, de Botton S, Chevret S, Caillot D, Raffoux E, Lemasle E, Marolleau JP, Berthon C, Pigneux A, Vey N, Reman O, Simon M, Recher C, Cahn JY, Hermine O, Castaigne S, Celli-Lebras K, Ifrah N, Preudhomme C, Terré C, Dombret H. Randomized Phase II Study of Clofarabine-Based Consolidation for Younger Adults With Acute Myeloid Leukemia in First Remission. Pubmed ID 28221862. J Clin Oncol. 2017 Apr 10;35(11):1223-1230. doi: 10.1200/JCO.2016.70.4551.

Scott BL, Pasquini MC, Logan BR, Wu J, Devine SM, Porter DL, Maziarz RT, Warlick ED, Fernandez HF, Alyea EP, Hamadani M, Bashey A, Giralt S, Geller NL, Leifer E, Le-Rademacher J, Mendizabal AM, Horowitz MM, Deeg HJ, Horwitz ME. Myeloablative Versus Reduced-Intensity Hematopoietic Cell Transplantation for Acute Myeloid Leukemia and Myelodysplastic Syndromes. Pubmed ID 28380315. J Clin Oncol. 2017 Apr 10;35(11):1154-1161. doi: 10.1200/JCO.2016.70.7091.

Tiseo M, Boni L, Ambrosio F, Camerini A, Baldini E, Cinieri S, Brighenti M, Zanelli F, Defraia E, Chiari R, Dazzi C, Tibaldi C, Turolla GM, D'Alessandro V, Zilembo N, Trolese AR, Grossi F, Riccardi F, Ardizzoni A. Italian, Multicenter, Phase III, Randomized Study of Cisplatin Plus Etoposide With or Without Bevacizumab as First-Line Treatment in Extensive-Disease Small-Cell Lung Cancer: The GOIRC-AIFA FARM6PMFJM Trial. Pubmed ID 28135143. J Clin Oncol. 2017 Apr 20;35(12):1281-1287. doi: 10.1200/JCO.2016.69.4844.

Seckl MJ, Ottensmeier CH, Cullen M, Schmid P, Ngai Y, Muthukumar D, Thompson J, Harden S, Middleton G, Fife KM, Crosse B, Taylor P, Nash S, Hackshaw A. Multicenter, Phase III, Randomized, Double-Blind, Placebo-Controlled Trial of Pravastatin Added to First-Line Standard Chemotherapy in Small-Cell Lung Cancer (LUNGSTAR). Pubmed ID 28240967. J Clin Oncol. 2017 May 10;35(14):1506-1514. doi: 10.1200/JCO.2016.69.7391.

Mason MD, Clarke NW, James ND, Dearnaley DP, Spears MR, Ritchie AWS, Attard G, Cross W, Jones RJ, Parker CC, Russell JM, Thalmann GN, Schiavone F, Cassoly E, Matheson D, Millman R, Rentsch CA, Barber J, Gilson C, Ibrahim A, Logue J, Lydon A, et al. Adding Celecoxib With or Without Zoledronic Acid for Hormone-Naïve Prostate Cancer: Long-Term Survival Results From an Adaptive, Multiarm, Multistage, Platform, Randomized Controlled Trial. Pubmed ID 28300506. J Clin Oncol. 2017 May 10;35(14):1530-1541. doi: 10.1200/JCO.2016.69.0677.

Bradstock KF, Link E, Di Iulio J, Szer J, Marlton P, Wei AH, Enno A, Schwarer A, Lewis ID, D'Rozario J, Coyle L, Cull G, Campbell P, Leahy MF, Hahn U, Cannell P, Tiley C, Lowenthal RM, Moore J, Cartwright K, Cunningham I, Taper J, et al. Idarubicin Dose Escalation During Consolidation Therapy for Adult Acute Myeloid Leukemia. Pubmed ID 28368672. J Clin Oncol. 2017 May 20;35(15):1678-1685. doi: 10.1200/JCO.2016.70.6374.

Yao JC, Guthrie KA, Moran C, Strosberg JR, Kulke MH, Chan JA, LoConte N, McWilliams RR, Wolin EM, Mattar B, McDonough S, Chen H, Blanke CD, Hochster HS. Phase III Prospective Randomized Comparison Trial of Depot Octreotide Plus Interferon Alfa-2b Versus Depot Octreotide Plus Bevacizumab in Patients With Advanced Carcinoid

Tumors: SWOG S0518. Pubmed ID 28384065. J Clin Oncol. 2017 May 20;35(15):1695-1703. doi: 10.1200/JCO.2016.70.4072.

Jones RJ, Hussain SA, Protheroe AS, Birtle A, Chakraborti P, Huddart RA, Jagdev S, Bahl A, Stockdale A, Sundar S, Crabb SJ, Dixon-Hughes J, Alexander L, Morris A, Kelly C, Stobo J, Paul J, Powles T. Randomized Phase II Study Investigating Pazopanib Versus Weekly Paclitaxel in Relapsed or Progressive Urothelial Cancer. Pubmed ID 28402747. J Clin Oncol. 2017 Jun 1;35(16):1770-1777. doi: 10.1200/JCO.2016.70.7828.

Catton CN, Lukka H, Gu CS, Martin JM, Supiot S, Chung PWM, Bauman GS, Bahary JP, Ahmed S, Cheung P, Tai KH, Wu JS, Parliament MB, Tsakiridis T, Corbett TB, Tang C, Dayes IS, Warde P, Craig TK, Julian JA, Levine MN. Randomized Trial of a Hypofractionated Radiation Regimen for the Treatment of Localized Prostate Cancer. Pubmed ID 28296582. J Clin Oncol. 2017 Jun 10;35(17):1884-1890. doi: 10.1200/JCO.2016.71.7397.

Zucca E, Conconi A, Martinelli G, Bouabdallah R, Tucci A, Vitolo U, Martelli M, Pettengell R, Salles G, Sebban C, Guillermo AL, Pinotti G, Devizzi L, Morschhauser F, Tilly H, Torri V, Hohaus S, Ferreri AJM, Zachée P, Bosly A, Haioun C, Stelitano C, et al. Final Results of the IELSG-19 Randomized Trial of Mucosa-Associated Lymphoid Tissue Lymphoma: Improved Event-Free and Progression-Free Survival With Rituximab Plus Chlorambucil Versus Either Chlorambucil or Rituximab Monotherapy. Pubmed ID 28355112. J Clin Oncol. 2017 Jun 10;35(17):1905-1912.doi:10.1200/JCO.2016.70.6994.

Arcangeli G, Saracino B, Arcangeli S, Gomellini S, Petrongari MG, Sanguineti G, Strigari L. Moderate Hypofractionation in High-Risk, Organ-Confined Prostate Cancer: Final Results of a Phase III Randomized Trial. Pubmed ID 28355113. J Clin Oncol. 2017 Jun 10;35(17):1891-1897. doi: 10.1200/JCO.2016.70.4189.

The Lancet

Bruix J, Qin S, Merle P, Granito A, Huang YH, Bodoky G, Pracht M, Yokosuka O, Rosmorduc O, Breder V, Gerolami R, Masi G, Ross PJ, Song T, Bronowicki JP, Ollivier-Hourmand I, Kudo M, Cheng AL, Llovet JM, Finn RS, LeBerre MA, Baumhauer A, et al. Regorafenib for patients with hepatocellular carcinoma who progressed on sorafenib treatment (RESORCE): a randomised, double-blind, placebo-controlled, phase 3 trial. Pubmed ID 27932229. Lancet. 2017 Jan 7;389(10064):56-66. doi: 10.1016/S0140-6736(16)32453-9.

Rittmeyer A, Barlesi F, Waterkamp D, Park K, Ciardiello F, von Pawel J, Gadgeel SM, Hida T, Kowalski DM, Dols MC, Cortinovis DL, Leach J, Polikoff J, Barrios C, Kabbinavar F, Frontera OA, De Marinis F, Turna H, Lee JS, Ballinger M, Kowanetz M, He P, et al. Atezolizumab versus docetaxel in patients with previously treated non-small-cell lung cancer (OAK): a phase 3, open-label, multicentre randomised controlled trial. Pubmed ID 27979383. Lancet. 2017 Jan 21;389(10066):255-265. doi: 10.1016/S0140-6736(16)32517-X.

Durie BG, Hoering A, Abidi MH, Rajkumar SV, Epstein J, Kahanic SP, Thakuri M, Reu F, Reynolds CM, Sexton R, Orlowski RZ, Barlogie B, Dispenzieri A. Bortezomib with lenalidomide and dexamethasone versus lenalidomide and dexamethasone alone in patients with newly diagnosed myeloma without intent for immediate autologous stem-cell transplant (SWOG S0777): a randomised, open-label, phase 3 trial. Pubmed ID 28017406. Lancet. 2017 Feb 4;389(10068):519-527. doi: 10.1016/S0140-6736(16)31594-X.

Soria JC, Tan DSW, Chiari R, Wu YL, Paz-Ares L, Wolf J, Geater SL, Orlov S, Cortinovis D, Yu CJ, Hochmair M, Cortot AB, Tsai CM, Moro-Sibilot D, Campelo RG, McCulloch T, Sen P, Dugan M, Pantano S, Branle F, Massacesi C, de Castro G Jr. First-line ceritinib versus platinum-based chemotherapy in advanced ALK-rearranged non-small-cell lung cancer (ASCEND-4): a randomised, open-label, phase 3 study. Pubmed ID 28126333. Lancet. 2017 Mar 4;389(10072):917-929. doi: 10.1016/S0140-6736(17)30123-X.

Kepreotes E, Whitehead B, Attia J, Oldmeadow C, Collison A, Searles A, Goddard B, Hilton J, Lee M, Mattes J. High-flow warm humidified oxygen versus standard low-flow nasal cannula oxygen for moderate bronchiolitis (HFWHO RCT): an open, phase 4, randomised controlled trial. Pubmed ID 28161016. Lancet. 2017 Mar 4;389(10072):930-939. doi: 10.1016/S0140-6736(17)30061-2.

Neoptolemos JP, Palmer DH, Ghaneh P, Psarelli EE, Valle JW, Halloran CM, Faluyi O, O'Reilly DA, Cunningham D, Wadsley J, Darby S, Meyer T, Gillmore R, Anthoney A, Lind P, Glimelius B, Falk S, Izbicki JR, Middleton GW, Cummins S, Ross PJ, Wasan H, et al. Comparison of adjuvant gemcitabine and capecitabine with gemcitabine monotherapy in patients with resected pancreatic cancer (ESPAC-4): a multicentre, open-label, randomised, phase 3 trial. Pubmed ID 28129987. Lancet. 2017 Mar 11;389(10073):1011-1024. doi: 10.1016/S0140-6736(16)32409-6.

Cameron D, Piccart-Gebhart MJ, Gelber RD, Procter M, Goldhirsch A, de Azambuja E, Castro G Jr, Untch M, Smith I, Gianni L, Baselga J, Al-Sakaff N, Lauer S, McFadden E, Leyland-Jones B, Bell R, Dowsett M, Jackisch C; Herceptin Adjuvant (HERA) Trial Study Team. 11 years' follow-up of trastuzumab after adjuvant chemotherapy in HER2-positive early breast cancer: final analysis of the HERceptin Adjuvant (HERA) trial. Pubmed ID 28215665. Lancet. 2017 Mar 25;389 (10075):1195-1205. doi: 10.1016/S0140-6736(16)32616-2.

Atkin W, Wooldrage K, Parkin DM, Kralj-Hans I, MacRae E, Shah U, Duffy S, Cross AJ. Long term effects of once-only flexible sigmoidoscopy screening after 17 years of follow-up: the UK Flexible Sigmoidoscopy Screening randomised controlled trial. Pubmed ID 28236467. Lancet. 2017 Apr 1;389(10076):1299-1311. doi: 10.1016/S0140-6736(17)30396-3.

le Roux CW, Astrup A, Fujioka K, Greenway F, Lau DCW, Van Gaal L, Ortiz RV, Wilding JPH, Skjøth TV, Manning LS, Pi-Sunyer X; SCALE Obesity Prediabetes NN8022-1839 Study Group. 3 years of liraglutide versus placebo for type 2 diabetes risk reduction and weight management in individuals with prediabetes: a randomised, double-blind trial. Pubmed ID 28237263. Lancet. 2017 Apr 8;389(10077):1399-1409. doi: 10.1016/S0140-6736(17)30069-7.

Fixation using Alternative Implants for the Treatment of Hip fractures (FAITH) Investigators. Fracture fixation in the operative management of hip fractures (FAITH): an international, multicentre, randomised controlled trial. Pubmed ID 28262269. Lancet. 2017 Apr 15;389(10078):1519-1527. doi: 10.1016/S0140-6736(17)30066-1.

Ohman EM, Roe MT, Steg PG, James SK, Povsic TJ, White J, Rockhold F, Plotnikov A, Mundl H, Strony J, Sun X, Husted S, Tendera M, Montalescot G, Bahit MC, Ardissino D, Bueno H, Claeys MJ, Nicolau JC, Cornel JH, Goto S, Kiss RG, et al. Clinically significant bleeding with low-dose rivaroxaban versus aspirin, in addition to P2Y12 inhibition, in acute coronary syndromes (GEMINI-ACS-1): a double-blind, multicentre, randomised trial. Pubmed ID 28325638. Lancet. 2017 May 6;389(10081):1799-1808. doi: 10.1016/S0140-6736(17)30751-1.

Chan FKL, Ching JYL, Tse YK, Lam K, Wong GLH, Ng SC, Lee V, Au KWL, Cheong PK, Suen BY, Chan H, Kee KM, Lo A, Wong VWS, Wu JCY, Kyaw MH. Gastrointestinal safety of celecoxib versus naproxen in patients with cardiothrombotic diseases and arthritis after upper gastrointestinal bleeding (CONCERN): an industry-independent, double-blind, double-dummy, randomised trial. Pubmed ID 28410791. Lancet. 2017 Jun 17;389(10087):2375-2382. doi: 10.1016/S0140-6736(17)30981-9.

New England Journal of Medicine

Hiatt WR, Fowkes FG, Heizer G, Berger JS, Baumgartner I, Held P, Katona BG, Mahaffey KW, Norgren L, Jones WS, Blomster J, Millegård M, Reist C, Patel MR; EUCLID Trial Steering Committee and Investigators. Ticagrelor versus Clopidogrel in Symptomatic Peripheral Artery Disease. Pubmed ID 27959717. N Engl J Med. 2017 Jan 5;376(1):32-40. doi: 10.1056/NEJMoa1611688.

Strosberg J, El-Haddad G, Wolin E, Hendifar A, Yao J, Chasen B, Mittra E, Kunz PL, Kulke MH, Jacene H, Bushnell D, O'Dorisio TM, Baum RP, Kulkarni HR, Caplin M, Lebtahi R, Hobday T, Delpassand E, Van Cutsem E, Benson A, Srirajaskanthan R, Pavel M, et al. Phase 3 Trial of (177)Lu-Dotatate for Midgut Neuroendocrine Tumors. Pubmed ID 28076709. N Engl J Med. 2017 Jan 12;376(2):125-135. doi: 10.1056/NEJMoa1607427.

Montalban X, Hauser SL, Kappos L, Arnold DL, Bar-Or A, Comi G, de Seze J, Giovannoni G, Hartung HP, Hemmer B, Lublin F, Rammohan KW, Selmaj K, Traboulsee A, Sauter A, Masterman D, Fontoura P, Belachew S, Garren H, Mairon N, Chin P, Wolinsky JS, et al. Ocrelizumab versus Placebo in Primary Progressive Multiple Sclerosis. Pubmed ID 28002688. N Engl J Med. 2017 Jan 19;376(3):209-220. doi: 10.1056/NEJMoa1606468.

Mehra MR, Naka Y, Uriel N, Goldstein DJ, Cleveland JC Jr, Colombo PC, Walsh MN, Milano CA, Patel CB, Jorde UP, Pagani FD, Aaronson KD, Dean DA, McCants K, Itoh A, Ewald GA, Horstmanshof D, Long JW, Salerno C; MOMENTUM 3 Investigators. A Fully Magnetically Levitated Circulatory Pump for Advanced Heart Failure. Pubmed ID 27959709. N Engl J Med. 2017 Feb 2;376(5):440-450. doi: 10.1056/NEJMoa1610426.

Rogers JG, Pagani FD, Tatooles AJ, Bhat G, Slaughter MS, Birks EJ, Boyce SW, Najjar SS, Jeevanandam V, Anderson AS, Gregoric ID, Mallidi H, Leadley K, Aaronson KD, Frazier OH, Milano CA. Intrapericardial Left Ventricular Assist Device for Advanced Heart Failure. Pubmed ID 28146651. N Engl J Med. 2017 Feb 2;376(5):451-460. doi: 10.1056/NEJMoa1602954.

Shipley WU, Seiferheld W, Lukka HR, Major PP, Heney NM, Grignon DJ, Sartor O, Patel MP, Bahary JP, Zietman AL, Pisansky TM, Zeitzer KL, Lawton CA, Feng FY, Lovett RD, Balogh AG, Souhami L, Rosenthal SA, Kerlin KJ, Dignam JJ, Pugh SL, Sandler HM, et al. Radiation with or without Antiandrogen Therapy in Recurrent Prostate Cancer. Pubmed ID 28146658. N Engl J Med. 2017 Feb 2;376(5):417-428. doi: 10.1056/NEJMoa1607529.

Mok TS, Wu Y-L, Ahn M-J, Garassino MC, Kim HR, Ramalingam SS, Shepherd FA, He Y, Akamatsu H, Theelen WS, Lee CK, Sebastian M, Templeton A, Mann H, Marotti M, Ghiorghiu S, Papadimitrakopoulou VA; AURA3 Investigators. Osimertinib or Platinum-Pemetrexed in EGFR T790M-Positive Lung Cancer. Pubmed ID 27959700. N Engl J Med. 2017 Feb 16;376(7):629-640. doi: 10.1056/NEJMoa1612674.

Agus MS, Wypij D, Hirshberg EL, Srinivasan V, Faustino EV, Luckett PM, Alexander JL, Asaro LA, Curley MA, Steil GM, Nadkarni VM; HALF-PINT Study Investigators and the PALISI Network. Tight Glycemic Control in Critically Ill Children. Pubmed ID 28118549. N Engl J Med. 2017 Feb 23;376(8):729-741. doi: 10.1056/NEJMoa1612348.

Kantarjian H, Stein A, Gökbuget N, Fielding AK, Schuh AC, Ribera JM, Wei A, Dombret H, Foà R, Bassan R, Arslan Önder, Sanz MA, Bergeron J, Demirkan F, Lech-Maranda E, Rambaldi A, Thomas X, Horst HA, Brüggemann M, Klapper W, Wood BL, Fleishman A, et al. Blinatumomab versus Chemotherapy for Advanced Acute Lymphoblastic Leukemia. Pubmed ID 28249141. N Engl J Med. 2017 Mar 2;376(9):836-847. doi: 10.1056/NEJMoa1609783.

Bellmunt J, de Wit R, Vaughn DJ, Fradet Y, Lee JL, Fong L, Vogelzang NJ, Climent MA, Petrylak DP, Choueiri TK, Necchi A, Gerritsen W, Gurney H, Quinn DI, Culine S, Sternberg CN, Mai Y, Poehlein CH, Perini RF, Bajorin DF; KEYNOTE-045 Investigators. Pembrolizumab as Second-Line Therapy for Advanced Urothelial Carcinoma. Pubmed ID 28212060. N Engl J Med. 2017 Mar 16;376(11):1015-1026. doi: 10.1056/NEJMoa1613683.

Perry JR, Laperriere N, O'Callaghan CJ, Brandes AA, Menten J, Phillips C, Fay M, Nishikawa R, Cairncross JG, Roa W, Osoba D, Rossiter JP, Sahgal A, Hirte H, Laigle-Donadey F, Franceschi E, Chinot O, Golfinopoulos V, Fariselli L, Wick A, Feuvret L, Back M, et al. Short-Course Radiation plus Temozolomide in Elderly Patients with Glioblastoma. Pubmed ID 28296618. N Engl J Med. 2017 Mar 16;376(11):1027-1037. doi: 10.1056/NEJMoa1611977.

Weitz JI, Lensing AWA, Prins MH, Bauersachs R, Beyer-Westendorf J, Bounameaux H, Brighton TA, Cohen AT, Davidson BL, Decousus H, Freitas MCS, Holberg G, Kakkar AK, Haskell L, van Bellen B, Pap AF, Berkowitz SD, Verhamme P, Wells PS, Prandoni P; EINSTEIN CHOICE Investigators. Rivaroxaban or Aspirin for Extended Treatment of Venous Thromboembolism. Pubmed ID 28316279. N Engl J Med. 2017 Mar 30;376(13):1211-1222. doi: 10.1056/NEJMoa1700518.

Smits PC, Abdel-Wahab M, Neumann FJ, Boxma-de Klerk BM, Lunde K, Schotborgh CE, Piroth Z, Horak D, Wlodarczak A, Ong PJ, Hambrecht R, Angerås O, Richardt G, Omerovic E; Compare-Acute Investigators. Fractional Flow Reserve-Guided Multivessel Angioplasty in Myocardial Infarction. Pubmed ID 28317428. N Engl J Med. 2017 Mar 30;376(13):1234-1244. doi: 10.1056/NEJMoa1701067.

Attal M, Lauwers-Cances V, Hulin C, Leleu X, Caillot D, Escoffre M, Arnulf B, Macro M, Belhadj K, Garderet L, Roussel M, Payen C, Mathiot C, Fermand JP, Meuleman N, Rollet S, Maglio ME, Zeytoonjian AA, Weller EA, Munshi N, Anderson KC, Richardson PG, et al. Lenalidomide, Bortezomib, and Dexamethasone with Transplantation for Myeloma. Pubmed ID 28379796. N Engl J Med. 2017 Apr 6;376(14):1311-1320. doi: 10.1056/NEJMoa1611750.

Ridker PM, Revkin J, Amarenco P, Brunell R, Curto M, Civeira F, Flather M, Glynn RJ, Gregoire J, Jukema JW, Karpov Y, Kastelein JJP, Koenig W, Lorenzatti A, Manga P, Masiukiewicz U, Miller M, Mosterd A, Murin J, Nicolau JC, Nissen

S, Ponikowski P, et al. Cardiovascular Efficacy and Safety of Bococizumab in High-Risk Patients. Pubmed ID 28304242. N Engl J Med. 2017 Apr 20;376(16):1527-1539. doi: 10.1056/NEJMoa1701488.

Ramanan AV, Dick AD, Jones AP, McKay A, Williamson PR, Compeyrot-Lacassagne S, Hardwick B, Hickey H, Hughes D, Woo P, Benton D, Edelsten C, Beresford MW; SYCAMORE Study Group. Adalimumab plus Methotrexate for Uveitis in Juvenile Idiopathic Arthritis. Pubmed ID 28445659. N Engl J Med. 2017 Apr 27;376(17):1637-1646. doi: 10.1056/NEJMoa1614160.

Sabatine MS, Giugliano RP, Keech AC, Honarpour N, Wiviott SD, Murphy SA, Kuder JF, Wang H, Liu T, Wasserman SM, Sever PS, Pedersen TR; FOURIER Steering Committee and Investigators. Evolocumab and Clinical Outcomes in Patients with Cardiovascular Disease. Pubmed ID 28304224. N Engl J Med. 2017 May 4;376(18):1713-1722. doi: 10.1056/NEJMoa1615664.

Packer M, O'Connor C, McMurray JJV, Wittes J, Abraham WT, Anker SD, Dickstein K, Filippatos G, Holcomb R, Krum H, Maggioni AP, Mebazaa A, Peacock WF, Petrie MC, Ponikowski P, Ruschitzka F, van Veldhuisen DJ, Kowarski LS, Schactman M, Holzmeister J; TRUE-AHF Investigators. Effect of Ularitide on Cardiovascular Mortality in Acute Heart Failure. Pubmed ID 28402745. N Engl J Med. 2017 May 18;376(20):1956-1964. doi: 10.1056/NEJMoa1601895.

Lincoff AM, Nicholls SJ, Riesmeyer JS, Barter PJ, Brewer HB, Fox KAA, Gibson CM, Granger C, Menon V, Montalescot G, Rader D, Tall AR, McErlean E, Wolski K, Ruotolo G, Vangerow B, Weerakkody G, Goodman SG, Conde D, McGuire DK, Nicolau JC, Leiva-Pons JL, et al. Evacetrapib and Cardiovascular Outcomes in High-Risk Vascular Disease. Pubmed ID 28514624. N Engl J Med. 2017 May 18;376(20):1933-1942. doi: 10.1056/NEJMoa1609581.

Masuda N, Lee SJ, Ohtani S, Im YH, Lee ES, Yokota I, Kuroi K, Im SA, Park BW, Kim SB, Yanagita Y, Ohno S, Takao S, Aogi K, Iwata H, Jeong J, Kim A, Park KH, Sasano H, Ohashi Y, Toi M. Adjuvant Capecitabine for Breast Cancer after Preoperative Chemotherapy. Pubmed ID 28564564. N Engl J Med. 2017 Jun 1;376(22):2147-2159. doi: 10.1056/NEJMoa1612645.

Faries MB, Thompson JF, Cochran AJ, Andtbacka RH, Mozzillo N, Zager JS, Jahkola T, Bowles TL, Testori A, Beitsch PD, Hoekstra HJ, Moncrieff M, Ingvar C, Wouters MWJM, Sabel MS, Levine EA, Agnese D, Henderson M, Dummer R, Rossi CR, Neves RI, Trocha SD, et al. Completion Dissection or Observation for Sentinel-Node Metastasis in Melanoma. Pubmed ID 28591523. N Engl J Med. 2017 Jun 8;376(23):2211-2222. doi: 10.1056/NEJMoa1613210.

Wykrzykowska JJ, Kraak RP, Hofma SH, van der Schaaf RJ, Arkenbout EK, IJsselmuiden AJ, Elias J, van Dongen IM, Tijssen RYG, Koch KT, Baan J Jr, Vis MM, de Winter RJ, Piek JJ, Tijssen JGP, Henriques JPS; AIDA Investigators. Bioresorbable Scaffolds versus Metallic Stents in Routine PCI Pubmed ID 28402237. N Engl J Med. 2017 Jun 15;376(24):2319-2328. doi: 10.1056/ NEJMoa1614954.

Kraft WK, Adeniyi-Jones SC, Chervoneva I, Greenspan JS, Abatemarco D, Kaltenbach K, Ehrlich ME. Buprenorphine for the Treatment of the Neonatal Abstinence Syndrome Pubmed ID 28468518. N Engl J Med. 2017 Jun 15;376(24):2341-2348. doi: 10.1056/NEJMoa1614835.

Carbone DP, Reck M, Paz-Ares L, Creelan B, Horn L, Steins M, Felip E, van den Heuvel MM, Ciuleanu TE, Badin F, Ready N, Hiltermann TJN, Nair S, Juergens R, Peters S, Minenza E, Wrangle JM, Rodriguez-Abreu D, Borghaei H, Blumenschein GR Jr, Villaruz LC, Havel L, et al. First-Line Nivolumab in Stage IV or Recurrent Non-Small-Cell Lung Cancer. Pubmed ID 28636851. N Engl J Med. 2017 Jun 22;376(25):2415-2426. doi: 10.1056/NEJMoa1613493.

von Minckwitz G, Procter M, de Azambuja E, Zardavas D, Benyunes M, Viale G, Suter T, Arahmani A, Rouchet N, Clark E, Knott A, Lang I, Levy C, Yardley DA, Bines J, Gelber RD, Piccart M, Baselga J; APHINITY Steering Committee and Investigators. Adjuvant Pertuzumab and Trastuzumab in Early HER2-Positive Breast Cancer. Pubmed ID 28581356. N Engl J Med. 2017 Jul 13;377(2)122-131. doi: 10.1056/NEJMoa1703643.

Fizazi K, Tran N, Fein L, Matsubara N, Rodriguez-Antolin A, Alekseev BY, Özgûroğlu M, Ye D, Feyerabend S, Protheroe A, De Porre P, Kheoh T, Park YC, Todd MB, Chi KN; LATITUDE Investigators. Abiraterone plus Prednisone

in Metastatic, Castration-Sensitive Prostate Cancer. Pubmed ID 28578607. N Engl J Med. 2017 Jul 27;377(4):352-360. doi: 10.1056/NEJMoa1704174.

James ND, de Bono JS, Spears MR, Clarke NW, Mason MD, Dearnaley DP, Ritchie AWS, Amos CL, Gilson C, Jones RJ, Matheson D, Millman R, Attard G, Chowdhury S, Cross WR, Gillessen S, Parker CC, Russell JM, Berthold DR, Brawley C, Adab F, Aung S, et al. Abiraterone for Prostate Cancer Not Previously Treated with Hormone Therapy. Pubmed ID 28578639. N Engl J Med. 2017 Jul 27;377(4):338-351. doi: 10.1056/NEJMoa1702900.

Stone RM, Mandrekar SJ, Sanford BL, Laumann K, Geyer S, Bloomfield CD, Thiede C, Prior TW, Döhner K, Marcucci G, Lo-Coco F, Klisovic RB, Wei A, Sierra J, Sanz MA, Brandwein JM, de Witte T, Niederwieser D, Appelbaum FR, Medeiros BC, Tallman MS, Krauter J, et al. Midostaurin plus Chemotherapy for Acute Myeloid Leukemia with a FLT3 Mutation. Pubmed ID 28644114. N Engl J Med. 2017 Aug 3;377(5):454-464. doi: 10.1056/NEJMoa1614359.

Robson M, Im SA, Senkus E, Xu B, Domchek SM, Masuda N, Delaloge S, Li W, Tung N, Armstrong A, Wu W, Goessl C, Runswick S, Conte P. Olaparib for Metastatic Breast Cancer in Patients with a Germline BRCA Mutation. Pubmed ID 28578601. N Engl J Med. 2017 Aug 10;377(6):523-533. doi: 10.1056/NEJMoa1706450.

Neal B, Perkovic V, Mahaffey KW, de Zeeuw D, Fulcher G, Erondu N, Shaw W, Law G, Desai M, Matthews DR; CANVAS Program Collaborative Group. Canagliflozin and Cardiovascular and Renal Events in Type 2 Diabetes. Pubmed ID 28605608. N Engl J Med. 2017 Aug 17;377(7):644-657. doi: 10.1056/NEJMoa1611925.

Marso SP, McGuire DK, Zinman B, Poulter NR, Emerson SS, Pieber TR, Pratley RE, Haahr PM, Lange M, Brown-Frandsen K, Moses A, Skibsted S, Kvist K, Buse JB; DEVOTE Study Group.. Efficacy and Safety of Degludec versus Glargine in Type 2 Diabetes. Pubmed ID 28605603. N Engl J Med. 2017 Aug 24;377(8):723-732. doi: 10.1056/NEJMoa1615692.

Peters S, Camidge DR, Shaw AT, Gadgeel S, Ahn JS, Kim DW, Ou SI, Pérol M, Dziadziuszko R, Rosell R, Zeaiter A, Mitry E, Golding S, Balas B, Noe J, Morcos PN, Mok T; ALEX Trial Investigators. Alectinib versus Crizotinib in Untreated ALK-Positive Non-Small-Cell Lung Cancer. Pubmed ID 28586279. N Engl J Med. 2017 Aug 31;377(9):829-838. doi: 10.1056/NEJMoa1704795.

Journal	PubmedID	KMcurve	LogrankTe	CoxPH	Regressio	Other	Otherdetail	loglogSurv	logcumH	Schoenfel
NEJM	27959717	y	n	y	n	n		n	y	n
NEJM	28076709	y	y	y	n	n		n	n	n
NEJM	28002688	y	y	y	n	n		n	n	n
NEJM	28146658	y	y	y	n	y	competing risks	n	n	n
NEJM	27959709	y	y	y	n	n		n	n	n
NEJM	28146651	y	y	y	y	n	Weibull model	y	n	n
NEJM	27959700	y	y	y	n	n		n	y	n
NEJM	28118549	y	n	y	n	n		n	n	n
NEJM	28249141	y	y	y	n	n		n	n	n
NEJM	28212060	y	y	y	n	n		y	n	n
NEJM	28296618	y	y	y	n	n		n	n	n
NEJM	28316279	y	n	y	n	n		y	n	y
NEJM	28317428	y	y	y	n	n		n	n	n
NEJM	28379796	y	y	y	n	y	competing risks	n	n	n
NEJM	28304242	y	y	y	n	n		n	n	n
NEJM	28445659	y	y	y	n	n		n	n	n
NEJM	28304224	y	y	y	n	y	landmark analyses	n	n	y
NEJM	28514624	y	y	y	n	n		n	n	n
NEJM	28402745	y	n	y	n	n		n	n	n
NEJM	28564564	y	y	y	n	n		n	n	n
NEJM	28591523	y	y	y	n	y	competing risks (Wei, Lin, Weiss	n	n	n
NEJM	28402237	y	y	y	n	y	landmark analyses	n	n	n
NEJM	28468518	n	n	n	n	y	two-sample van Elteren test (ext			
NEJM	28636851	y	y	y	n	n		n	n	n
NEJM	28578639	y	y	y	y	y	restricted cubic splines, competi	n	n	n
NEJM	28578601	y	y	y	n	n		y	n	n
NEJM	28578607	y	y	y	n	n		y	n	n
NEJM	28581356	y	y	y	n	n		n	n	n
NEJM	28586279	y	y	y	n	y	competing risks	n	n	n
NEJM	28605603	y	n	y	n	n		n	n	n
NEJM	28605608	y	y	y	y	n	AFT models mentioned but no re	n	n	n
NEJM	28644114	y	y	y	n	y	competing risks, landmark analy	n	n	n
Lancet	27932229	y	y	y	n	n		n	n	n
Lancet	27979383	y	y	y	n	n		n	n	n
Lancet	28017406	y	y	y	n	n		n	n	n
Lancet	28161016	y	y	y	n	n		n	n	n
Lancet	28126333	y	y	y	n	n		n	n	n
Lancet	28129987	y	y	y	n	n		n	n	y
Lancet	28215665	y	y	y	n	y	competing risks, landmark analy	n	n	n
Lancet	28236467	y	y	y	y	y	segmented Poisson, stratified C	n	n	n
Lancet	28237263	y	n	y	y	n	Weibull regression	n	n	n
Lancet	28262269	y	n	y	n	n		n	n	n
Lancet	28325638	y	y	y	n	n	landmark analyses	n	n	n
Lancet	28410791	y	y	y	n	n		n	n	n
JCO	28034081	y	y	y	n	n		n	n	n
JCO	28034079	y	y	y	n	n		n	n	n
JCO	28056202	y	y	y	n	n		n	n	n
JCO	27918718	y	y	y	n	n		n	n	n
JCO	28129526	y	y	y	n	n	landmark analyses - figures 2B a	n	n	y
JCO	27937096	y	y	y	n	n		n	n	n
JCO	28029326	y	y	y	n	n		n	n	n
JCO	28199818	y	y	y	n	n		n	n	n
JCO	27400939	y	y	y	n	y	competing risks	n	n	n
JCO	28135150	y	y	y	n	n		n	n	n
JCO	28113032	y	y	y	n	n		n	n	n
JCO	28380315	y	y	y	n	y	competing risks	n	n	n
JCO	28221862	y	y	y	n	y	competing risks	n	n	n
JCO	28135143	y	y	y	n	n		n	n	n
JCO	28300506	y	y	y	y	y	parametric survival models, com	n	n	n
JCO	28240967	y	y	y	n	n		n	n	n
JCO	28384065	y	y	y	n	n		n	n	n
JCO	28368672	y	y	y	n	n	landmark in appendix	n	n	n
JCO	28402747	y	y	y	n	n		n	n	n
JCO	28355113	y	y	n	y	y	flexible parmetric PH models (R			
JCO	28355112	y	y	y	n	n		n	n	n
JCO	28296582	y	y	y	n	n		n	n	n

TandG	timeinterac	othPHtest	othPHdeta	ssc_method	prim_meth	prim_result1	prim_method2	prim_result2	npt_aware
n	n	n		logrank	cox	1.02, 0.92, 1.13, 0.65	logrank		yes
n	n	y	unspecifie	exponential	logrank	18%, 3%, 0.001	cox	0.21, 0.13, 0.33	yes
n	n	n		logrank	cox	0.76, 0.59, 0.98, 0.03	logrank	29.6%, 35.7%	no
n	n	n		logrank	cox	0.77, 0.59, 0.99, 0.04	logrank	76.3%, 71.3%	no
n	n	n		proportion	cox	0.55, 0.32, 0.95, 0.04	logrank	.03	no
n	n	n		simulation	logrank	.67	weibull	.01	no
n	y	n		logrank	cox	0.3, 0.23, 0.41, 0.001			yes
n	n	n		exponential	PH regress	19.4,19.4,0.58			no
n	n	n		logrank	cox	0.71,0.55, 0.93, 0.01	logrank	7.7,4.4, 0.009	no
n	n	y	unspecifie	logrank	cox	0.73,0.59,0.91,0.002			yes
n	n	n		logrank	cox	0.67,0.56,0.8, 0.001			no
n	y	n		logrank	cox	0.34, 0.20, 0.59,0.001	cox	0.26,0.14,0.47,	yes
n	n	n		proportion	cox	0.35,0.22,0.55,0.001			no
n	n	n		logrank	cox	0.65,0.53,0.8,0.001			no
n	n	y	graphical s	logrank	cox	0.88,0.76,1.02,0.08			yes
n	y	n		proportion	cox	0.25,0.12,0.49,0.001			yes
n	n	n		exponential	cox	0.85,0.79,0.92,0.001			yes
y	n	y	unspecifie	logrank	cox	1.03,0.93,1.15,0.58			yes
n	n	y	unspecifie	logrank	cox	1.03,0.85,1.25,0.75			yes
n	n	n		logrank	cox	0.7,0.53,0.92,0.01			no
n	n	y	unspecifie	simulation	logrank	.42	cox	1.08,0.88,1.34,	yes
n	n	n		proportion	cox	1.12,0.85,1.48,0.43			no
				two sample ttest	ttest	13,7,21,0.001			na
n	n	n		simulation	cox	1.15,0.91,1.45,0.25			yes
y	y	n		simulation	cox	0.63,0.52,0.76,0.001			yes
n	y	n		logrank	cox	0.58,0.43,0.80,0.001			yes
n	n	n		cox	cox	0.62,0.51,0.76,0.001			yes
n	n	n		logrank	cox	0.81,0.66,1.00,0.045			no
n	y	n		logrank	cox	0.47,0.34,0.65	logrank	.001	yes
n	n	n		cox	cox	0.91,0.78,1.06,0.001			no
n	n	y	unspecifie	logrank	cox	0.86,0.75,0.97,0.02			yes
n	n	n		logrank	logrank	25.6,74.7,0.009			no
n	n	n		logrank	cox	0.63,0.5,0.79,0.0001			no
n	n	n		couldn't determine	cox	0.73,0.62,0.87,0.0003			no
n	n	y	Kolmogoro	logrank	cox	0.712,0.560,0.906,0.0037			yes
n	n	n		logrank	cox	0.9,0.7,1.2,0.61			no
n	n	n		logrank	cox	0.55,0.42,0.73,0.00001			yes
n	n	n		logrank	cox	0.82,0.68,0.98,0.032			yes
n	y	n	interaction	exponential	cox	0.76,0.68,0.86,0.0001			yes
n	t	y	not fully sp	logrank	cox	0.74,0.70,0.80,0.0001			yes
n	n	n		logrank	weibull	0.21,0.13,0.34,0.0001			no
n	n	n		cox	cox	0.83,0.63,1.09,0.18			no
n	n	n		couldn't determine	cox	1.09,0.80,1.50,0.584			no
y	n	n		logrank	logrank	5.6%,12.3%,0.008	cox	0.44,0.23,0.82,	yes
n	y	n		logrank	logrank	.3667	cox	1.11, 0.88,1.39,	yes
n	n	n		logrank	cox	1.07,0.81,1.43,0.63			no
n	n	y	acknowled	logrank	cox	0.91,0.73,1.13,0.31			yes
n	n	n		couldn't determine	cox	1.06,0.72,1.56,0.74			no
y	y	n		logrank	logrank	0.654,0.002	km	31.2,22.8,40.0,	yes
n	n	n		couldn't determine	proportion	52,46	cox	1.27,0.98,1.65,	no
n	n	y	piecewise	simulation	cox	1.12,0.89,1.42,0.33			yes
n	n	n		logrank	cox	0.66,0.46,0.95,0.012			no
n	n	y	unspecifie	proportion	logrank	97.3,80,0.001			yes
n	n	n		simulation	logrank	0.7,0.7,0.964	cox	0.98,0.79,1.22	yes
n	n	n		logrank	cox	0.93,0.80,1.07,0.315			no
n	n	y	unspecifie	proportion	proportions	0.677,0.775,0.07			yes
n	n	y	unspecifie	logrank	cox	0.65,0.43,0.98,0.041			yes
n	n	n		logrank	cox	0.78,0.58,1.06,.113			no
n	n	y	unspecifie	simulation	cox	0.98,0.8,1.2,0.847			yes
n	n	n		logrank	cox	1.01,0.88,1.16,0.9			no
n	n	n		logrank	cox	0.93,0.73,1.18,0.55			no
n	n	n		logrank	logrank	47,40,56,35,28,44,.045			no
n	n	n		logrank	cox	1.09,0.85,1.4,.67			no
				logrank	logrank	86,79,.68			
n	n	n		logrank	logrank	.0009	cox	0.54,0.38,0.77	no
n	n	y	ref: the Co	cox	cox	0.96,0.77,1.2			yes

npt_report	date_registered	date_start	date_finish	date_publish	SigEffect	CoxUsed	CoxInferer	PHaware	PHreport	Parametric	Landmark
no	26/11/2012	4/12/2012	26/09/2016	5/01/2017		Yes	Yes	Yes			
no	16/04/2012	1/09/2012	1/07/2015	12/01/2017	Yes	Yes	Yes	Yes			
no	3/10/2010	2/03/2011	23/07/2015	19/01/2017	Yes	Yes	Yes				
no	27/01/2003	1/02/1998	1/08/2015	2/02/2017	Yes	Yes	Yes				
no	25/08/2014	2/09/2014	3/01/2017	2/02/2017	Yes	Yes	Yes				
no	21/07/2010	1/09/2010	1/05/2014	2/02/2017		Yes				Yes	
no	2/06/2014	4/08/2014	15/04/2016	16/02/2017	Yes	Yes	Yes	Yes			
no	29/03/2012	1/04/2012	1/09/2016	23/02/2017		Yes	Yes				
no	17/12/2013	3/01/2014	29/12/2015	2/03/2017	Yes	Yes	Yes				
no	3/10/2014	22/10/2014	7/09/2016	16/03/2017	Yes	Yes	Yes	Yes			
no	5/06/2007	1/05/2007	1/03/2016	16/03/2017	Yes	Yes	Yes				
no	17/02/2014	5/03/2014	22/09/2016	30/03/2017	Yes	Yes	Yes	Yes			
no	22/07/2011	22/07/2011	31/10/2016	30/03/2017	Yes	Yes	Yes				
no	30/08/2010	1/10/2010	1/09/2015	6/04/2017	Yes	Yes	Yes				
no	4/11/2013	29/10/2013	22/03/2017	20/04/2017		Yes	Yes	Yes			
no	24/06/2011	21/10/2011	14/12/2016	27/04/2017	Yes	Yes	Yes	Yes			
no	9/01/2013	8/02/2013	11/11/2016	4/05/2017	Yes	Yes	Yes	Yes			Yes
no	19/09/2012	1/10/2012	12/10/2015	18/05/2017		Yes	Yes	Yes			
no	9/08/2012	1/07/2012	1/11/2015	18/05/2017		Yes	Yes	Yes			
no	6/10/2007	1/02/2007	20/01/2017	1/06/2017	Yes	Yes	Yes				
no	3/01/2006	1/12/2004	9/05/2017	8/06/2017		Yes	Yes	Yes			
no	21/05/2013	1/08/2013	16/05/2017	15/06/2017		Yes	Yes				Yes
na	17/10/2011	1/11/2011	1/06/2017	15/06/2017	Yes						
no	22/01/2014	25/03/2014	1/07/2016	22/06/2017		Yes	Yes	Yes			
yes	22/12/2005	8/07/2005	27/06/2017	27/07/2017	Yes	Yes	Yes	Yes	Yes	Yes	
no	4/12/2013	27/03/2014	9/12/2016	10/08/2017	Yes	Yes	Yes	Yes			
no	26/10/2012	12/02/2013	31/10/2016	27/07/2017	Yes	Yes	Yes	Yes			
no	24/05/2011	8/11/2011	19/12/2016	13/07/2017	Yes	Yes	Yes				
no	3/03/2014	19/08/2014	9/02/2017	31/08/2017		Yes	Yes	Yes			
no	10/10/2013	29/10/2013	16/10/2016	24/08/2017	Yes	Yes	Yes				
no	15/12/2009	9/12/2009	22/02/2017	17/08/2017	Yes	Yes	Yes	Yes		Yes	
no	2/04/2008	2/04/2008	1/07/2016	3/08/2017	Yes	Yes					Yes
no	24/01/2013	14/05/2013	29/02/2016	7/01/2017	Yes	Yes	Yes				
no	11/12/2013	11/03/2014	7/07/2016	21/01/2017	Yes	Yes	Yes				
yes	26/03/2008	1/04/2008	1/07/2016	4/02/2017	Yes	Yes	Yes	Yes	Yes		
no	26/06/2012	16/07/2012	1/05/2015	4/03/2017		Yes	Yes				
no	10/04/2013	9/07/2013	24/06/2016	4/03/2017	Yes	Yes	Yes	Yes			
yes	15/01/2009	13/10/2008	30/01/2017	11/03/2017	Yes	Yes	Yes	Yes	Yes		
yes	27/01/2003	1/11/2001	1/06/2015	25/03/2017	Yes	Yes	Yes	Yes	Yes		Yes
yes	6/04/2000	1/07/1995	31/12/2015	1/04/2017	Yes	Yes	Yes	Yes	Yes	Yes	
no	7/01/2011	1/06/2011	2/03/2015	8/04/2017	Yes	Yes				Yes	
no	30/09/2008	1/03/2009	1/03/2016	15/04/2017		Yes	Yes				
no	18/11/2014	20/04/2015	14/10/2016	6/05/2017		Yes	Yes				Yes
yes	12/09/2005	1/06/2009	1/12/2016	17/06/2017	Yes	Yes	Yes	Yes	Yes		
no	27/01/2010	1/07/2010	1/04/2015	1/01/2017		Yes	Yes	Yes			
no	30/07/2009	1/03/2009	1/07/2015	1/01/2017		Yes					
no	10/05/2010	31/07/2010	16/09/2016	10/01/2017		Yes		Yes			
no	2/07/2012	29/06/2012	21/01/2016	20/01/2017		Yes					
no	18/06/2008	1/06/2002	1/09/2012	1/02/2017	Yes	Yes		Yes			Yes
no	20/10/2011	1/01/2012	1/01/2016	1/02/2017	Yes	Yes					
no	16/11/2009	1/03/2010	1/11/2014	10/02/2017		Yes	Yes	Yes			
no	18/04/2013	8/07/2014	18/04/2016	20/02/2017	Yes	Yes	Yes				
no	5/06/2007	1/08/2007	1/09/2012	20/02/2017	Yes	Yes		Yes			
no	27/01/2003	1/12/1998	4/05/2016	10/03/2017		Yes	Yes	Yes			
no	2/11/2005	1/12/2005	1/09/2014	1/04/2017		Yes	Yes				
no	21/04/2011	1/06/2011	16/01/2017	10/04/2017		Yes		Yes			
no	3/07/2009	1/03/2009	1/04/2016	10/04/2017	Yes	Yes	Yes	Yes			
no	8/10/2008	11/09/2009	31/12/2015	20/04/2017		Yes	Yes				
yes	22/12/2005	1/07/2005	12/09/2016	10/05/2017		Yes	Yes	Yes	Yes	Yes	
no	12/02/2007	1/01/2007	1/11/2013	10/05/2017		Yes	Yes				
no	6/12/2007	1/12/2007	28/04/2016	20/05/2017		Yes	Yes				
no	5/08/2005	1/07/2003	31/07/2013	20/05/2017	Yes	Yes					Yes
no	13/03/2012	1/05/2012	1/07/2015	1/06/2017		Yes	Yes				
		1/01/2002	1/01/2016	10/06/2017						Yes	
no	21/09/2005	1/01/2003	1/04/2017	10/06/2017	Yes	Yes	Yes				
no	20/03/2006	1/05/2006	1/01/2016	10/06/2017		Yes	Yes	Yes			

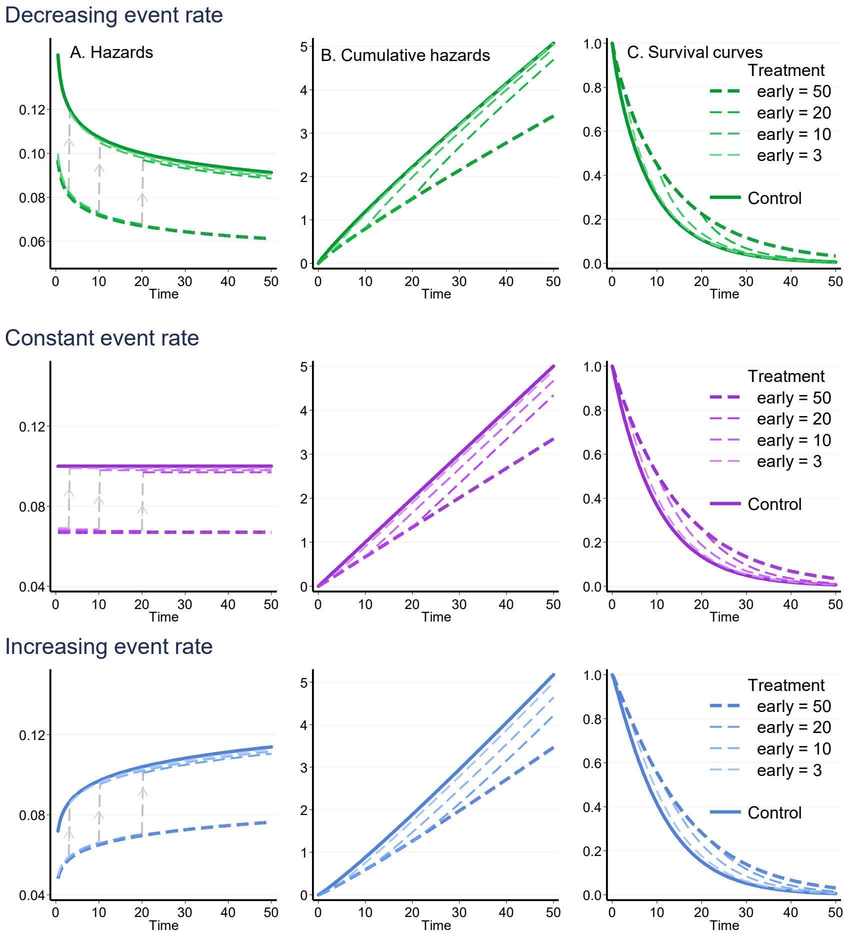
For: Impact of a non-constant baseline hazard on detection of time-dependent treatment effects

Supplementary Information

Supplementary Methods

Data-generating processess for simulation scenarios

Early effect that ceases



Supplementary Figure S1: Baseline hazard function, cumulative hazard curves and survival curves for the early effect that ceases non-PH under three event rates. The control group is indicated by the lightest dashed line and the treatment group is shown by the darkest solid lines. Increasing effect times of 3, 10 and 20 months are indicated by the increased shading and decreased dashed lines. Decreasing, constant and increasing event rate scenarios are indicated by the green, purple and blue lines respectively.

Supplementary Results

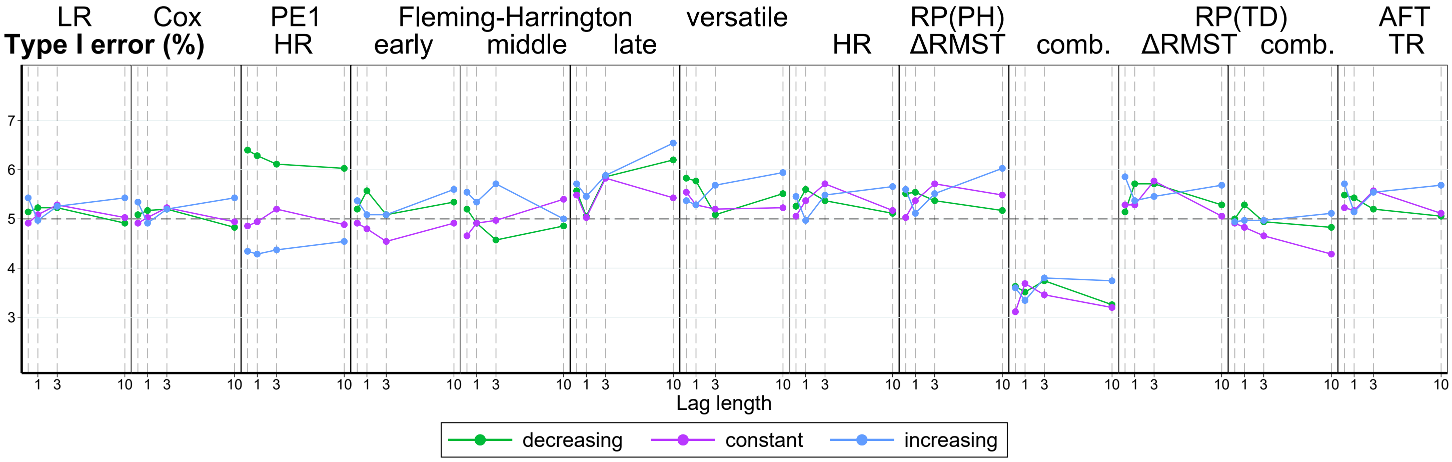
Type I error

When comparing performance measures such as power for the stipulated PH and non-PH scenarios with a known treatment effect, it is important to assess all analytical approaches are controlling the Type I error level at the same or similar nominal value when there is truly no effect. We compared the empirical type I error of the tests of regression-based treatment effect estimate and equal survival function under the null treatment effect by simulation. In these simulations, there was no treatment effect (ie HR =1) in both periods specified by the data- generating models. Pooled replicates for each event rate type and all change point times are presented in Table S1. For the majority of the tests, the empirical Type I errors are within or close to the nominal two-sided 5% significance level. The Type I error of the RP(PH) combined test is conservative under both types of non-PH. A minor Type I error inflation is observed for the FH(0,1) test weighted for late effects (Type I error: 5.7% (95% CI 5.5%, 5.8%)), with even smaller increases in the Type I error above the nominal level also being observed for the versatile test and the regression coefficient estimates for the RP(PH) ΔRMST, the RP(TD)ΔRMST and the AFT TR. Similar minor increases in Type I error rate have been reported in other simulation studies comparing the power of tests for treatment effect under non-PH scenarios.

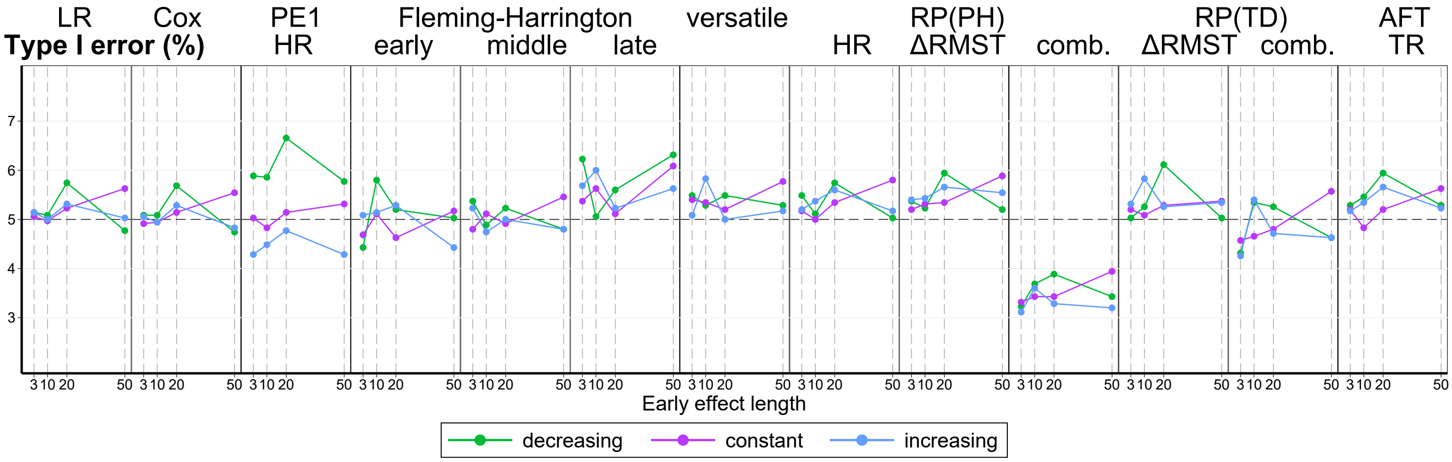
Supplementary Figures S2 and S3 further present the results of this empirical assessment of Type I error by the decreasing, constant and increasing event rate scenarios using the change points from the lag to effect and early effect that ceases non-PH scenario under investigation for each of the analysis methods. Comparisons of the power of the analysis methods presented below needs to be undertaken with the conservative Type I error for the RP(PH) combined test, and minor inflation of the empirical Type I errors for versatile test and the regression coefficient estimates for the RP(PH) ΔRMST, the RP(TD) ΔRMST and the AFT TR in mind.

Table S1: Empirical Type I error (%) of the test of treatment effect or equal survival functions

		Size (95% CI)
Estimands of treatment effect		
Cox Proportional Hazards HR	Cox HR	5.1 (5.0, 5.3)
Piecewise Exponential HR	PE1 HR	5.2 (5.0, 5.3)
Royston-Parmar (PH) HR	RP(PH)	5.3 (5.1, 5.5)
Royston-Parmar (PH) ΔRMST	RP(PH) ΔRMST	5.4 (5.2, 5.6)
Royston-Parmar (TD) ΔRMST	RP(TD) ΔRMST	5.4 (5.3, 5.6)
Accelerated Failure Time model TR	AFT TR	5.4 (5.2, 5.5)
Tests of equal survival functions		
Logrank	LR	5.1 (5.0, 5.3)
Fleming Harrington (1,0) early effects	FH early	5.1 (4.9, 5.2)
Fleming Harrington (1,1) middle effects	FH middle	5.1 (4.9, 5.2)
Fleming Harrington (0,1) late effects	FH late	5.7 (5.5, 5.8)
Versatile test	versatile	5.4 (5.3, 5.6)



Supplementary Figure S2: Empirical type I error of the tests of regression-based treatment effect estimate and equal survival functions under the null treatment effect by event rate for change point times used in the increasing lag until effect data-generating model. Decreasing, constant and increasing event rate scenarios are indicated by the green, purple and blue lines respectively.



Supplementary Figure S3: Empirical type I error of the tests of regression-based treatment effect estimate and equal survival functions under the null treatment effect by event rate for change point times used in the early effect that ceases data-generating model. Decreasing, constant and increasing event rate scenarios are indicated by the green, purple and blue lines respectively.

Lag until treatment effect

Table S2: Summary of event numbers during the inactive and active phases of the treatment effect in the simulation investigating the effect of a lag until treatment effect

Lag until effect	Event rate	Number events inactive phase Mean (range)	Number events active phase Mean (range)	Total number of events (N)
None	Decreasing	N/A	198 (190,202)	198 (190,202)
	Constant	N/A	198 (188,202)	198 (188,202)
	Increasing	N/A	198 (189,202)	198 (189,202)
One	Decreasing	28 (11,46)	170 (151,188)	198 (191,202)
	Constant	19 (6,35)	179 (188,202)	198 (190,202)
	Increasing	14 (2,25)	185 (170,198)	198 (190,202)
Three	Decreasing	67 (46,95)	131 (103,153)	198 (190,202)
	Constant	52 (34,72)	146 (128,167)	198 (189,202)
	Increasing	42 (26,69)	156 (131,172)	198 (190,202)
Ten	Decreasing	141 (118,161)	58 (38, 79)	199 (192,202)
	Constant	128 (107,150)	71 (49, 93)	199 (192,202)
	Increasing	118 (98,143)	81 (56,100)	199 (192,202)

Table S3: Bias (MCSE), % Coverage (MCSE) and % Power (MCSE) for the scenario investigating the lag until treatment effect that ceases for the decreasing baseline hazard

		Hazard Ratio					ΔRMST		Time Ratio
	Effect time	Cox PH	PE1	PE2	LM	RP(PH)	RP(PH)	RP(TD)	Weibull AFT
Bias	Zero	0.01 (0.003)	-0.01 (0.003)	-0.02 (0.004)	0.00 (0.004)	0.00 (0.003)	-0.3 (0.03)	-0.3 (0.03)	0.04 (0.004)
	One	0.06 (0.003)	0.05 (0.003)	-0.01 (0.004)	0.00 (0.004)	0.06 (0.003)	-0.9 (0.03)	-0.9 (0.03)	-0.01 (0.004)
	Three	0.14 (0.003)	0.13 (0.003)	-0.02 (0.004)	0.00 (0.004)	0.14 (0.003)	-1.6 (0.04)	-1.7 (0.04)	-0.09 (0.004)
	Ten	0.29 (0.003)	0.26 (0.004)	0.19 (0.004)	0.22 (0.004)	0.28 (0.003)	-3.3 (0.03)	-3.3 (0.03)	-0.25 (0.004)
Coverage	Zero	94.8 (0.5)	93.8 (0.5)	94.3 (0.5)	95.4 (0.5)	94.6 (0.5)	94.3 (0.5)	94.8 (0.5)	93.8 (0.5)
	One	92.8 (0.6)	92.4 (0.6)	93.9 (0.5)	94.9 (0.5)	92.6 (0.6)	90.2 (0.7)	89.9 (0.7)	95.3 (0.5)
	Three	82.4 (0.9)	83.6 (0.8)	93.4 (0.6)	94.7 (0.5)	82.9 (0.8)	80.4 (0.9)	79.7 (0.9)	90.9 (0.6)
	Ten	48.0 (1.1)	54.4 (1.1)	79.1 (0.9)	75.5 (1.0)	48.7 (1.1)	35.3 (1.1)	35.8 (1.1)	65.2 (1.1)
Power	Zero	77.5 (0.9)	81.1 (0.9)	69.2 (1.0)	64.9 (1.1)	78.8 (0.9)	78.3 (0.9)	78.5 (0.9)	78.9 (0.9)
	One	66.1 (1.1)	70.2 (1.0)	66.2 (1.1)	61.8 (1.1)	67.2 (1.0)	67.2 (1.1)	66.9 (1.1)	66.8 (1.1)
	Three	43.4 (1.1)	50.2 (1.1)	66.8 (1.1)	61.5 (1.1)	44.2 (1.1)	44.7 (1.1)	44.3 (1.1)	46.0 (1.1)
	Ten	13.1 (0.8)	20.0 (0.9)	26.5 (1.0)	17.0 (0.8)	13.9 (0.8)	14.1 (0.8)	14.4 (0.8)	15.6 (0.8)

Table S4: Bias (MCSE), % Coverage (MCSE) and % Power (MCSE) for the scenario investigating the lag until treatment effect that ceases for the constant baseline hazard

		Hazard Ratio					ΔRMST		Time Ratio
	Effect time	Cox PH	PE1	PE2	LM	RP(PH)	RP(PH)	RP(TD)	Weibull AFT
Bias	Zero	0.00 (0.003)	0.00 (0.003)	0.00 (0.004)	0.00 (0.004)	-0.01 (0.003)	-0.1 (0.03)	-0.1 (0.03)	0.00 (0.003)
	One	0.04 (0.003)	0.04 (0.003)	0.01 (0.004)	0.00 (0.004)	0.04 (0.003)	-0.6 (0.04)	-0.6 (0.04)	-0.04 (0.003)
	Three	0.11 (0.003)	0.11 (0.003)	0.00 (0.004)	0.00 (0.004)	0.11 (0.003)	-1.3 (0.04)	-1.4 (0.04)	-0.10 (0.003)
	Ten	0.26 (0.003)	0.25 (0.003)	0.19 (0.004)	0.21 (0.004)	0.26 (0.003)	-3.0 (0.03)	-3.1 (0.03)	-0.24 (0.003)
Coverage	Zero	95.7 (0.5)	95.7 (0.5)	95.4 (0.5)	95.8 (0.4)	95.7 (0.5)	95.2 (0.5)	95.2 (0.5)	95.2 (0.5)
	One	93.4 (0.6)	93.7 (0.5)	95.2 (0.5)	95.3 (0.5)	93.2 (0.6)	92.9 (0.6)	92.8 (0.6)	94.1 (0.5)
	Three	87.3 (0.7)	87.5 (0.7)	95.3 (0.5)	94.8 (0.5)	87.4 (0.7)	85.1 (0.8)	84.2 (0.8)	90.0 (0.7)
	Ten	56.3 (1.1)	59.4 (1.1)	77.5 (0.9)	76.7 (0.9)	56.5 (1.1)	48.6 (1.1)	46.3 (1.1)	62.7 (1.1)
Power	Zero	80.7 (0.9)	81.4 (0.9)	71.3 (1.0)	69.8 (1.0)	81.3 (0.9)	81.3 (0.9)	81.4 (0.9)	81.7 (0.9)
	One	70.9 (1.0)	72.3 (1.0)	69.8 (1.0)	68.4 (1.0)	71.8 (1.0)	72.1 (1.0)	71.6 (1.0)	71.8 (1.0)
	Three	52.0 (1.1)	54.5 (1.1)	68.2 (1.0)	67.2 (1.1)	53.0 (1.1)	53.3 (1.1)	50.8 (1.1)	53.7 (1.1)
	Ten	17.0 (0.8)	21.5 (0.9)	26.4 (1.0)	20.0 (0.9)	17.9 (0.9)	18.0 (0.9)	16.8 (0.8)	20.0 (0.9)

Table S5: Bias (MCSE), % Coverage (MCSE) and % Power (MCSE) for the scenario investigating the lag until treatment effect that ceases for the increasing baseline hazard

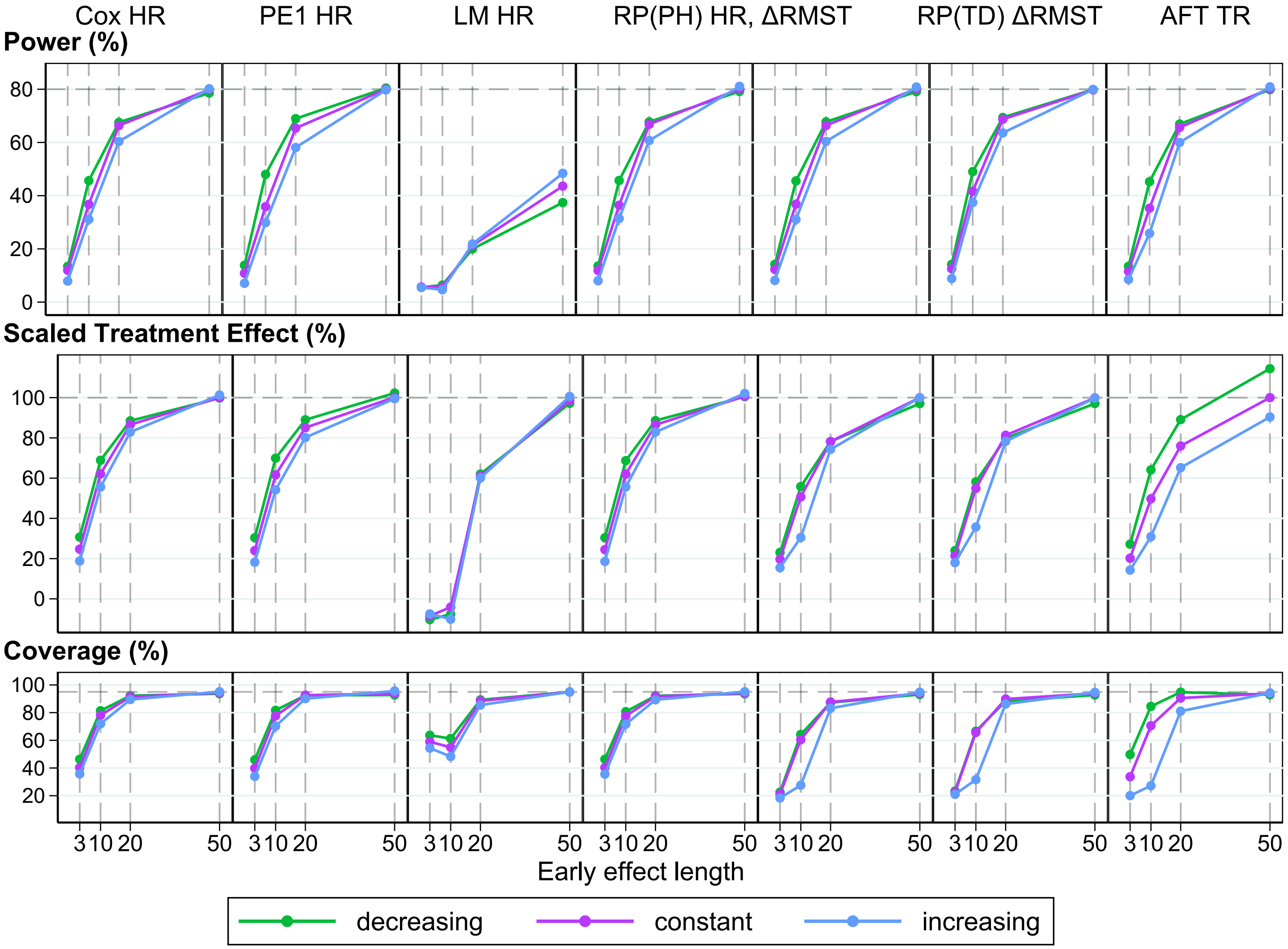
		Hazard Ratio					ΔRMST		Time Ratio
	Effect time	Cox PH	PE1	PE2	LM	RP(PH)	RP(PH)	RP(TD)	Weibull AFT
Bias	Zero	0.01 (0.003)	0.03 (0.003)	0.03 (0.003)	0.01 (0.004)	0.01 (0.003)	-0.3 (0.04)	-0.3 (0.04)	-0.05 (0.003)
	One	0.03 (0.003)	0.05 (0.003)	0.02 (0.003)	0.00 (0.004)	0.03 (0.003)	-0.5 (0.03)	-0.6 (0.03)	-0.06 (0.003)
	Three	0.09 (0.003)	0.10 (0.003)	0.02 (0.003)	0.00 (0.004)	0.08 (0.003)	-1.1 (0.03)	-1.2 (0.04)	-0.10 (0.003)
	Ten	0.25 (0.003)	0.25 (0.003)	0.20 (0.004)	0.21 (0.004)	0.25 (0.003)	-2.9 (0.03)	-3.0 (0.03)	-0.25 (0.003)
Coverage	Zero	94.4 (0.5)	95.1 (0.5)	95.5 (0.5)	94.8 (0.5)	94.3 (0.5)	94.1 (0.5)	93.6 (0.5)	93.1 (0.6)
	One	94.1 (0.5)	94.3 (0.5)	95.4 (0.5)	95.1 (0.5)	94.1 (0.5)	92.9 (0.6)	92.4 (0.6)	92.3 (0.6)
	Three	90.7 (0.7)	89.8 (0.7)	96.1 (0.4)	95.3 (0.5)	90.3 (0.7)	87.8 (0.7)	86.4 (0.8)	88.3 (0.7)
	Ten	57.1 (1.1)	57.8 (1.1)	74.7 (1.0)	74.7 (1.0)	58.3 (1.1)	49.5 (1.1)	46.2 (1.1)	54.3 (1.1)
Power	Zero	78.9 (0.9)	76.9 (0.9)	66.4 (1.1)	69.7 (1.0)	78.9 (0.9)	79.3 (0.9)	79.2 (0.9)	79.8 (0.9)
	One	74.3 (1.0)	72.6 (1.0)	69.5 (1.0)	71.5 (1.0)	74.9 (1.0)	74.9 (1.0)	73.9 (1.0)	75.2 (1.0)
	Three	60.5 (1.1)	58.5 (1.1)	67.6 (1.0)	70.1 (1.0)	62.1 (1.1)	62.2 (1.1)	58.0 (1.1)	62.5 (1.1)
	Ten	18.8 (0.9)	20.4 (0.9)	24.6 (1.0)	21.5 (0.9)	19.4 (0.9)	19.4 (0.9)	18.0 (0.9)	22.0 (0.9)

Supplementary Results cont'd

Early effect that ceases

Power of the regression model approaches

Figure S4 presents the results for the non-PH scenario of an early effect that ceases. Seven different modelling approaches were compared. For the constant hazard event rate scenario, the average number of events during the effective treatment period were 22%, 56% and 82% of the total number of events observed for the early effect times of three, ten and twenty months respectively. For the decreasing hazard event rate, the average number of events during the period when there was an early effect were 28%, 63% and 85%, and for the increasing hazard event rate, the average number of events during the effective period were 18%, 52% and 80% of the total number of events observed for the early effect times of three, ten and twenty units respectively. Supplementary Table S6 presents a summary of event numbers during the active and inactive phases of treatment effect for this early effect that ceases non-PH scenario.



Supplementary Figure S4: The power (%), scaled treatment effect magnitude (%) and coverage (%) are presented as relative to that anticipated at the design stafe of the trial assuming PH. The early effect period lengths investigated were $t_{early} = 3, 10, 20$ and 50 months, with the setting $t_{early} = 50$ representing PH.

When the treatment was constantly effective throughout the follow up period ($t_{early} = 50$) equivalent to a PH data generating model, we observed power at or very close to the design model values of 80% for all estimates of treatment effect except for the LM method. There was substantial decreased power for this period-specific estimate partly due to less than half of the events

Table S6: Summary of event numbers during the active and inactive phases of the treatment effect in the simulation investigating an early treatment effect that ceases

Early effect time	Event rate	Number events active phase Mean (range)	Number events inactive phase Mean (range)	Total number of events (N)
Three	Decreasing	57 (38,78)	143 (121,163)	201 (195,202)
	Constant	45 (28,66)	156 (135,173)	201 (196,202)
	Increasing	36 (15,54)	165 (146,184)	201 (194,202)
Ten	Decreasing	126 (103,148)	75 (53, 95)	198 (191,202)
	Constant	113 (90,138)	87 (64,110)	198 (190,202)
	Increasing	104 (77,128)	97 (73,125)	198 (190,202)
Twenty	Decreasing	168 (147,184)	32 (16,54)	198 (190,202)
	Constant	162 (142,181)	38 (19,57)	198 (189,202)
	Increasing	158 (140,175)	42 (25,61)	198 (190,202)
Fifty	Decreasing	198 (188,202)	N/A	198 (188,202)
	Constant	198 (191,202)	N/A	198 (191,202)
	Increasing	198 (191,202)	N/A	198 (192,202)

being used in the estimation of HR after the prespecified cutpoint of $t_{LM} = 10$ was applied under all event rates, and partly due to the inclusion of more events from the no treatment effect period. For all methods, there was an appreciable loss of power in the early effect non-PH scenario. A decreasing event rate was able to offset the lower power seen as a result of fewer events occurring during the period when the treatment effect had ceased, relative to the number of events observed under a constant event rate. Conversely, the losses in power observed under an increasing event rate in the presence of an early effect that ceases were greater as a result of more events occurring during the period where the treatment had no effect. This pattern of relative power loss was observed for all three estimands.

Scaled Treatment Effect (STE) estimates of regression model approaches

The results comparing the magnitude of treatment effect estimates are presented in the middle panel of Figure S4. For the STE under the PH scenario ($t_{early} = 50$), estimates close to the design model values are obtained for the HR and ΔRMST estimands. Non-constant event rates affect the magnitude of the TR estimated from an AFT model. A decreasing event rate resulted in STEs greater than were observed with a constant event rate, and an increasing event rate resulted in STEs lower than estimated under constant event rates.

Coverage of regression model approaches

Coverage of the estimators for the treatment effect used in the design model is presented in the bottom panel of Figure S4. Under PH, coverage at the design model value of 95% was observed when the treatment effect persisted throughout the analysis period ($t_{early} = 50$). The presence of an early effect that ceases quickly causes a dramatic decrease in the observed coverage. In contrast, having a treatment that stops being effective later has far less impact and most of the nominal coverage is maintained. Non-constant event rates have minimal impact on coverage for the estimates of HR, but for an increasing event rate, estimates of ΔRMST were more affected. The effect of non-constant event rates was most noticeable for the coverage estimates for the TR from an AFT model, consistent with the observed effect of non-constant event rates on the STEs. The summary estimates for bias, coverage and power with the Monte Carlo standard errors (MCSEs) for simulations under this scenario for the decreasing, constant and increasing baseline hazards are presented in Supplementary Tables S7, S8 and S9 respectively.

Table S7: Bias (MCSE), % Coverage (MCSE) and % Power (MCSE) for the scenario investigating the early treatment effect that ceases for the decreasing baseline hazard

		Hazard Ratio					ΔRMST		Time Ratio
	Effect time	Cox PH	PE1	PE2	LM	RP(PH)	RP(PH)	RP(TD)	Weibull AFT
Bias	Three	0.29 (0.003)	0.29 (0.003)	0.41 (0.006)	0.40 (0.006)	0.29 (0.003)	-3.5 (0.03)	-3.5 (0.03)	-0.29 (0.003)
	Ten	0.14 (0.003)	0.14 (0.003)	0.40 (0.006)	0.40 (0.006)	0.14 (0.003)	-2.1 (0.03)	-2.0 (0.03)	-0.14 (0.003)
	Twenty	0.06 (0.003)	0.06 (0.003)	0.16 (0.005)	0.15 (0.006)	0.06 (0.003)	-1.1 (0.03)	-1.1 (0.03)	-0.05 (0.003)
	Fifty	0.01 (0.003)	-0.01 (0.004)	-0.01 (0.006)	0.00 (0.006)	0.00 (0.003)	-0.3 (0.04)	-0.3 (0.04)	0.04 (0.004)
Coverage	Three	45.7 (1.1)	45 (1.1)	62.5 (1.1)	63.5 (1.1)	45.3 (1.1)	22.5 (0.9)	23.4 (0.9)	49.8 (1.1)
	Ten	80.8 (0.9)	81 (0.9)	60.4 (1.1)	60.7 (1.1)	80.1 (0.9)	64.3 (1.1)	66.6 (1.1)	84.5 (0.8)
	Twenty	92.1 (0.6)	92.3 (0.6)	88.9 (0.7)	89.1 (0.7)	91.8 (0.6)	87.3 (0.7)	88.3 (0.7)	94.8 (0.5)
	Fifty	93.7 (0.5)	92.7 (0.6)	94.5 (0.5)	95 (0.5)	93.6 (0.5)	93.1 (0.6)	92.8 (0.6)	92.9 (0.6)
Power	Three	13.4 (0.8)	13.9 (0.8)	6.1 (0.5)	5.6 (0.5)	13.6 (0.8)	14.2 (0.8)	14.2 (0.8)	13.4 (0.8)
	Ten	45.7 (1.1)	48.0 (1.1)	6.6 (0.6)	6.4 (0.5)	45.7 (1.1)	45.5 (1.1)	49.0 (1.1)	45.3 (1.1)
	Twenty	67.6 (1.0)	68.9 (1.0)	17.4 (0.8)	20.0 (0.9)	67.8 (1.0)	67.8 (1.0)	69.4 (1.0)	66.9 (1.1)
	Fifty	78.6 (0.9)	80.4 (0.9)	40.8 (1.1)	37.4 (1.1)	79.2 (0.9)	79.1 (0.9)	79.9 (0.9)	79.8 (0.9)

Table S8: Bias (MCSE), % Coverage (MCSE) and % Power (MCSE) for the scenario investigating the early treatment effect that ceases for the constant baseline hazard

		Hazard Ratio					ΔRMST		Time Ratio
	Effect time	Cox PH	PE1	PE2	LM	RP(PH)	RP(PH)	RP(TD)	Weibull AFT
Bias	Three	0.31 (0.003)	0.32 (0.003)	0.41 (0.005)	0.4 (0.005)	0.31 (0.003)	-3.7 (0.03)	-3.6 (0.03)	-0.32 (0.003)
	Ten	0.17 (0.003)	0.18 (0.003)	0.39 (0.005)	0.39 (0.005)	0.17 (0.003)	-2.3 (0.03)	-2.1 (0.03)	-0.19 (0.003)
	Twenty	0.07 (0.003)	0.08 (0.003)	0.18 (0.005)	0.16 (0.005)	0.07 (0.003)	-1.1 (0.03)	-1.0 (0.03)	-0.09 (0.003)
	Fifty	0.01 (0.003)	0.01 (0.003)	0.00 (0.005)	0.00 (0.005)	0.00 (0.003)	-0.2 (0.04)	-0.2 (0.04)	-0.01 (0.003)
Coverage	Three	39.5 (1.1)	39.2 (1.1)	57.4 (1.1)	58.6 (1.1)	39.7 (1.1)	21.0 (0.9)	22.5 (0.9)	33.7 (1.1)
	Ten	77.4 (0.9)	76.8 (0.9)	54.0 (1.1)	54.5 (1.1)	76.8 (0.9)	60.5 (1.1)	65.7 (1.1)	70.6 (1.0)
	Twenty	91.4 (0.6)	92.4 (0.6)	88.3 (0.7)	88.3 (0.7)	91.3 (0.6)	87.7 (0.7)	89.8 (0.7)	90.5 (0.7)
	Fifty	94.4 (0.5)	93.9 (0.5)	94.8 (0.5)	95.1 (0.5)	94.3 (0.5)	93.8 (0.5)	93.9 (0.5)	93.8 (0.5)
Power	Three	12.0 (0.7)	10.9 (0.7)	6.3 (0.5)	5.7 (0.5)	11.9 (0.7)	12.3 (0.7)	12.6 (0.7)	11.6 (0.7)
	Ten	36.8 (1.1)	35.9 (1.1)	5.7 (0.5)	5.6 (0.5)	36.4 (1.1)	36.8 (1.1)	41.7 (1.1)	35.3 (1.1)
	Twenty	66.3 (1.1)	65.4 (1.1)	18.3 (0.9)	21.3 (0.9)	66.8 (1.1)	66.4 (1.1)	68.8 (1.0)	65.7 (1.1)
	Fifty	79.8 (0.9)	80.0 (0.9)	44.7 (1.1)	43.6 (1.1)	80.1 (0.9)	80.2 (0.9)	79.8 (0.9)	80.3 (0.9)

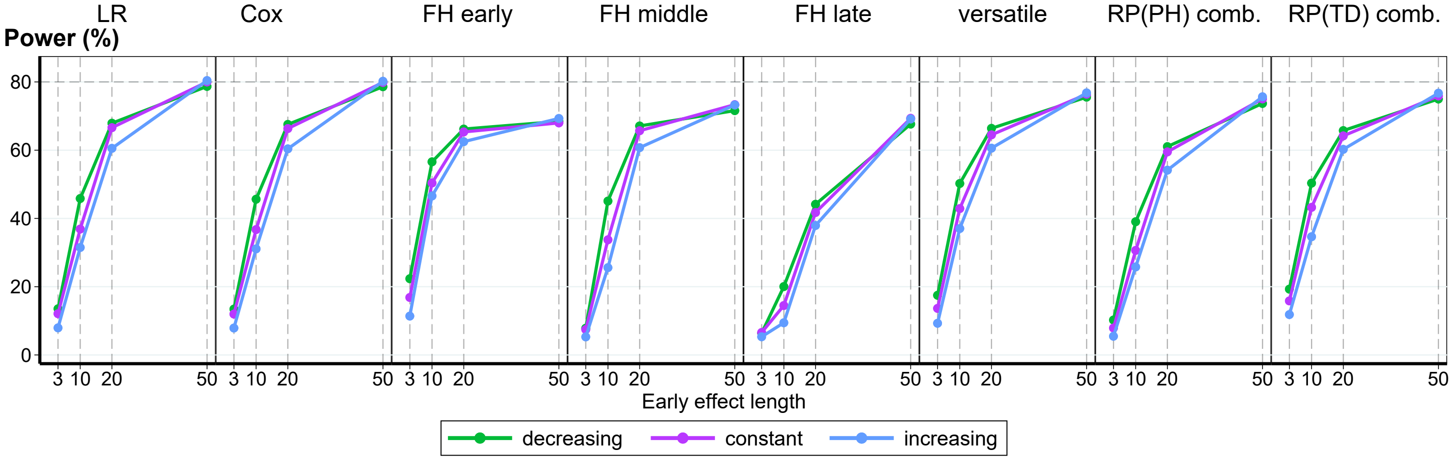
Table S9: Bias (MCSE), % Coverage (MCSE) and % Power (MCSE) for the scenario investigating the early treatment effect that ceases for the increasing baseline hazard

		Hazard Ratio					ΔRMST		Time Ratio
	Effect time	Cox PH	PE1	PE2	LM	RP(PH)	RP(PH)	RP(TD)	Weibull AFT
Bias	Three	0.33 (0.003)	0.34 (0.003)	0.40 (0.005)	0.40 (0.005)	0.33 (0.003)	-3.9 (0.03)	-3.7 (0.03)	-0.34 (0.003)
	Ten	0.20 (0.003)	0.21 (0.003)	0.42 (0.004)	0.42 (0.005)	0.20 (0.003)	-3.2 (0.03)	-3.0 (0.03)	-0.27 (0.002)
	Twenty	0.08 (0.003)	0.10 (0.003)	0.19 (0.004)	0.17 (0.005)	0.08 (0.003)	-1.3 (0.03)	-1.1 (0.03)	-0.13 (0.003)
	Fifty	0.00 (0.003)	0.01 (0.003)	0.01 (0.005)	-0.01 (0.005)	0.00 (0.003)	-0.2 (0.03)	-0.2 (0.03)	-0.04 (0.003)
Coverage	Three	34.8 (1.1)	33.2 (1.1)	53.0 (1.1)	54.1 (1.1)	34.8 (1.1)	18.4 (0.9)	21.0 (0.9)	20.0 (0.9)
	Ten	71.1 (1.0)	69.2 (1.0)	47.5 (1.1)	47.9 (1.1)	71.0 (1.0)	27.5 (1.0)	31.6 (1.0)	27.2 (1.0)
	Twenty	89.1 (0.7)	89.8 (0.7)	86.3 (0.8)	85.2 (0.8)	89.2 (0.7)	83.2 (0.8)	86.2 (0.8)	81.1 (0.9)
	Fifty	95.3 (0.5)	95.6 (0.5)	95.9 (0.4)	95.1 (0.5)	95.2 (0.5)	94.8 (0.5)	94.7 (0.5)	94.3 (0.5)
Power	Three	7.9 (0.6)	7.1 (0.6)	4.6 (0.5)	5.5 (0.5)	8.1 (0.6)	8.1 (0.6)	8.9 (0.6)	8.5 (0.6)
	Ten	31.1 (1.0)	29.9 (1.0)	3.5 (0.4)	4.7 (0.5)	31.4 (1.0)	31.0 (1.0)	37.5 (1.1)	25.9 (1.0)
	Twenty	60.4 (1.1)	58.1 (1.1)	17.6 (0.9)	21.9 (0.9)	60.8 (1.1)	60.4 (1.1)	63.6 (1.1)	60.0 (1.1)
	Fifty	80.2 (0.9)	79.8 (0.9)	47.0 (1.1)	48.4 (1.1)	81.1 (0.9)	80.8 (0.9)	79.9 (0.9)	80.9 (0.9)

Power of the tests of equal survival curves

Supplementary Figure S5 presents the results of investigating the effect of non-constant hazard rates in the presence of an early effect that ceases for seven tests of equal survival functions. In the scenario equivalent to PH, only the LR and Cox tests achieve the power values anticipated under the design model, with the versatile test and the combined tests showing a small decrease in power. Under PH, all three FH tests (using early, middle and late weightings) had lower power than the expected 80%. The FH early test, with weighting emphasising earlier events in the survival curve, obtained the highest power when the treatment was only effective for short initial periods of 3% and 10% of study duration. The versatile test obtained the next highest power in the presence of the shorter effective periods but also had a power value closer to that observed for the LR and Cox tests when the treatment effect length was longer or persisted for the entire follow up. The RP(TD) combined test was closest to the versatile test, with allowing for a time-dependent treatment effect improving the power values slightly at each of the times investigated, relative to the RP(PH) combined test.

In general, the effect of non-constant event rates on the power of tests was consistent with what we observed for the modelling approaches. Decreases in the power loss were observed for a decreasing event rate compared to a constant event rate. An increasing event rate resulted in greater power losses than observed under a constant event rate. Whilst most changes in power observed attributable to a non-constant event rate were relatively modest for this simulation, depending on the test and the length of the effective period under consideration differences in power values ±5% were observed.



Supplementary Figure S5: Effect of non-constant event rates on the power of seven tests of equal survival function. The power of the z-test for the HR treatment effect from the Cox PH model is included in the panel as a comparator. The early effect period lengths investigated were $t_{early} = 3, 10, 20$ and 50 months, with the setting $t_{early} = 50$ representing PH.

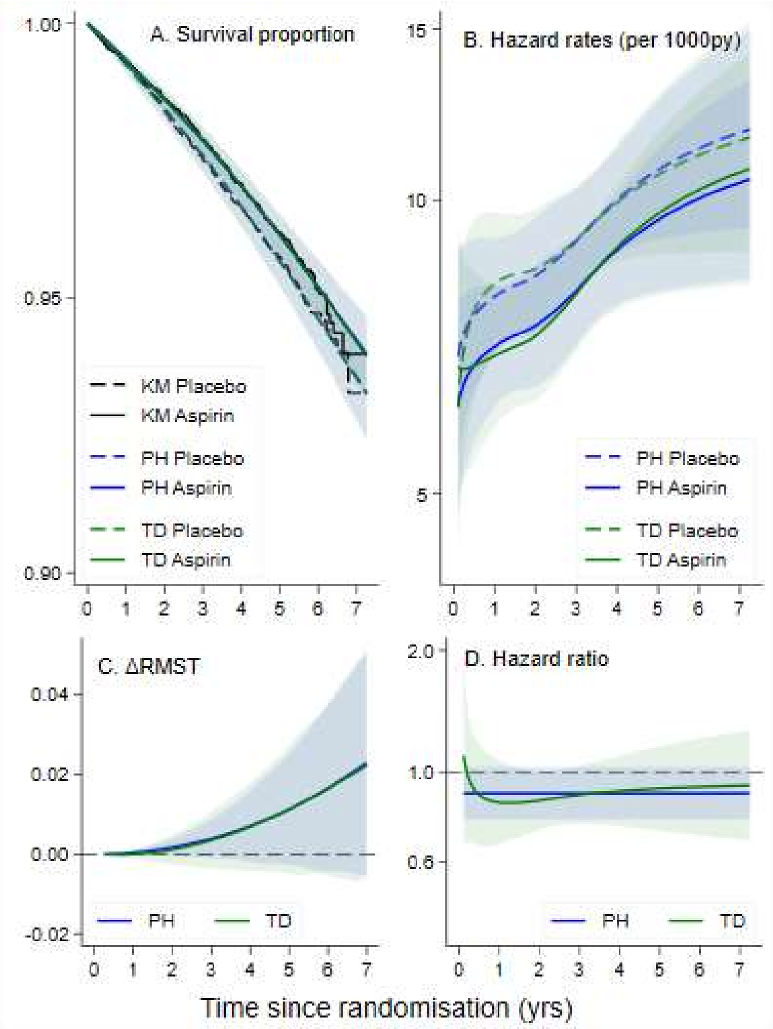
```
/*      Simulations example: lag length

        Three hazard functions scenarios
        *constant     - lambdas(0.1) gammas(1.0)
        *decreasing   - lambdas(0.15) gammas(0.9)
        *increasing   - lambdas(0.07) gammas(1.1)

        Using n=202 for each run
        Decreasing lag to effect; lag time = 3
*/

version 15
set more off
clear

cd "`yourpath'"
*number of simulations
local sims = 100
local trt_eff1 = 0
local trt_eff2 = -0.4

*set seed for reproducibility
set seed 50621

*lag length
local laglen 3

*event rate scenario - decreasing
local lambda_1 0.15
local gamma_1 0.9
local evtype_1 "dec"

local timetot = 50
local time1 = `laglen'
local time2 = `timetot' - `time1'

clear
set obs 202
generate id = _n
*treatment variable, probability of 50% into each arm
generate trt = rbinomial(1,0.5)
*generate a time change
gen tchange1= cond(trt==1, `laglen', 500)

*user-defined hazard function
survsim survtime died, ///
        hazard( (`lambda_`rate'':*`gamma_`rate'':*#t:^(`gamma_`rate'':-1) ) :*      ///
        ( exp( (`trt_eff1':*trt):*(#t:<tchange1) :+                 ///
                (`trt_eff2':*trt):*(#t:>=tchange1) ) ) ) maxtime(50)  nodes(50)
```

```
*number of events pre and post lag time
count if died==1 & survtime <=`laglen'
local n_pre = r(N)
count if died==1 & survtime >`laglen' & survtime<=`timetot'
local n_post = r(N)
*by treatment group
count if died==1 & survtime <=`laglen' & trt==0
local n_pre_c = r(N)
count if died==1 & survtime <=`laglen' & trt==1
local n_pre_t = r(N)
count if died==1 & survtime >`laglen' & survtime<=`timetot' & trt==0
local n_post_c = r(N)
count if died==1 & survtime >`laglen' & survtime<=`timetot' & trt==1
local n_post_t = r(N)

stset survtime, failure(died = 1) id(id)

*Hazard Ratio from a Cox PH model
stcox trt, iterate(200)

*p-value from test of PH using Schoenfeld residuals
estat phtest

*Beta coefficients from a Cox PH model
stcox trt, iterate(200) nohr

*p-value from the logrank test and fleming-harrington tests
*equal weighting on all events
sts test trt, fh(0 0)

*early events weighting
sts test trt, fh(1 0)

*middle events weighting
sts test trt, fh(1 1)

*later events weighting
sts test trt, fh(0 1)


*versatile tests for equal, early or late (Karrison, 2016)
verswlr trt
local ver_std_p = r(pval)

*Cox model with time interaction for t>3 using stsplit
*includes landmark analyses

stsplit time_gt3, at(3)
*landmark analysis at t=3
```

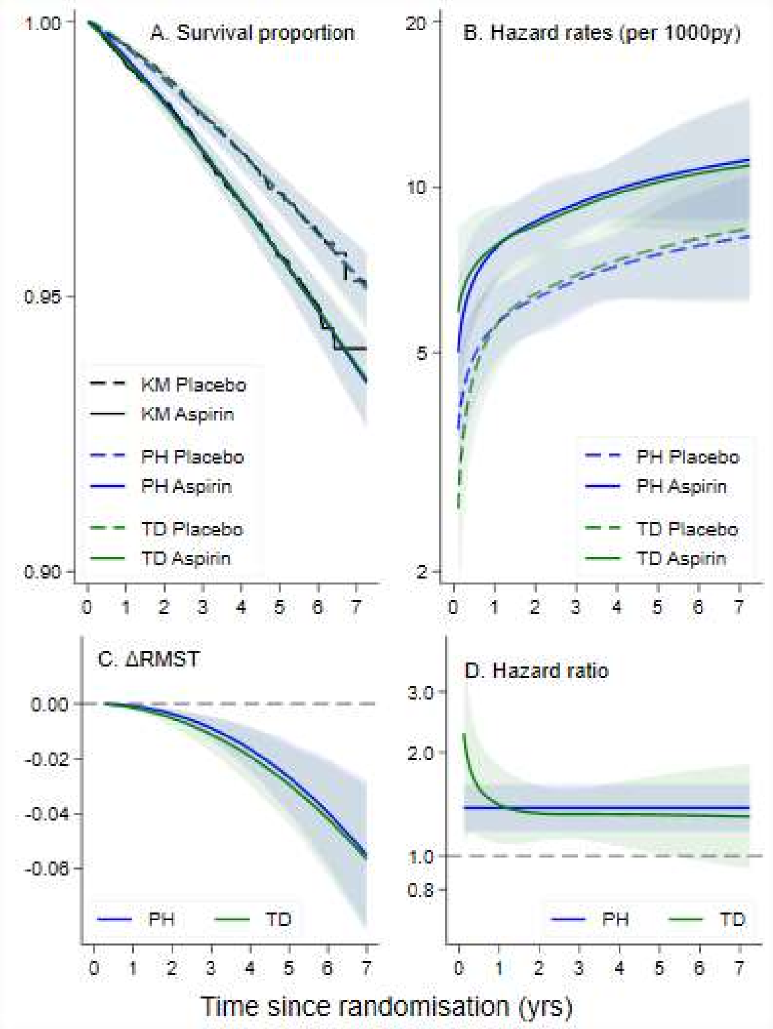
```
*logrank test
sts test trt if time_gt3==3

stcox trt if time_gt3==3

*piecewise exponential model (at t=3)
*one estimate of treatment effect
streg trt ibn.time_gt3, dist(exponential) nocons


*two estimates of treatment effect
streg trt ibn.time_gt3 trt#ibn.time_gt3, dist(exponential) nocons

*estimate of treatment effect in interval time>3 (same as trt output)
lincom _b[_t:trt]  + _b[_t:1.trt#3.time_gt3], hr

*estimate of treatment effect when in interval time<=3
lincom  _b[_t:trt]  + _b[_t:1.trt#0.time_gt3], hr

stjoin

*Weibull shape and scale
streg trt, distribution(weibull) iterate(200)

*Time Ratio from accelerated failure time (AFT) model - Weibull
streg trt, distribution(weibull) time tr iterate(200)


*Estimate HR from RP model with 5 df - PH model parametric (ie no tvc option)
stpm2 trt, scale(hazard) df(5) failconvlininit eform iterate(200)

*Estimate diff in RMST from RP model with 5 df - PH model above
*using t* to be maximum uncensored event time
centile _t if _d==1, centile(100)
predictnl diff = ///
        predict(rmst at(trt 1) tmax(`t') ) - predict(rmst at(trt 0) tmax(`t') ), ///
        se(diff_se) p(diff_p) ci(diff_lci diff_uci)

*using t=50 to be the event time
predictnl diffm = ///
        predict(rmst at(trt 1) tmax(50) ) - predict(rmst at(trt 0) tmax(50) ), ///
        se(diffm_se) p(diffm_p) ci(diffm_lci diffm_uci)

*Royston-Parmar (RP) test for a generalized treatment effect (p-value)
*using default 5df (and as above) and specifying PH model
stctest rp trt, df(5) dftvc(0)

*Estimate HR from RP model with 5 df using dftvc(2) option
stpm2 trt, scale(hazard) df(5) tvc(trt) dftvc(2) failconvlininit eform
```

```
*Estimate diff in RMST from RP model with 5 df using tvc(2) option
*using t* to be maximum uncensored event time
centile _t if _d==1, centile(100)
local t = r(c_1)
predictnl diff_t = ///
        predict(rmst at(trt 1) tmax(`t') ) - predict(rmst at(trt 0) tmax(`t') ), ///
         p(diff_p_t) ci(diff_lci_t diff_uci_t) se(diff_se_t)

*using t=50 to be the event time
predictnl diffm_t = ///
        predict(rmst at(trt 1) tmax(50) ) - predict(rmst at(trt 0) tmax(50) ), ///
         p(diffm_p_t) ci(diffm_lci_t diffm_uci_t) se(diffm_se_t)


*Royston-Parmar (RP) test for a generalized treatment effect (p-value)
*using default 5df and using tvc(2) option
stctest rp trt, df(5) dftvc(2)

exit
```

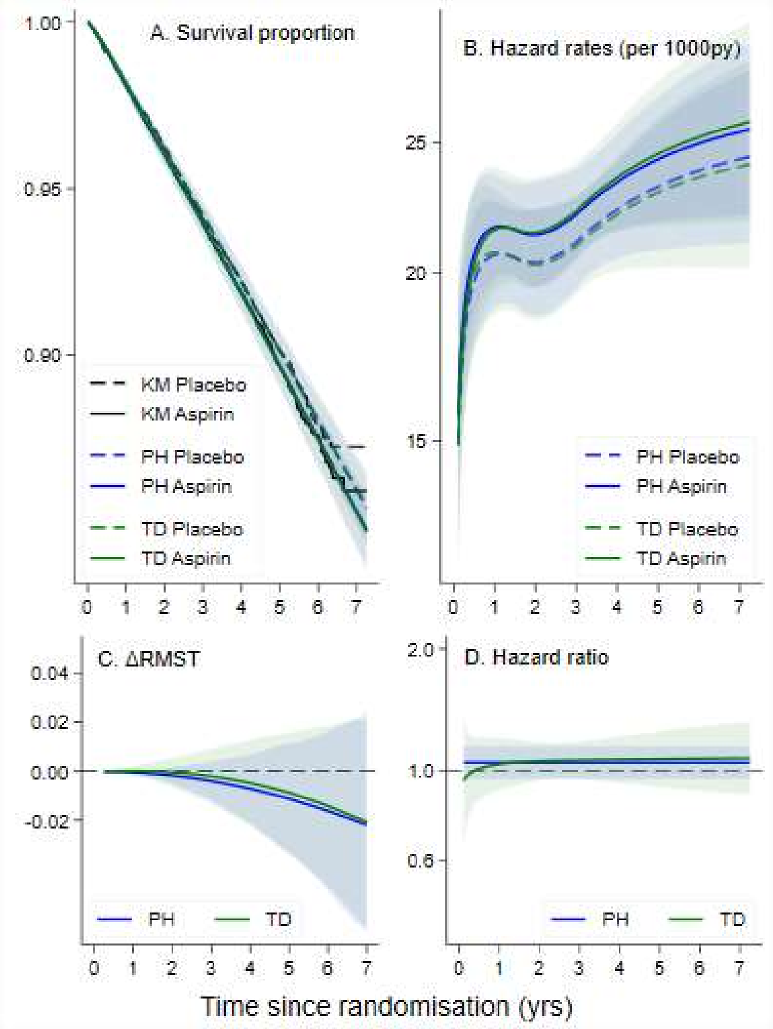
Supplementary material: Examining evidence for time-dependent treatment effects using alternative regression-based methods in clinical trials



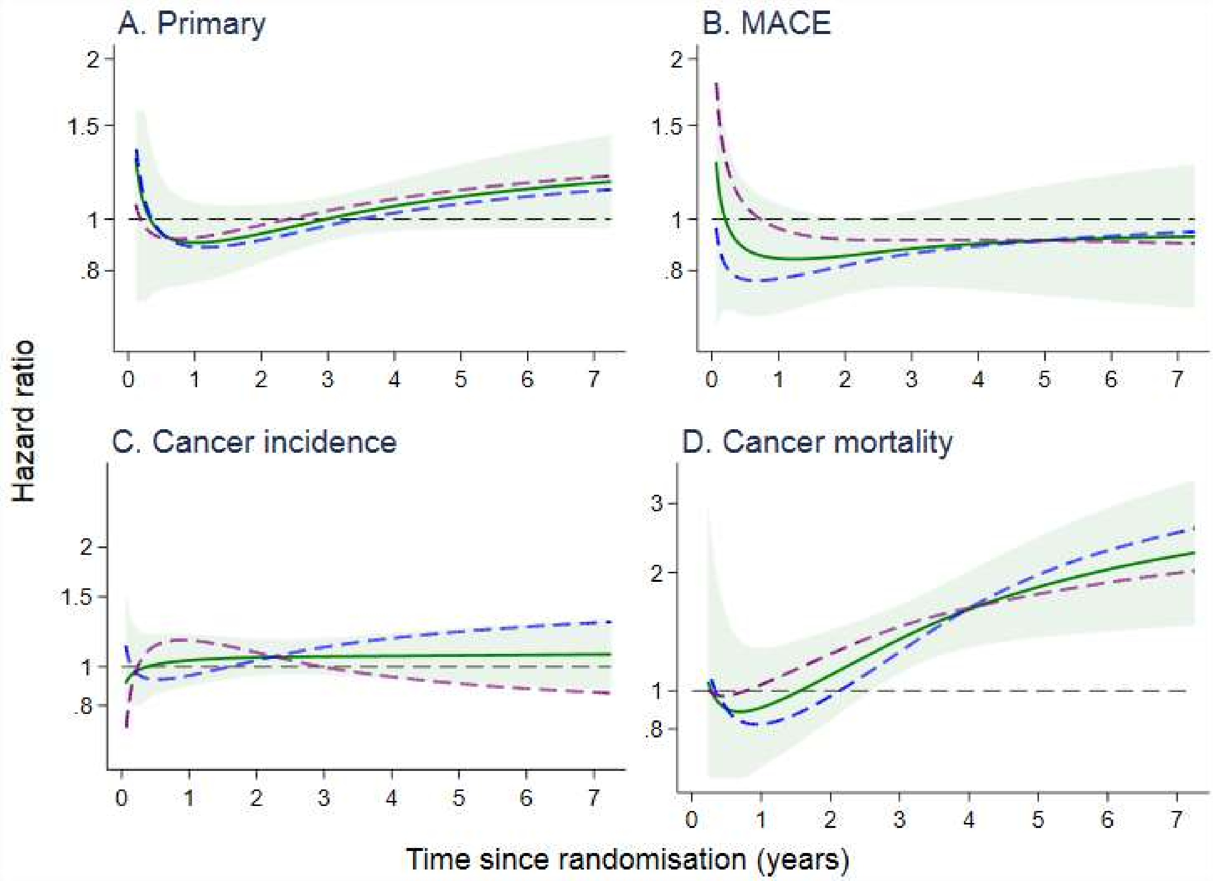
Supplementary Figure S1: Survival curves (panel A) and hazard rates (panel B) by treatment arm, and difference in RMST (ΔRMST; panel C) and HR (panel D) over time from PH and TD analysis models for the MACE endpoint.



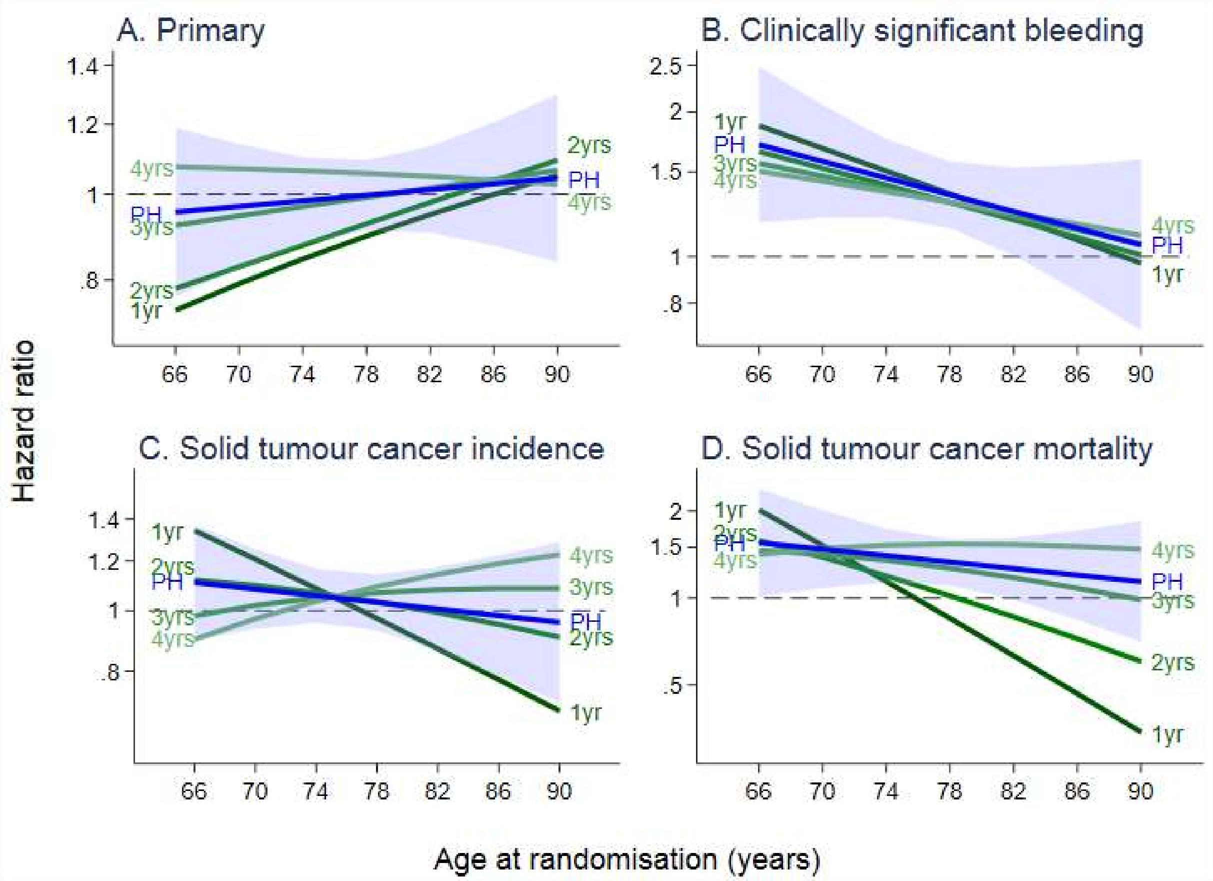
Supplementary Figure S2: Survival curves (panel A) and hazard rates (panel B) by treatment arm, and difference in RMST (ΔRMST; panel C) and HR (panel D) over time from PH and TD analysis models for the clinically significant bleeding endpoint.



Supplementary Figure S3: Survival curves (panel A) and hazard rates (panel B) by treatment arm, and difference in RMST (ΔRMST; panel C) and HR (panel D) over time from PH and TD analysis models for the cancer incidence endpoint.



Supplementary Figure S4: Effect of binary covariate sex on the HR(t) of treatment from TD analysis models for (A) the primary, (B) MACE, (C) cancer incidence and (D) cancer mortality endpoints. The overall estimated HR(t) for treatment effect is the solid green line with the shaded green area indicating the 95% CI width. The HR(t) for treatment effect estimated from females only is indicated by a purple dashed line, and the HR(t) for treatment effect estimated from males only indicated by the blue dashed line.



Supplementary Figure S5: Effect of covariate age at randomisation with treatment from PH and TD analysis models for (A) the primary, (B) clinically significant bleeding, (C) solid tumour cancer incidence and (D) solid tumour cancer mortality endpoints. The estimated age by treatment interaction effect from the PH model is the solid blue line. The interaction treatment effect from the TD model at yearly intervals is indicated by the green lines with color intensity decreasing over time.

Supplementary Table S1: Overall and yearly incremental treatment effect estimates for the primary endpoint. Estimates are from regression-based modelling approaches assuming PH (Cox, Weibull and FPM PH) or allowing for TD treatment effects (FMP TD and pseudo-observations).

	HR (95% CI)			ΔRMST (95% CI)		
	Cox PH	Weibull PH	FPM PH	FPM PH	FPM TD	pseudo-observations
Overall	1.01 (0.92, 1.11)	1.01 (0.92, 1.11)	1.01 (0.92, 1.11)	-0.006 (-0.047, 0.035)	-0.006 (-0.047, 0.035)	-0.006 (-0.047, 0.035)
0-1 years	0.87 (0.65, 1.18)	0.87 (0.65, 1.18)	0.87 (0.65, 1.18)	0.000 (-0.000, 0.001)	0.000 (-0.001, 0.001)	0.000 (-0.001, 0.001)
0-2 years	0.99 (0.83, 1.18)	0.99 (0.83, 1.18)	0.99 (0.83, 1.18)	0.000 (-0.003, 0.004)	0.001 (-0.004, 0.005)	0.001 (-0.004, 0.005)
0-3 years	0.95 (0.83, 1.08)	0.95 (0.83, 1.08)	0.95 (0.83, 1.08)	0.003 (-0.004, 0.010)	0.002 (-0.007, 0.010)	0.001 (-0.008, 0.009)
0-4 years	0.96 (0.86, 1.07)	0.96 (0.86, 1.07)	0.96 (0.86, 1.07)	0.005 (-0.007, 0.017)	0.005 (-0.009, 0.019)	0.005 (-0.010, 0.019)
0-5 years	0.99 (0.89, 1.09)	0.99 (0.89, 1.09)	0.99 (0.89, 1.09)	0.003 (-0.016, 0.022)	0.006 (-0.015, 0.027)	0.006 (-0.015, 0.027)
0-6 years	1.01 (0.92, 1.11)	1.01 (0.92, 1.11)	1.01 (0.92, 1.11)	-0.003 (-0.032, 0.027)	0.002 (-0.027, 0.032)	0.002 (-0.028, 0.032)

Supplementary Table S2: Overall and yearly incremental treatment effect estimates for the clinically significant bleeding endpoint. Estimates are from regression-based modelling approaches assuming PH (Cox, Weibull and FPM PH) or allowing for TD treatment effects (FMP TD and pseudo-observations).

	HR (95% CI)			ΔRMST (95% CI)		
	Cox PH	Weibull PH	FPM PH	FPM PH	FPM TD	pseudo-observations
Overall	1.38 (1.18, 1.62)	1.38 (1.18, 1.62)	1.38 (1.18, 1.62)	-0.050 (-0.075, -0.026)	-0.052 (-0.077, -0.027)	-0.053 (-0.078, -0.028)
0-1 years	1.84 (1.25, 2.70)	1.84 (1.25, 2.70)	1.84 (1.25, 2.70)	-0.002 (-0.003, -0.001)	-0.001 (-0.003, -0.000)	-0.001 (-0.003, -0.000)
0-2 years	1.56 (1.20, 2.04)	1.56 (1.20, 2.04)	1.56 (1.20, 2.04)	-0.005 (-0.008, -0.002)	-0.005 (-0.009, -0.002)	-0.006 (-0.009, -0.002)
0-3 years	1.37 (1.12, 1.68)	1.37 (1.12, 1.68)	1.37 (1.12, 1.68)	-0.008 (-0.014, -0.003)	-0.011 (-0.017, -0.004)	-0.011 (-0.017, -0.004)
0-4 years	1.41 (1.18, 1.69)	1.41 (1.18, 1.69)	1.41 (1.18, 1.69)	-0.018 (-0.027, -0.008)	-0.019 (-0.029, -0.009)	-0.019 (-0.030, -0.009)
0-5 years	1.38 (1.17, 1.63)	1.38 (1.17, 1.63)	1.38 (1.17, 1.63)	-0.027 (-0.040, -0.013)	-0.029 (-0.043, -0.015)	-0.029 (-0.044, -0.014)
0-6 years	1.38 (1.17, 1.62)	1.38 (1.17, 1.62)	1.38 (1.17, 1.62)	-0.039 (-0.059, -0.020)	-0.042 (-0.062, -0.021)	-0.042 (-0.062, -0.021)

Supplementary Table S3: Overall and yearly incremental treatment effect estimates for the major adverse cardiovascular endpoint. Estimates are from regression-based modelling approaches assuming PH (Cox, Weibull and FPM PH) or allowing for TD treatment effects (FMP TD and pseudo-observations).

	HR (95% CI)			ΔRMST (95% CI)		
	Cox PH	Weibull PH	FPM PH	FPM PH	FPM TD	pseudo-observations
Overall	0.89 (0.77, 1.03)	0.89 (0.77, 1.03)	0.89 (0.77, 1.03)	0.021 (-0.005, 0.048)	0.021 (-0.005, 0.048)	0.020 (-0.008, 0.047)
0-1 years	1.07 (0.76, 1.53)	1.07 (0.76, 1.53)	1.07 (0.76, 1.53)	-0.000 (-0.001, 0.001)	-0.001 (-0.002, 0.001)	-0.000 (-0.002, 0.001)
0-2 years	0.86 (0.68, 1.09)	0.86 (0.68, 1.09)	0.86 (0.68, 1.09)	0.002 (-0.001, 0.005)	0.000 (-0.004, 0.004)	0.000 (-0.004, 0.004)
0-3 years	0.85 (0.70, 1.03)	0.85 (0.70, 1.03)	0.85 (0.70, 1.03)	0.005 (-0.001, 0.012)	0.003 (-0.004, 0.010)	0.004 (-0.003, 0.011)
0-4 years	0.87 (0.73, 1.03)	0.87 (0.73, 1.03)	0.87 (0.73, 1.03)	0.008 (-0.002, 0.018)	0.007 (-0.004, 0.018)	0.008 (-0.003, 0.019)
0-5 years	0.88 (0.75, 1.03)	0.88 (0.75, 1.03)	0.88 (0.75, 1.03)	0.012 (-0.002, 0.027)	0.011 (-0.004, 0.027)	0.012 (-0.004, 0.028)
0-6 years	0.88 (0.76, 1.03)	0.88 (0.76, 1.03)	0.88 (0.76, 1.03)	0.017 (-0.003, 0.037)	0.016 (-0.005, 0.037)	0.017 (-0.004, 0.038)

Supplementary Table S4: Overall and yearly incremental treatment effect estimates for the cancer incidence endpoint. Estimates are from regression-based modelling approaches assuming PH (Cox, Weibull and FPM PH) or allowing for TD treatment effects (FMP TD and pseudo-observations).

	HR (95% CI)			ΔRMST (95% CI)		
	Cox PH	Weibull PH	FPM PH	FPM PH	FPM TD	pseudo-observations
Overall	1.05 (0.95, 1.15)	1.05 (0.95, 1.15)	1.05 (0.95, 1.15)	-0.020 (-0.061, 0.021)	-0.018 (-0.061, 0.021)	-0.019 (-0.061, 0.021)
0-1 years	0.99 (0.80, 1.22)	0.99 (0.80, 1.22)	0.99 (0.80, 1.22)	0.000 (-0.002, 0.002)	0.000 (-0.002, 0.002)	0.000 (-0.002, 0.002)
0-2 years	1.06 (0.91, 1.22)	1.06 (0.91, 1.22)	1.06 (0.91, 1.22)	-0.002 (-0.007, 0.003)	-0.001 (-0.007, 0.005)	-0.001 (-0.007, 0.005)
0-3 years	1.03 (0.91, 1.15)	1.03 (0.91, 1.15)	1.03 (0.91, 1.15)	-0.002 (-0.012, 0.008)	-0.002 (-0.013, 0.009)	-0.002 (-0.014, 0.009)
0-4 years	1.04 (0.94, 1.14)	1.04 (0.94, 1.14)	1.04 (0.94, 1.14)	-0.006 (-0.022, 0.010)	-0.005 (-0.022, 0.013)	-0.005 (-0.023, 0.010)
0-5 years	1.04 (0.95, 1.15)	1.04 (0.95, 1.15)	1.04 (0.95, 1.15)	-0.010 (-0.033, 0.013)	-0.009 (-0.034, 0.017)	-0.008 (-0.034, 0.017)
0-6 years	1.04 (0.94, 1.14)	1.04 (0.94, 1.14)	1.04 (0.94, 1.14)	-0.013 (-0.045, 0.019)	-0.012 (-0.045, 0.021)	-0.014 (-0.048, 0.019)

Supplementary Table S5: Overall and yearly incremental treatment effect estimates for the cancer mortality endpoint. Estimates are from regression-based modelling approaches assuming PH (Cox, Weibull and FPM PH) or allowing for TD treatment effects (FPM TD and pseudo-observations).

	HR (95% CI)			ΔRMST (95% CI)		
	Cox PH	Weibull PH	FPM PH	FPM PH	FPM TD	pseudo-observations
Overall	1.36 (1.13, 1.63)	1.36 (1.13, 1.63)	1.36 (1.13, 1.63)	-0.032 (-0.052, 0.013)	-0.029 (-0.048, 0.010)	-0.029 (-0.049, 0.010)
0-1 years	0.90 (0.47, 1.73)	0.90 (0.47, 1.73)	0.90 (0.47, 1.73)	0.000 (-0.001, 0.001)	0.000 (-0.001, 0.001)	0.000 (-0.001, 0.001)
0-2 years	1.04 (0.72, 1.50)	1.04 (0.72, 1.50)	1.04 (0.72, 1.50)	-0.000 (-0.002, 0.002)	0.000 (-0.002, 0.002)	0.000 (-0.002, 0.002)
0-3 years	1.06 (0.82, 1.38)	1.06 (0.82, 1.38)	1.06 (0.82, 1.38)	-0.001 (-0.004, 0.003)	-0.000 (-0.004, 0.004)	-0.000 (-0.005, 0.004)
0-4 years	1.20 (0.96, 1.50)	1.20 (0.96, 1.50)	1.20 (0.96, 1.50)	-0.005 (-0.011, 0.001)	-0.002 (-0.009, 0.005)	-0.002 (-0.009, 0.005)
0-5 years	1.27 (1.04, 1.55)	1.27 (1.04, 1.55)	1.27 (1.04, 1.55)	-0.012 (-0.022, -0.002)	-0.007 (-0.018, 0.003)	-0.007 (-0.018, 0.003)
0-6 years	1.36 (1.13, 1.64)	1.36 (1.13, 1.64)	1.36 (1.13, 1.64)	-0.024 (-0.038, -0.009)	-0.017 (-0.038, -0.003)	-0.018 (-0.033, -0.003)

SUPPLEMENTARY MATERIAL

Complementing the Kaplan-Meier plot to enable assessment of treatment effects consistent with proportional hazards

Citation references for the 65 articles included in the Kaplan-Meier plot recommendations review:

Journal of Clinical Oncology

Powles T, Huddart RA, Elliott T, Sarker SJ, Ackerman C, Jones R, Hussain S, Crabb S, Jagdev S, Chester J, Hilman S, Beresford M, Macdonald G, Santhanam S, Frew JA, Stockdale A, Hughes S, Berney D, Chowdhury S. Phase III, Double-Blind, Randomized Trial That Compared Maintenance Lapatinib Versus Placebo After First-Line Chemotherapy in Patients With Human Epidermal Growth Factor Receptor 1/2-Positive Metastatic Bladder Cancer. Pubmed ID 28034079. J Clin Oncol. 2017 Jan;35(1):48-55. doi: 10.1200/JCO.2015.66.3468.

Beer TM, Kwon ED, Drake CG, Fizazi K, Logothetis C, Gravis G, Ganju V, Polikoff J, Saad F, Humanski P, Piulats JM, Gonzalez Mella P, Ng SS, Jaeger D, Parnis FX, Franke FA, Puente J, Carvajal R, Sengeløv L, McHenry MB, Varma A, van den Eertwegh AJ, et al. Randomized, Double-Blind, Phase III Trial of Ipilimumab Versus Placebo in Asymptomatic or Minimally Symptomatic Patients With Metastatic Chemotherapy-Naive Castration-Resistant Prostate Cancer. Pubmed ID 28034081. J Clin Oncol. 2017 Jan;35(1):40-47. doi: 10.1200/JCO.2016.69.1584.

Perez EA, Barrios C, Eiermann W, Toi M, Im YH, Conte P, Martin M, Pienkowski T, Pivot X, Burris H 3rd, Petersen JA, Stanzel S, Strasak A, Patre M, Ellis P. Trastuzumab Emtansine With or Without Pertuzumab Versus Trastuzumab Plus Taxane for Human Epidermal Growth Factor Receptor 2-Positive, Advanced Breast Cancer: Primary Results From the Phase III MARIANNE Study. Pubmed ID 28056202. J Clin Oncol. 2017 Jan 10;35(2):141-148. doi: 10.1200/JCO.2016.67.4887.

Cloughesy T, Finocchiaro G, Belda-Iniesta C, Recht L, Brandes AA, Pineda E, Mikkelsen T, Chinot OL, Balana C, Macdonald DR, Westphal M, Hopkins K, Weller M, Bais C, Sandmann T, Bruey JM, Koeppen H, Liu B, Verret W, Phan SC, Shames DS. Randomized, Double-Blind, Placebo-Controlled, Multicenter Phase II Study of Onartuzumab Plus Bevacizumab Versus Placebo Plus Bevacizumab in Patients With Recurrent Glioblastoma: Efficacy, Safety, and Hepatocyte Growth Factor and O(6)-Methylguanine-DNA Methyltransferase Biomarker Analyses. Pubmed ID 27918718. J Clin Oncol. 2017 Jan 20;35(3):343-351. doi: 10.1200/JCO.2015.64.7685.

Spigel DR, Edelman MJ, O'Byrne K, Paz-Ares L, Mocci S, Phan S, Shames DS, Smith D, Yu W, Paton VE, Mok T. Results From the Phase III Randomized Trial of Onartuzumab Plus Erlotinib Versus Erlotinib in Previously Treated Stage IIIB or IV Non-Small-Cell Lung Cancer: METLung. Pubmed ID 27937096. J Clin Oncol. 2017 Feb;35(4):412-420. doi: 10.1200/JCO.2016.69.2160.

Pigneux A, Béné MC, Guardiola P, Recher C, Hamel JF, Sauvezie M, Harousseau JL, Tournilhac O, Witz F, Berthou C, Escoffre-Barbe M, Guyotat D, Fegueux N, Himberlin C, Hunault M, Delain M, Lioure B, Jourdan E, Bauduer F, Dreyfus F, Cahn JY, Sotto JJ, et al. Addition of Androgens Improves Survival in Elderly Patients With Acute Myeloid Leukemia: A GOELAMS Study. Pubmed ID 28129526. J Clin Oncol. 2017 Feb;35(4):387-393. doi: 10.1200/JCO.2016.67.6213.

van Imhoff GW, McMillan A, Matasar MJ, Radford J, Ardeshna KM, Kuliczkowski K, Kim W, Hong X, Goerloev JS, Davies A, Barrigón MDC, Ogura M, Leppä S, Fennessy M, Liao Q, van der Holt B, Lisby S, Hagenbeek A. Ofatumumab Versus Rituximab Salvage Chemoimmunotherapy in Relapsed or Refractory Diffuse Large B-Cell Lymphoma: The ORCHARRD Study. Pubmed ID 28029326. J Clin Oncol. 2017 Feb 10;35(5):544-551. doi: 10.1200/JCO.2016.69.0198.

Platzbecker U, Avvisati G, Cicconi L, Thiede C, Paoloni F, Vignetti M, Ferrara F, Divona M, Albano F, Efficace F, Fazi P, Sborgia M, Di Bona E, Breccia M, Borlenghi E, Cairoli R, Rambaldi A, Melillo L, La Nasa G, Fiedler W, Brossart P, Hertenstein B, et al. Improved Outcomes With Retinoic Acid and Arsenic Trioxide Compared With Retinoic Acid and Chemotherapy in Non-High-Risk Acute Promyelocytic Leukemia: Final Results of the Randomized Italian-German APL0406 Trial. Pubmed ID 27400939. J Clin Oncol. 2017 Feb 20;35(6):605-612. doi: 10.1200/JCO.2016.67.1982.

Choueiri TK, Halabi S, Sanford BL, Hahn O, Michaelson MD, Walsh MK, Feldman DR, Olencki T, Picus J, Small EJ, Dakhil S, George DJ, Morris MJ. Cabozantinib Versus Sunitinib As Initial Targeted Therapy for Patients With Metastatic Renal Cell Carcinoma of Poor or Intermediate Risk: The Alliance A031203 CABOSUN Trial. Pubmed ID 28199818. J Clin Oncol. 2017 Feb 20;35(6):591-597. doi: 10.1200/JCO.2016.70.7398.

Agarwala SS, Lee SJ, Yip W, Rao UN, Tarhini AA, Cohen GI, Reintgen DS, Evans TL, Brell JM, Albertini MR, Atkins MB, Dakhil SR, Conry RM, Sosman JA, Flaherty LE, Sondak VK, Carson WE, Smylie MG, Pappo AS, Kefford RF, Kirkwood JM. Phase III Randomized Study of 4 Weeks of High-Dose Interferon-Î±-2b in Stage T2bNO, T3a-bNO, T4a-bNO, and T1-4N1a-2a (microscopic) Melanoma: A Trial of the Eastern Cooperative Oncology Group-American College of Radiology Imaging Network Cancer Research Group (E1697). Pubmed ID 28135150. J Clin Oncol. 2017 Mar 10;35(8):885-892. doi: 10.1200/JCO.2016.70.2951.

Smith I, Yardley D, Burris H, De Boer R, Amadori D, McIntyre K, Ejlertsen B, Gnant M, Jonat W, Pritchard KI, Dowsett M, Hart L, Poggio S, Comarella L, Salomon H, Wamil B, O'Shaughnessy J. Comparative Efficacy and Safety of Adjuvant Letrozole Versus Anastrozole in Postmenopausal Patients With Hormone Receptor-Positive, Node-Positive Early Breast Cancer: Final Results of the Randomized Phase III Femara Versus Anastrozole Clinical Evaluation (FACE) Trial. Pubmed ID 28113032. J Clin Oncol. 2017 Apr 1;35(10):1041-1048. doi: 10.1200/JCO.2016.69.2871.

Thomas X, de Botton S, Chevret S, Caillot D, Raffoux E, Lemasle E, Marolleau JP, Berthon C, Pigneux A, Vey N, Reman O, Simon M, Recher C, Cahn JY, Hermine O, Castaigne S, Celli-Lebras K, Ifrah N, Preudhomme C, Terré C, Dombret H. Randomized Phase II Study of Clofarabine-Based Consolidation for Younger Adults With Acute Myeloid Leukemia in First Remission. Pubmed ID 28221862. J Clin Oncol. 2017 Apr 10;35(11):1223-1230. doi: 10.1200/JCO.2016.70.4551.

Scott BL, Pasquini MC, Logan BR, Wu J, Devine SM, Porter DL, Maziarz RT, Warlick ED, Fernandez HF, Alyea EP, Hamadani M, Bashey A, Giralt S, Geller NL, Leifer E, Le-Rademacher J, Mendizabal AM, Horowitz MM, Deeg HJ, Horwitz ME. Myeloablative Versus Reduced-Intensity Hematopoietic Cell Transplantation for Acute Myeloid Leukemia and Myelodysplastic Syndromes. Pubmed ID 28380315. J Clin Oncol. 2017 Apr 10;35(11):1154-1161. doi: 10.1200/JCO.2016.70.7091.

Tiseo M, Boni L, Ambrosio F, Camerini A, Baldini E, Cinieri S, Brighenti M, Zanelli F, Defraia E, Chiari R, Dazzi C, Tibaldi C, Turolla GM, D'Alessandro V, Zilembo N, Trolese AR, Grossi F, Riccardi F, Ardizzoni A. Italian, Multicenter, Phase III, Randomized Study of Cisplatin Plus Etoposide With or Without Bevacizumab as First-Line Treatment in Extensive-Disease Small-Cell Lung Cancer: The GOIRC-AIFA FARM6PMFJM Trial. Pubmed ID 28135143. J Clin Oncol. 2017 Apr 20;35(12):1281-1287. doi: 10.1200/JCO.2016.69.4844.

Seckl MJ, Ottensmeier CH, Cullen M, Schmid P, Ngai Y, Muthukumar D, Thompson J, Harden S, Middleton G, Fife KM, Crosse B, Taylor P, Nash S, Hackshaw A. Multicenter, Phase III, Randomized, Double-Blind, Placebo-Controlled Trial of Pravastatin Added to First-Line Standard Chemotherapy in Small-Cell Lung Cancer (LUNGSTAR). Pubmed ID 28240967. J Clin Oncol. 2017 May 10;35(14):1506-1514. doi: 10.1200/JCO.2016.69.7391.

Mason MD, Clarke NW, James ND, Dearnaley DP, Spears MR, Ritchie AWS, Attard G, Cross W, Jones RJ, Parker CC, Russell JM, Thalmann GN, Schiavone F, Cassoly E, Matheson D, Millman R, Rentsch CA, Barber J, Gilson C, Ibrahim A, Logue J, Lydon A, et al. Adding Celecoxib With or Without Zoledronic Acid for Hormone-Naïve Prostate Cancer: Long-Term Survival Results From an Adaptive, Multiarm, Multistage, Platform, Randomized Controlled Trial. Pubmed ID 28300506. J Clin Oncol. 2017 May 10;35(14):1530-1541. doi: 10.1200/JCO.2016.69.0677.

Bradstock KF, Link E, Di Iulio J, Szer J, Marlton P, Wei AH, Enno A, Schwarer A, Lewis ID, D'Rozario J, Coyle L, Cull G, Campbell P, Leahy MF, Hahn U, Cannell P, Tiley C, Lowenthal RM, Moore J, Cartwright K, Cunningham I, Taper J, et al. Idarubicin Dose Escalation During Consolidation Therapy for Adult Acute Myeloid Leukemia. Pubmed ID 28368672. J Clin Oncol. 2017 May 20;35(15):1678-1685. doi: 10.1200/JCO.2016.70.6374.

Yao JC, Guthrie KA, Moran C, Strosberg JR, Kulke MH, Chan JA, LoConte N, McWilliams RR, Wolin EM, Mattar B, McDonough S, Chen H, Blanke CD, Hochster HS. Phase III Prospective Randomized Comparison Trial of Depot Octreotide Plus Interferon Alfa-2b Versus Depot Octreotide Plus Bevacizumab in Patients With Advanced Carcinoid

Tumors: SWOG S0518. Pubmed ID 28384065. J Clin Oncol. 2017 May 20;35(15):1695-1703. doi: 10.1200/JCO.2016.70.4072.

Jones RJ, Hussain SA, Protheroe AS, Birtle A, Chakraborti P, Huddart RA, Jagdev S, Bahl A, Stockdale A, Sundar S, Crabb SJ, Dixon-Hughes J, Alexander L, Morris A, Kelly C, Stobo J, Paul J, Powles T. Randomized Phase II Study Investigating Pazopanib Versus Weekly Paclitaxel in Relapsed or Progressive Urothelial Cancer. Pubmed ID 28402747. J Clin Oncol. 2017 Jun 1;35(16):1770-1777. doi: 10.1200/JCO.2016.70.7828.

Catton CN, Lukka H, Gu CS, Martin JM, Supiot S, Chung PWM, Bauman GS, Bahary JP, Ahmed S, Cheung P, Tai KH, Wu JS, Parliament MB, Tsakiridis T, Corbett TB, Tang C, Dayes IS, Warde P, Craig TK, Julian JA, Levine MN. Randomized Trial of a Hypofractionated Radiation Regimen for the Treatment of Localized Prostate Cancer. Pubmed ID 28296582. J Clin Oncol. 2017 Jun 10;35(17):1884-1890. doi: 10.1200/JCO.2016.71.7397.

Zucca E, Conconi A, Martinelli G, Bouabdallah R, Tucci A, Vitolo U, Martelli M, Pettengell R, Salles G, Sebban C, Guillermo AL, Pinotti G, Devizzi L, Morschhauser F, Tilly H, Torri V, Hohaus S, Ferreri AJM, Zachée P, Bosly A, Haioun C, Stelitano C, et al. Final Results of the IELSG-19 Randomized Trial of Mucosa-Associated Lymphoid Tissue Lymphoma: Improved Event-Free and Progression-Free Survival With Rituximab Plus Chlorambucil Versus Either Chlorambucil or Rituximab Monotherapy. Pubmed ID 28355112. J Clin Oncol. 2017 Jun 10;35(17):1905-1912.doi:10.1200/JCO.2016.70.6994.

Arcangeli G, Saracino B, Arcangeli S, Gomellini S, Petrongari MG, Sanguineti G, Strigari L. Moderate Hypofractionation in High-Risk, Organ-Confined Prostate Cancer: Final Results of a Phase III Randomized Trial. Pubmed ID 28355113. J Clin Oncol. 2017 Jun 10;35(17):1891-1897. doi: 10.1200/JCO.2016.70.4189.

The Lancet

Bruix J, Qin S, Merle P, Granito A, Huang YH, Bodoky G, Pracht M, Yokosuka O, Rosmorduc O, Breder V, Gerolami R, Masi G, Ross PJ, Song T, Bronowicki JP, Ollivier-Hourmand I, Kudo M, Cheng AL, Llovet JM, Finn RS, LeBerre MA, Baumhauer A, et al. Regorafenib for patients with hepatocellular carcinoma who progressed on sorafenib treatment (RESORCE): a randomised, double-blind, placebo-controlled, phase 3 trial. Pubmed ID 27932229. Lancet. 2017 Jan 7;389(10064):56-66. doi: 10.1016/S0140-6736(16)32453-9.

Rittmeyer A, Barlesi F, Waterkamp D, Park K, Ciardiello F, von Pawel J, Gadgeel SM, Hida T, Kowalski DM, Dols MC, Cortinovis DL, Leach J, Polikoff J, Barrios C, Kabbinavar F, Frontera OA, De Marinis F, Turna H, Lee JS, Ballinger M, Kowanetz M, He P, et al. Atezolizumab versus docetaxel in patients with previously treated non-small-cell lung cancer (OAK): a phase 3, open-label, multicentre randomised controlled trial. Pubmed ID 27979383. Lancet. 2017 Jan 21;389(10066):255-265. doi: 10.1016/S0140-6736(16)32517-X.

Durie BG, Hoering A, Abidi MH, Rajkumar SV, Epstein J, Kahanic SP, Thakuri M, Reu F, Reynolds CM, Sexton R, Orlowski RZ, Barlogie B, Dispenzieri A. Bortezomib with lenalidomide and dexamethasone versus lenalidomide and dexamethasone alone in patients with newly diagnosed myeloma without intent for immediate autologous stem-cell transplant (SWOG S0777): a randomised, open-label, phase 3 trial. Pubmed ID 28017406. Lancet. 2017 Feb 4;389(10068):519-527. doi: 10.1016/S0140-6736(16)31594-X.

Soria JC, Tan DSW, Chiari R, Wu YL, Paz-Ares L, Wolf J, Geater SL, Orlov S, Cortinovis D, Yu CJ, Hochmair M, Cortot AB, Tsai CM, Moro-Sibilot D, Campelo RG, McCulloch T, Sen P, Dugan M, Pantano S, Branle F, Massacesi C, de Castro G Jr. First-line ceritinib versus platinum-based chemotherapy in advanced ALK-rearranged non-small-cell lung cancer (ASCEND-4): a randomised, open-label, phase 3 study. Pubmed ID 28126333. Lancet. 2017 Mar 4;389(10072):917-929. doi: 10.1016/S0140-6736(17)30123-X.

Kepreotes E, Whitehead B, Attia J, Oldmeadow C, Collison A, Searles A, Goddard B, Hilton J, Lee M, Mattes J. High-flow warm humidified oxygen versus standard low-flow nasal cannula oxygen for moderate bronchiolitis (HFWHO RCT): an open, phase 4, randomised controlled trial. Pubmed ID 28161016. Lancet. 2017 Mar 4;389(10072):930-939. doi: 10.1016/S0140-6736(17)30061-2.

Neoptolemos JP, Palmer DH, Ghaneh P, Psarelli EE, Valle JW, Halloran CM, Faluyi O, O'Reilly DA, Cunningham D, Wadsley J, Darby S, Meyer T, Gillmore R, Anthoney A, Lind P, Glimelius B, Falk S, Izbicki JR, Middleton GW, Cummins S, Ross PJ, Wasan H, et al. Comparison of adjuvant gemcitabine and capecitabine with gemcitabine monotherapy in patients with resected pancreatic cancer (ESPAC-4): a multicentre, open-label, randomised, phase 3 trial. Pubmed ID 28129987. Lancet. 2017 Mar 11;389(10073):1011-1024. doi: 10.1016/S0140-6736(16)32409-6.

Cameron D, Piccart-Gebhart MJ, Gelber RD, Procter M, Goldhirsch A, de Azambuja E, Castro G Jr, Untch M, Smith I, Gianni L, Baselga J, Al-Sakaff N, Lauer S, McFadden E, Leyland-Jones B, Bell R, Dowsett M, Jackisch C; Herceptin Adjuvant (HERA) Trial Study Team. 11 years' follow-up of trastuzumab after adjuvant chemotherapy in HER2-positive early breast cancer: final analysis of the HERceptin Adjuvant (HERA) trial. Pubmed ID 28215665. Lancet. 2017 Mar 25;389 (10075):1195-1205. doi: 10.1016/S0140-6736(16)32616-2.

Atkin W, Wooldrage K, Parkin DM, Kralj-Hans I, MacRae E, Shah U, Duffy S, Cross AJ. Long term effects of once-only flexible sigmoidoscopy screening after 17 years of follow-up: the UK Flexible Sigmoidoscopy Screening randomised controlled trial. Pubmed ID 28236467. Lancet. 2017 Apr 1;389(10076):1299-1311. doi: 10.1016/S0140-6736(17)30396-3.

le Roux CW, Astrup A, Fujioka K, Greenway F, Lau DCW, Van Gaal L, Ortiz RV, Wilding JPH, Skjøth TV, Manning LS, Pi-Sunyer X; SCALE Obesity Prediabetes NN8022-1839 Study Group. 3 years of liraglutide versus placebo for type 2 diabetes risk reduction and weight management in individuals with prediabetes: a randomised, double-blind trial. Pubmed ID 28237263. Lancet. 2017 Apr 8;389(10077):1399-1409. doi: 10.1016/S0140-6736(17)30069-7.

Fixation using Alternative Implants for the Treatment of Hip fractures (FAITH) Investigators. Fracture fixation in the operative management of hip fractures (FAITH): an international, multicentre, randomised controlled trial. Pubmed ID 28262269. Lancet. 2017 Apr 15;389(10078):1519-1527. doi: 10.1016/S0140-6736(17)30066-1.

Ohman EM, Roe MT, Steg PG, James SK, Povsic TJ, White J, Rockhold F, Plotnikov A, Mundl H, Strony J, Sun X, Husted S, Tendera M, Montalescot G, Bahit MC, Ardissino D, Bueno H, Claeys MJ, Nicolau JC, Cornel JH, Goto S, Kiss RG, et al. Clinically significant bleeding with low-dose rivaroxaban versus aspirin, in addition to P2Y12 inhibition, in acute coronary syndromes (GEMINI-ACS-1): a double-blind, multicentre, randomised trial. Pubmed ID 28325638. Lancet. 2017 May 6;389(10081):1799-1808. doi: 10.1016/S0140-6736(17)30751-1.

Chan FKL, Ching JYL, Tse YK, Lam K, Wong GLH, Ng SC, Lee V, Au KWL, Cheong PK, Suen BY, Chan H, Kee KM, Lo A, Wong VWS, Wu JCY, Kyaw MH. Gastrointestinal safety of celecoxib versus naproxen in patients with cardiothrombotic diseases and arthritis after upper gastrointestinal bleeding (CONCERN): an industry-independent, double-blind, double-dummy, randomised trial. Pubmed ID 28410791. Lancet. 2017 Jun 17;389(10087):2375-2382. doi: 10.1016/S0140-6736(17)30981-9.

New England Journal of Medicine

Hiatt WR, Fowkes FG, Heizer G, Berger JS, Baumgartner I, Held P, Katona BG, Mahaffey KW, Norgren L, Jones WS, Blomster J, Millegård M, Reist C, Patel MR; EUCLID Trial Steering Committee and Investigators. Ticagrelor versus Clopidogrel in Symptomatic Peripheral Artery Disease. Pubmed ID 27959717. N Engl J Med. 2017 Jan 5;376(1):32-40. doi: 10.1056/NEJMoa1611688.

Strosberg J, El-Haddad G, Wolin E, Hendifar A, Yao J, Chasen B, Mittra E, Kunz PL, Kulke MH, Jacene H, Bushnell D, O'Dorisio TM, Baum RP, Kulkarni HR, Caplin M, Lebtahi R, Hobday T, Delpassand E, Van Cutsem E, Benson A, Srirajaskanthan R, Pavel M, et al. Phase 3 Trial of (177)Lu-Dotatate for Midgut Neuroendocrine Tumors. Pubmed ID 28076709. N Engl J Med. 2017 Jan 12;376(2):125-135. doi: 10.1056/NEJMoa1607427.

Montalban X, Hauser SL, Kappos L, Arnold DL, Bar-Or A, Comi G, de Seze J, Giovannoni G, Hartung HP, Hemmer B, Lublin F, Rammohan KW, Selmaj K, Traboulsee A, Sauter A, Masterman D, Fontoura P, Belachew S, Garren H, Mairon N, Chin P, Wolinsky JS, et al. Ocrelizumab versus Placebo in Primary Progressive Multiple Sclerosis. Pubmed ID 28002688. N Engl J Med. 2017 Jan 19;376(3):209-220. doi: 10.1056/NEJMoa1606468.

Mehra MR, Naka Y, Uriel N, Goldstein DJ, Cleveland JC Jr, Colombo PC, Walsh MN, Milano CA, Patel CB, Jorde UP, Pagani FD, Aaronson KD, Dean DA, McCants K, Itoh A, Ewald GA, Horstmanshof D, Long JW, Salerno C; MOMENTUM 3 Investigators. A Fully Magnetically Levitated Circulatory Pump for Advanced Heart Failure. Pubmed ID 27959709. N Engl J Med. 2017 Feb 2;376(5):440-450. doi: 10.1056/NEJMoa1610426.

Rogers JG, Pagani FD, Tatooles AJ, Bhat G, Slaughter MS, Birks EJ, Boyce SW, Najjar SS, Jeevanandam V, Anderson AS, Gregoric ID, Mallidi H, Leadley K, Aaronson KD, Frazier OH, Milano CA. Intrapericardial Left Ventricular Assist Device for Advanced Heart Failure. Pubmed ID 28146651. N Engl J Med. 2017 Feb 2;376(5):451-460. doi: 10.1056/NEJMoa1602954.

Shipley WU, Seiferheld W, Lukka HR, Major PP, Heney NM, Grignon DJ, Sartor O, Patel MP, Bahary JP, Zietman AL, Pisansky TM, Zeitzer KL, Lawton CA, Feng FY, Lovett RD, Balogh AG, Souhami L, Rosenthal SA, Kerlin KJ, Dignam JJ, Pugh SL, Sandler HM, et al. Radiation with or without Antiandrogen Therapy in Recurrent Prostate Cancer. Pubmed ID 28146658. N Engl J Med. 2017 Feb 2;376(5):417-428. doi: 10.1056/NEJMoa1607529.

Mok TS, Wu Y-L, Ahn M-J, Garassino MC, Kim HR, Ramalingam SS, Shepherd FA, He Y, Akamatsu H, Theelen WS, Lee CK, Sebastian M, Templeton A, Mann H, Marotti M, Ghiorghiu S, Papadimitrakopoulou VA; AURA3 Investigators. Osimertinib or Platinum-Pemetrexed in EGFR T790M-Positive Lung Cancer. Pubmed ID 27959700. N Engl J Med. 2017 Feb 16;376(7):629-640. doi: 10.1056/NEJMoa1612674.

Agus MS, Wypij D, Hirshberg EL, Srinivasan V, Faustino EV, Luckett PM, Alexander JL, Asaro LA, Curley MA, Steil GM, Nadkarni VM; HALF-PINT Study Investigators and the PALISI Network. Tight Glycemic Control in Critically Ill Children. Pubmed ID 28118549. N Engl J Med. 2017 Feb 23;376(8):729-741. doi: 10.1056/NEJMoa1612348.

Kantarjian H, Stein A, Gökbuget N, Fielding AK, Schuh AC, Ribera JM, Wei A, Dombret H, Foà R, Bassan R, Arslan Önder, Sanz MA, Bergeron J, Demirkan F, Lech-Maranda E, Rambaldi A, Thomas X, Horst HA, Brüggemann M, Klapper W, Wood BL, Fleishman A, et al. Blinatumomab versus Chemotherapy for Advanced Acute Lymphoblastic Leukemia. Pubmed ID 28249141. N Engl J Med. 2017 Mar 2;376(9):836-847. doi: 10.1056/NEJMoa1609783.

Bellmunt J, de Wit R, Vaughn DJ, Fradet Y, Lee JL, Fong L, Vogelzang NJ, Climent MA, Petrylak DP, Choueiri TK, Necchi A, Gerritsen W, Gurney H, Quinn DI, Culine S, Sternberg CN, Mai Y, Poehlein CH, Perini RF, Bajorin DF; KEYNOTE-045 Investigators. Pembrolizumab as Second-Line Therapy for Advanced Urothelial Carcinoma. Pubmed ID 28212060. N Engl J Med. 2017 Mar 16;376(11):1015-1026. doi: 10.1056/NEJMoa1613683.

Perry JR, Laperriere N, O'Callaghan CJ, Brandes AA, Menten J, Phillips C, Fay M, Nishikawa R, Cairncross JG, Roa W, Osoba D, Rossiter JP, Sahgal A, Hirte H, Laigle-Donadey F, Franceschi E, Chinot O, Golfinopoulos V, Fariselli L, Wick A, Feuvret L, Back M, et al. Short-Course Radiation plus Temozolomide in Elderly Patients with Glioblastoma. Pubmed ID 28296618. N Engl J Med. 2017 Mar 16;376(11):1027-1037. doi: 10.1056/NEJMoa1611977.

Weitz JI, Lensing AWA, Prins MH, Bauersachs R, Beyer-Westendorf J, Bounameaux H, Brighton TA, Cohen AT, Davidson BL, Decousus H, Freitas MCS, Holberg G, Kakkar AK, Haskell L, van Bellen B, Pap AF, Berkowitz SD, Verhamme P, Wells PS, Prandoni P; EINSTEIN CHOICE Investigators. Rivaroxaban or Aspirin for Extended Treatment of Venous Thromboembolism. Pubmed ID 28316279. N Engl J Med. 2017 Mar 30;376(13):1211-1222. doi: 10.1056/NEJMoa1700518.

Smits PC, Abdel-Wahab M, Neumann FJ, Boxma-de Klerk BM, Lunde K, Schotborgh CE, Piroth Z, Horak D, Wlodarczak A, Ong PJ, Hambrecht R, Angerås O, Richardt G, Omerovic E; Compare-Acute Investigators. Fractional Flow Reserve-Guided Multivessel Angioplasty in Myocardial Infarction. Pubmed ID 28317428. N Engl J Med. 2017 Mar 30;376(13):1234-1244. doi: 10.1056/NEJMoa1701067.

Attal M, Lauwers-Cances V, Hulin C, Leleu X, Caillot D, Escoffre M, Arnulf B, Macro M, Belhadj K, Garderet L, Roussel M, Payen C, Mathiot C, Fermand JP, Meuleman N, Rollet S, Maglio ME, Zeytoonjian AA, Weller EA, Munshi N, Anderson KC, Richardson PG, et al. Lenalidomide, Bortezomib, and Dexamethasone with Transplantation for Myeloma. Pubmed ID 28379796. N Engl J Med. 2017 Apr 6;376(14):1311-1320. doi: 10.1056/NEJMoa1611750.

Ridker PM, Revkin J, Amarenco P, Brunell R, Curto M, Civeira F, Flather M, Glynn RJ, Gregoire J, Jukema JW, Karpov Y, Kastelein JJP, Koenig W, Lorenzatti A, Manga P, Masiukiewicz U, Miller M, Mosterd A, Murin J, Nicolau JC, Nissen

S, Ponikowski P, et al. Cardiovascular Efficacy and Safety of Bococizumab in High-Risk Patients. Pubmed ID 28304242. N Engl J Med. 2017 Apr 20;376(16):1527-1539. doi: 10.1056/NEJMoa1701488.

Ramanan AV, Dick AD, Jones AP, McKay A, Williamson PR, Compeyrot-Lacassagne S, Hardwick B, Hickey H, Hughes D, Woo P, Benton D, Edelsten C, Beresford MW; SYCAMORE Study Group. Adalimumab plus Methotrexate for Uveitis in Juvenile Idiopathic Arthritis. Pubmed ID 28445659. N Engl J Med. 2017 Apr 27;376(17):1637-1646. doi: 10.1056/NEJMoa1614160.

Sabatine MS, Giugliano RP, Keech AC, Honarpour N, Wiviott SD, Murphy SA, Kuder JF, Wang H, Liu T, Wasserman SM, Sever PS, Pedersen TR; FOURIER Steering Committee and Investigators. Evolocumab and Clinical Outcomes in Patients with Cardiovascular Disease. Pubmed ID 28304224. N Engl J Med. 2017 May 4;376(18):1713-1722. doi: 10.1056/NEJMoa1615664.

Packer M, O'Connor C, McMurray JJV, Wittes J, Abraham WT, Anker SD, Dickstein K, Filippatos G, Holcomb R, Krum H, Maggioni AP, Mebazaa A, Peacock WF, Petrie MC, Ponikowski P, Ruschitzka F, van Veldhuisen DJ, Kowarski LS, Schactman M, Holzmeister J; TRUE-AHF Investigators. Effect of Ularitide on Cardiovascular Mortality in Acute Heart Failure. Pubmed ID 28402745. N Engl J Med. 2017 May 18;376(20):1956-1964. doi: 10.1056/NEJMoa1601895.

Lincoff AM, Nicholls SJ, Riesmeyer JS, Barter PJ, Brewer HB, Fox KAA, Gibson CM, Granger C, Menon V, Montalescot G, Rader D, Tall AR, McErlean E, Wolski K, Ruotolo G, Vangerow B, Weerakkody G, Goodman SG, Conde D, McGuire DK, Nicolau JC, Leiva-Pons JL, et al. Evacetrapib and Cardiovascular Outcomes in High-Risk Vascular Disease. Pubmed ID 28514624. N Engl J Med. 2017 May 18;376(20):1933-1942. doi: 10.1056/NEJMoa1609581.

Masuda N, Lee SJ, Ohtani S, Im YH, Lee ES, Yokota I, Kuroi K, Im SA, Park BW, Kim SB, Yanagita Y, Ohno S, Takao S, Aogi K, Iwata H, Jeong J, Kim A, Park KH, Sasano H, Ohashi Y, Toi M. Adjuvant Capecitabine for Breast Cancer after Preoperative Chemotherapy. Pubmed ID 28564564. N Engl J Med. 2017 Jun 1;376(22):2147-2159. doi: 10.1056/NEJMoa1612645.

Faries MB, Thompson JF, Cochran AJ, Andtbacka RH, Mozzillo N, Zager JS, Jahkola T, Bowles TL, Testori A, Beitsch PD, Hoekstra HJ, Moncrieff M, Ingvar C, Wouters MWJM, Sabel MS, Levine EA, Agnese D, Henderson M, Dummer R, Rossi CR, Neves RI, Trocha SD, et al. Completion Dissection or Observation for Sentinel-Node Metastasis in Melanoma. Pubmed ID 28591523. N Engl J Med. 2017 Jun 8;376(23):2211-2222. doi: 10.1056/NEJMoa1613210.

Wykrzykowska JJ, Kraak RP, Hofma SH, van der Schaaf RJ, Arkenbout EK, IJsselmuiden AJ, Elias J, van Dongen IM, Tijssen RYG, Koch KT, Baan J Jr, Vis MM, de Winter RJ, Piek JJ, Tijssen JGP, Henriques JPS; AIDA Investigators. Bioresorbable Scaffolds versus Metallic Stents in Routine PCI Pubmed ID 28402237. N Engl J Med. 2017 Jun 15;376(24):2319-2328. doi: 10.1056/ NEJMoa1614954.

Carbone DP, Reck M, Paz-Ares L, Creelan B, Horn L, Steins M, Felip E, van den Heuvel MM, Ciuleanu TE, Badin F, Ready N, Hiltermann TJN, Nair S, Juergens R, Peters S, Minenza E, Wrangle JM, Rodriguez-Abreu D, Borghaei H, Blumenschein GR Jr, Villaruz LC, Havel L, et al. First-Line Nivolumab in Stage IV or Recurrent Non-Small-Cell Lung Cancer. Pubmed ID 28636851. N Engl J Med. 2017 Jun 22;376(25):2415-2426. doi: 10.1056/NEJMoa1613493.

von Minckwitz G, Procter M, de Azambuja E, Zardavas D, Benyunes M, Viale G, Suter T, Arahmani A, Rouchet N, Clark E, Knott A, Lang I, Levy C, Yardley DA, Bines J, Gelber RD, Piccart M, Baselga J; APHINITY Steering Committee and Investigators. Adjuvant Pertuzumab and Trastuzumab in Early HER2-Positive Breast Cancer. Pubmed ID 28581356. N Engl J Med. 2017 Jul 13;377(2)122-131. doi: 10.1056/NEJMoa1703643.

Fizazi K, Tran N, Fein L, Matsubara N, Rodriguez-Antolin A, Alekseev BY, Özgûroğlu M, Ye D, Feyerabend S, Protheroe A, De Porre P, Kheoh T, Park YC, Todd MB, Chi KN; LATITUDE Investigators. Abiraterone plus Prednisone in Metastatic, Castration-Sensitive Prostate Cancer. Pubmed ID 28578607. N Engl J Med. 2017 Jul 27;377(4):352-360. doi: 10.1056/NEJMoa1704174.

James ND, de Bono JS, Spears MR, Clarke NW, Mason MD, Dearnaley DP, Ritchie AWS, Amos CL, Gilson C, Jones RJ, Matheson D, Millman R, Attard G, Chowdhury S, Cross WR, Gillessen S, Parker CC, Russell JM, Berthold DR, Brawley

C, Adab F, Aung S, et al. Abiraterone for Prostate Cancer Not Previously Treated with Hormone Therapy. Pubmed ID 28578639. N Engl J Med. 2017 Jul 27;377(4):338-351. doi: 10.1056/NEJMoa1702900.

Stone RM, Mandrekar SJ, Sanford BL, Laumann K, Geyer S, Bloomfield CD, Thiede C, Prior TW, Döhner K, Marcucci G, Lo-Coco F, Klisovic RB, Wei A, Sierra J, Sanz MA, Brandwein JM, de Witte T, Niederwieser D, Appelbaum FR, Medeiros BC, Tallman MS, Krauter J, et al. Midostaurin plus Chemotherapy for Acute Myeloid Leukemia with a FLT3 Mutation. Pubmed ID 28644114. N Engl J Med. 2017 Aug 3;377(5):454-464. doi: 10.1056/NEJMoa1614359.

Robson M, Im SA, Senkus E, Xu B, Domchek SM, Masuda N, Delaloge S, Li W, Tung N, Armstrong A, Wu W, Goessl C, Runswick S, Conte P. Olaparib for Metastatic Breast Cancer in Patients with a Germline BRCA Mutation. Pubmed ID 28578601. N Engl J Med. 2017 Aug 10;377(6):523-533. doi: 10.1056/NEJMoa1706450.

Neal B, Perkovic V, Mahaffey KW, de Zeeuw D, Fulcher G, Erondu N, Shaw W, Law G, Desai M, Matthews DR; CANVAS Program Collaborative Group. Canagliflozin and Cardiovascular and Renal Events in Type 2 Diabetes. Pubmed ID 28605608. N Engl J Med. 2017 Aug 17;377(7):644-657. doi: 10.1056/NEJMoa1611925.

Marso SP, McGuire DK, Zinman B, Poulter NR, Emerson SS, Pieber TR, Pratley RE, Haahr PM, Lange M, Brown-Frandsen K, Moses A, Skibsted S, Kvist K, Buse JB; DEVOTE Study Group.. Efficacy and Safety of Degludec versus Glargine in Type 2 Diabetes. Pubmed ID 28605603. N Engl J Med. 2017 Aug 24;377(8):723-732. doi: 10.1056/NEJMoa1615692.

Peters S, Camidge DR, Shaw AT, Gadgeel S, Ahn JS, Kim DW, Ou SI, Pérol M, Dziadziuszko R, Rosell R, Zeaiter A, Mitry E, Golding S, Balas B, Noe J, Morcos PN, Mok T; ALEX Trial Investigators. Alectinib versus Crizotinib in Untreated ALK-Positive Non-Small-Cell Lung Cancer. Pubmed ID 28586279. N Engl J Med. 2017 Aug 31;377(9):829-838. doi: 10.1056/NEJMoa1704795.

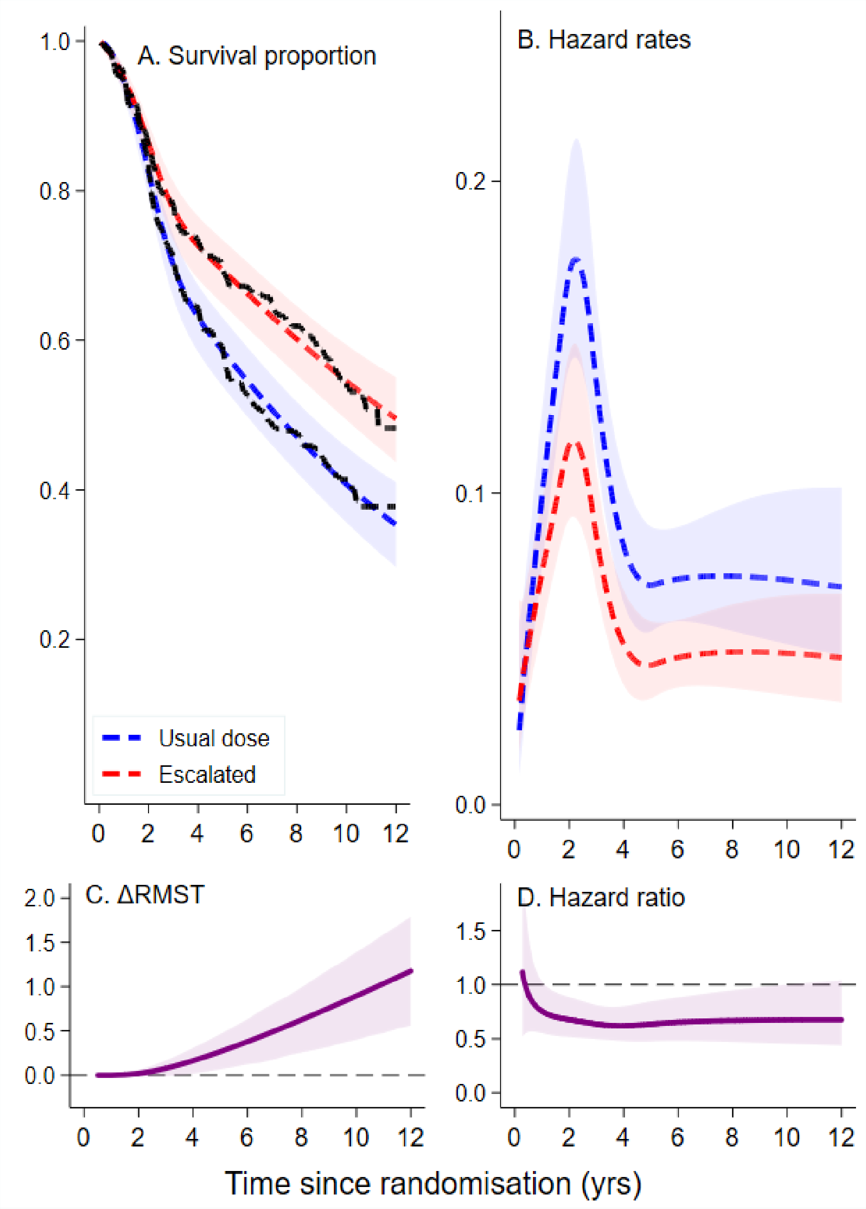
Review of Recommendations dataset

ID	Journal	Pubmed ID	General graphing			Survival curve specific recommendations				
			clear	xaxis	yaxis	step	risk	censor	event	uncert
3	NEJM	27959717	Yes	Yes	Yes	Yes	Yes	No	No	No
6	NEJM	28076709	Yes	Yes	Yes	Yes	Yes	Yes	No	No
8	NEJM	28002688	Yes	Yes	Yes	Yes	Yes	No	No	No
16	NEJM	28146658	Yes	Yes	Yes	Yes	Yes	No	No	No
18	NEJM	27959709	Yes	Yes	Yes	Yes	Yes	Yes	No	No
19	NEJM	28146651	Yes	Yes	Yes	Yes	Yes	No	No	No
24	NEJM	27959700	Yes	Yes	Yes	Yes	Yes	Yes	No	No
28	NEJM	28118549	Yes	Yes	Yes	Yes	No	No	No	No
33	NEJM	28249141	Yes	Yes	Yes	Yes	Yes	Yes	No	No
38	NEJM	28212060	Yes	Yes	Yes	Yes	Yes	Yes	No	No
39	NEJM	28296618	Yes	Yes	Yes	Yes	Yes	No	No	No
46	NEJM	28316279	Yes	Yes	Yes	Yes	Yes	No	No	No
48	NEJM	28317428	Yes	Yes	Yes	Yes	Yes	No	No	No
50	NEJM	28379796	Yes	Yes	Yes	Yes	Yes	Yes	No	No
59	NEJM	28304242	Yes	Yes	Yes	Yes	Yes	No	No	No
64	NEJM	28445659	Yes	Yes	Yes	Yes	Yes	No	No	No
66	NEJM	28304224	Yes	Yes	Yes	Yes	Yes	No	No	Yes
76	NEJM	28514624	Yes	Yes	Yes	Yes	Yes	No	No	No
78	NEJM	28402745	Yes	Yes	Yes	Yes	Yes	Yes	No	No
86	NEJM	28564564	Yes	Yes	Yes	Yes	Yes	Yes	No	No
87	NEJM	28591523	Yes	Yes	Yes	Yes	Yes	Yes	No	No
91	NEJM	28402237	Yes	Yes	Yes	Yes	Yes	No	No	No
95	NEJM	28636851	Yes	Yes	Yes	Yes	Yes	Yes	No	No
108	NEJM	28578639	Yes	Yes	Yes	Yes	Yes	No	No	No
109	NEJM	28578601	Yes	Yes	Yes	Yes	Yes	Yes	No	No
110	NEJM	28578607	Yes	Yes	Yes	Yes	Yes	No	No	No
111	NEJM	28581356	Yes	Yes	Yes	Yes	Yes	No	No	No
112	NEJM	28586279	Yes	Yes	Yes	Yes	Yes	Yes	No	No
113	NEJM	28605603	Yes	Yes	Yes	Yes	No	No	No	No

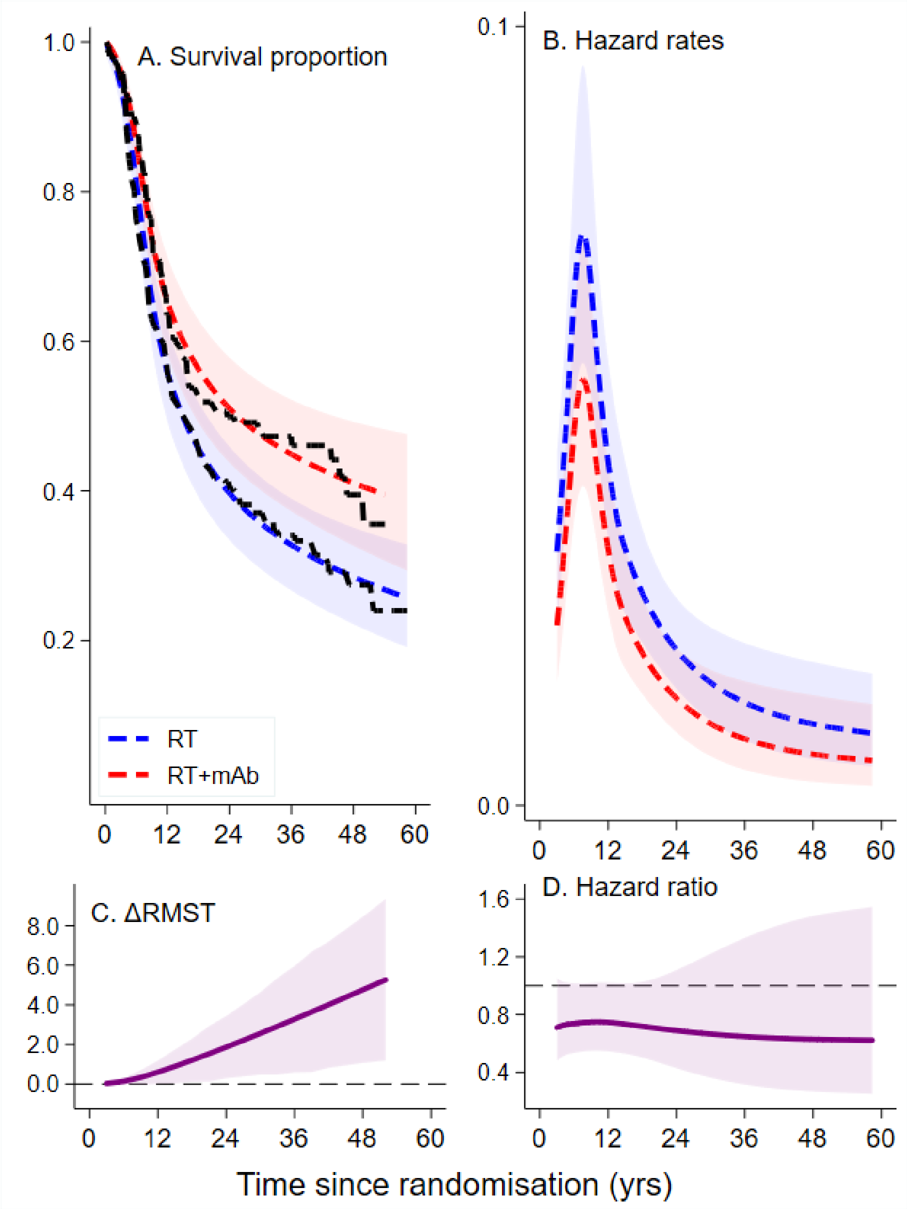
115	NEJM	28605608	Yes	Yes	Yes	Yes	Yes	No	No	No
117	NEJM	28644114	Yes	Yes	Yes	Yes	Yes	No	No	No
204	Lancet	27932229	Yes	Yes	Yes	Yes	Yes	Yes	No	No
210	Lancet	27979383	Yes	Yes	Yes	Yes	Yes	Yes	No	No
216	Lancet	28017406	Yes	Yes	Yes	Yes	Yes	Yes	No	No
233	Lancet	28161016	Yes	Yes	Yes	Yes	Yes	No	No	No
234	Lancet	28126333	Yes	Yes	Yes	Yes	Yes	Yes	No	No
238	Lancet	28129987	Yes	Yes	Yes	Yes	Yes	Yes	No	No
242	Lancet	28215665	Yes	Yes	Yes	Yes	Yes	No	No	No
246	Lancet	28236467	Yes	Yes	Yes	Yes	Yes	No	No	No
249	Lancet	28237263	Yes	Yes	Yes	Yes	Yes	Yes	No	No
252	Lancet	28262269	Yes	Yes	No	Yes	Yes	Yes	No	No
264	Lancet	28325638	Yes	Yes	Yes	Yes	Yes	No	No	No
282	Lancet	28410791	Yes	Yes	Yes	Yes	Yes	No	No	No
287	JCO	28034081	Yes	Yes	Yes	Yes	Yes	Yes	No	No
289	JCO	28034079	Yes	Yes	Yes	Yes	Yes	No	No	No
300	JCO	28056202	Yes	Yes	Yes	Yes	Yes	Yes	No	No
313	JCO	27918718	Yes	Yes	Yes	Yes	Yes	Yes	No	No
318	JCO	28129526	Yes	Yes	Yes	Yes	Yes	No	No	No
321	JCO	27937096	Yes	Yes	Yes	Yes	Yes	Yes	No	No
326	JCO	28029326	Yes	Yes	Yes	Yes	Yes	No	No	No
335	JCO	28199818	Yes	Yes	Yes	Yes	Yes	Yes	No	No
343	JCO	27400939	Yes	Yes	Yes	Yes	Yes	No	No	No
355	JCO	28135150	Yes	Yes	Yes	Yes	Yes	No	No	No
365	JCO	28113032	Yes	Yes	Yes	Yes	Yes	No	No	No
371	JCO	28380315	Yes	Yes	Yes	Yes	Yes	No	No	No
376	JCO	28221862	Yes	Yes	Yes	Yes	Yes	Yes	No	No
388	JCO	28135143	Yes	Yes	Yes	Yes	Yes	No	No	No
405	JCO	28300506	Yes	Yes	Yes	Yes	Yes	No	No	No
407	JCO	28240967	Yes	Yes	Yes	Yes	Yes	No	No	No
409	JCO	28384065	Yes	Yes	Yes	Yes	No	Yes	No	No
412	JCO	28368672	Yes	Yes	Yes	Yes	Yes	No	No	No
421	JCO	28402747	Yes	Yes	Yes	Yes	Yes	Yes	No	No
434	JCO	28355113	Yes	Yes	Yes	Yes	Yes	Yes	No	No
435	JCO	28355112	Yes	Yes	Yes	Yes	Yes	No	No	No
437	JCO	28296582	Yes	Yes	Yes	Yes	Yes	No	No	No

SUPPLEMENTARY MATERIAL

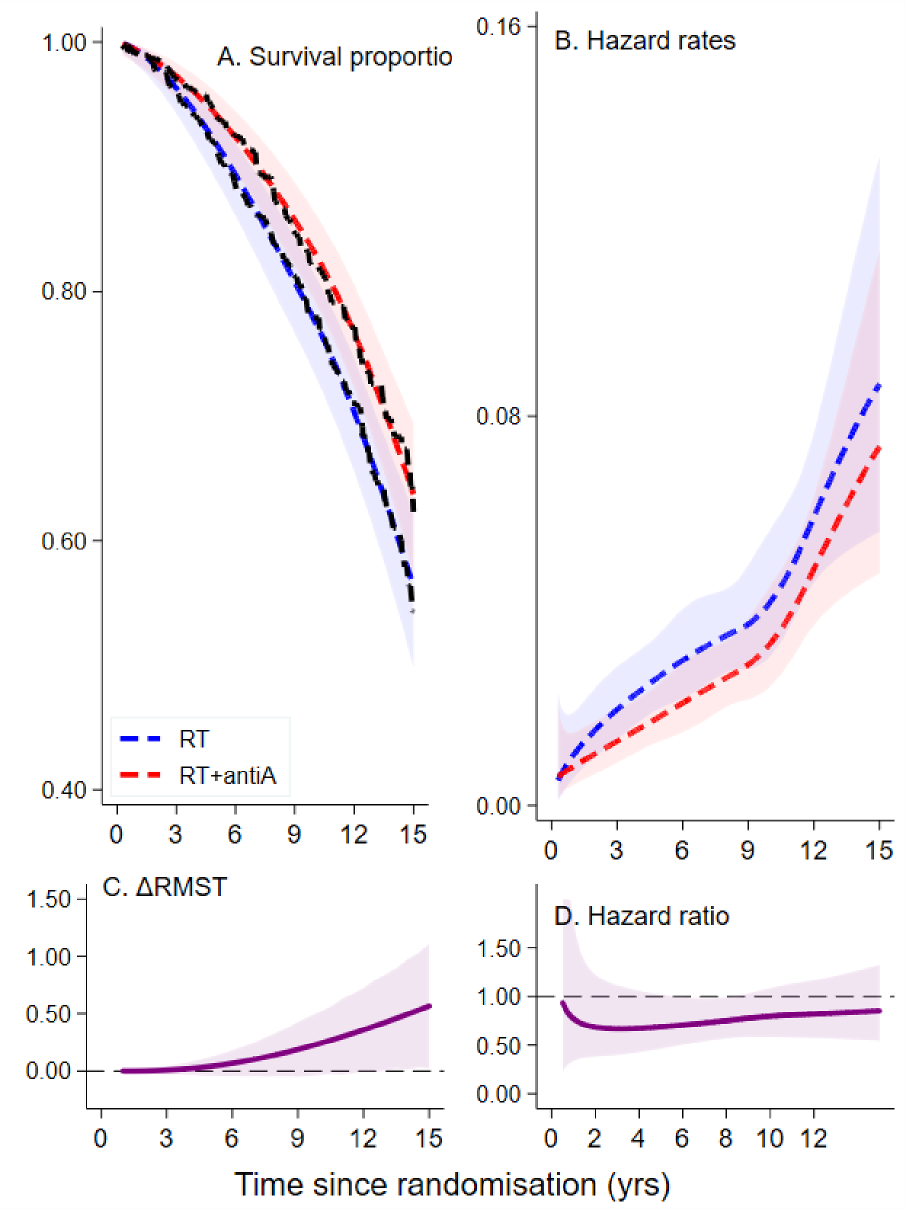
Complementing the Kaplan-Meier plot to enable assessment of treatment effects consistent with proportional hazards



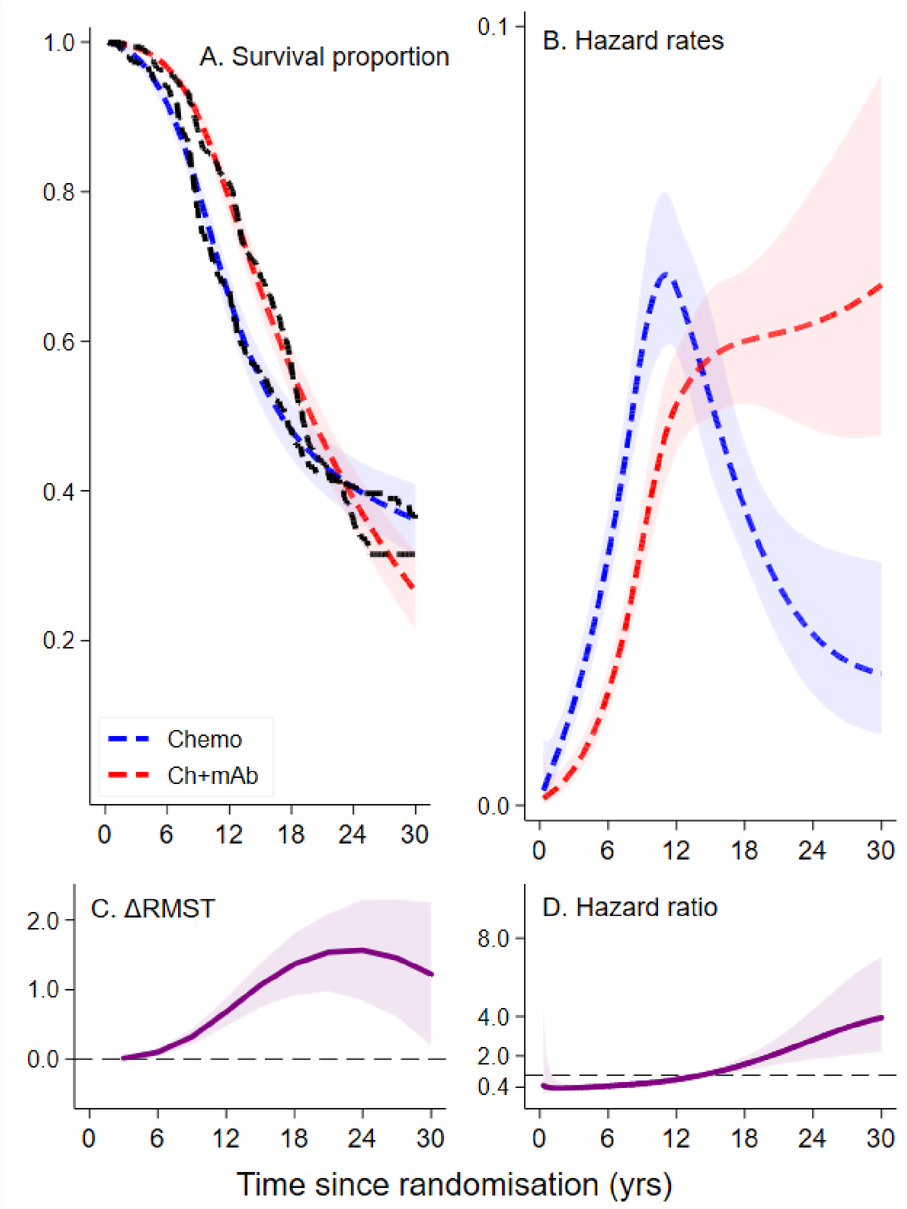
Supplementary Figure 1: RT01 Trial showing good proportionality, clear separation between arms of trial



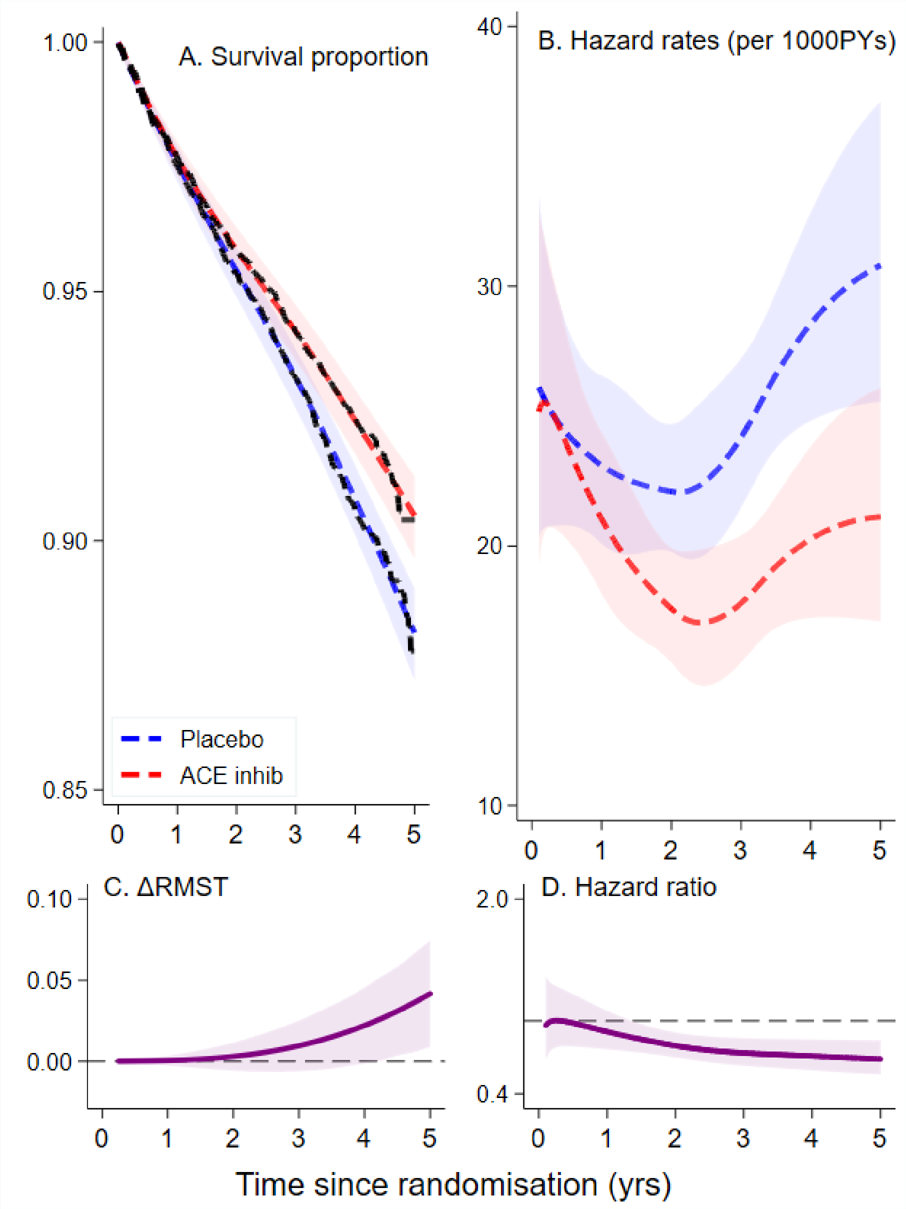
Supplementary Figure 2: Head and neck cancer trial showing good proportionality and treatment effect difference in the presence of some overlap of survival curve 95% CIs



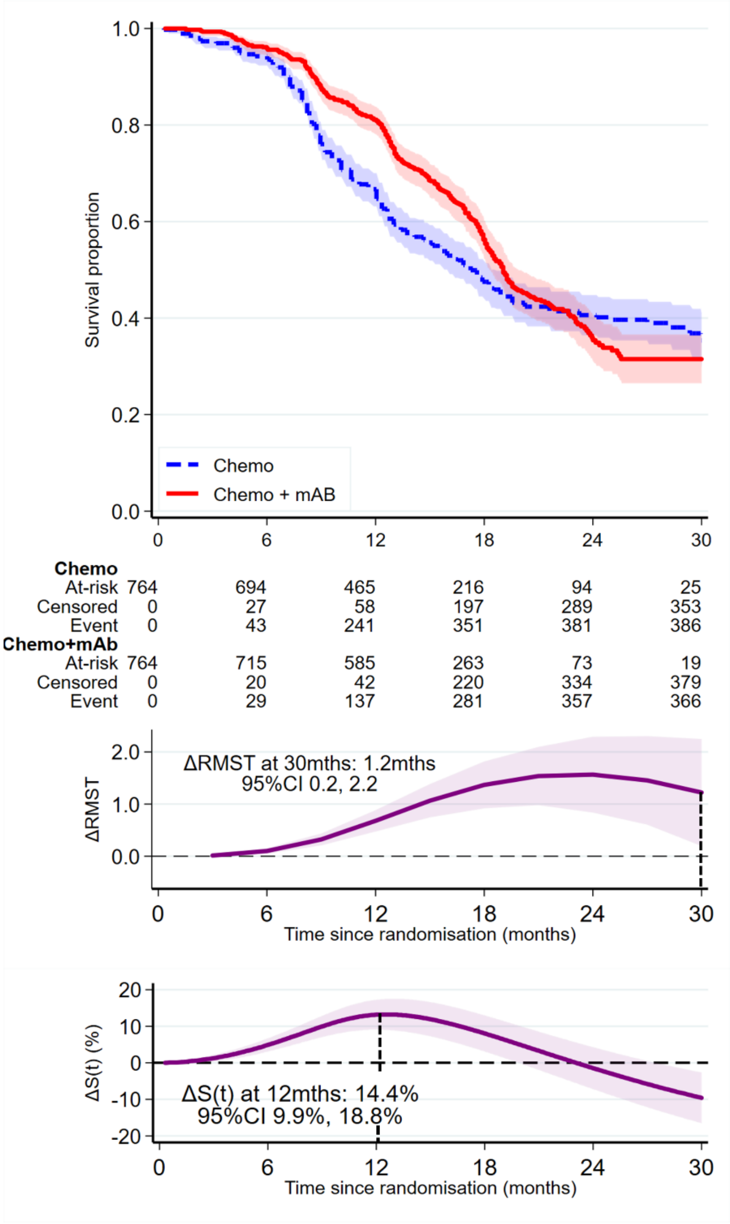
Supplementary Figure 3: Prostate cancer trial showing good proportionality and treatment effect difference in the presence of clear overlap of survival curve 95% CIs



Supplementary Figure 4: Ovarian cancer trial showing clear non-proportionality of treatment effect difference



Supplementary Figure 5: Cardiovascular events trial showing minor non-proportionality of treatment effect difference, increasing benefit



	0	6	12	18	24	30
Chemo						
At-risk	764	694	465	216	94	25
Censored	0	27	58	197	289	353
Event	0	43	241	351	381	386
Chemo+mAb						
At-risk	764	715	585	263	73	19
Censored	0	20	42	220	334	379
Event	0	29	137	281	357	366

Supplementary Figure 6: The ICON7 trial with two alternative treatment effect measures provided in the original report, the difference in restricted mean survival time (ΔRMST) and the difference in survival curves (ΔS(t)) at 12 months. The reported point estimate of treatment effect time is indicated by the black dashed line extending upwards from the x-axis.

SUPPLEMENTARY MATERIAL

Complementing the Kaplan-Meier plot to enable assessment of treatment effects consistent with proportional hazards

Stata code to create a Kaplan-Meier graph and complementary HR(t) plot for a trial with a treatment group and a comparison group.

Uses portions of the Stata code provided at https://github.com/tpmorris/kmunicate from the KMunicate paper.

Morris, T. P., Jarvis, C. I., Cragg, W., Phillips, P. P. J., Choodari-Oskooei, B., & Sydes, M. R. (2019). Proposals on Kaplan–Meier plots in medical research and a survey of stakeholder views: KMunicate. *BMJ Open, 9*(9), e030215. doi:10.1136/bmjopen-2019-030215

```
/*Graphical presentation paper

Figure 4 – (left) non proportional hazards ICON7

RECOMMENDED PRESENTATIONS

Dataset (trial_data.dta) contains trial data and estimated baseline and treatment group
survival curves with associated 95% CIs from
   •  Kaplan-Meier survival curves: skm0,skmlb0,skmub0,skm1,skmlb1,skmub1
   •  FPM TD model baseline hazard has 4df, time-dependenent trt effect with 2df
      (stpm2 trt, scale(hazard) df(4) tvc(trt) dftvc(2))
         o  hrrptd, hrrptd_lci, hrrptd_uci

See Tim Morris KMunicate paper for extended risk table code
*/
version 15.1
capture log close
clear

use trial_data.dta, clear

*get estimate of centile times
centile _t if _d==1, centile(2.5(2.5)10 20(10)100)

*recode to match "opt" groups as 1 and 2
recode trt (0=1) (1=2), gen(trt2)

* First create row labels for risk table (need to modify according to # groups -
clunky)
local times 0(6)30 // times at which you want to summarise
local groups 1 2 // labels for groups
forval j = `times' {
    foreach i of local groups {
        quietly count if trt2==`i' & _t >= `j'
            local risk_`i'_`j' = r(N)
         quietly count if trt2==`i' & _t < `j' & !_d
           local cens_`i'_`j' = r(N)
        quietly count if trt2==`i' & _t < `j' & _d
            local ev_`i'_`j' = r(N)
    }
      local opt `opt' `j' `"  " " "`risk_1_`j''" "`cens_1_`j''" "`ev_1_`j''" " "
"`risk_2_`j''" "`cens_2_`j''" "`ev_2_`j''" "'
}
di "`opt'"

quietly {
```

```
stcox trt
mat cox = r(table)
mat list cox

local c_hr  = round(cox[1,1],0.01)
local c_p   = round(cox[4,1],0.01)
local c_lb  = round(cox[5,1],0.01)
local c_ub  = round(cox[6,1],0.01)

noisily di "HR=" %03.2f `c_hr' ", 95%CI " %03.2f `c_lb' ","  %03.2f `c_ub' "
p="  %03.2f `c_p'
}
*HR=0.83, 95%CI 0.72,0.96 p=0.01

*control (blue) and treatment (red) group colors
local con_color_area "`"blue*1%20"'"
local trt_color_area "`"red*1%20"'"
local con_color_line "`"blue*1%100"'"
local trt_color_line "`"red*1%100"'"

local trteff_color_area "`"purple*1%10"'"
local trteff_color_line "`"purple*1%100"'"

*reference and model line types
local con_pattern "dash"
local trt_pattern "solid"

local alook_con "sort fc(`con_color_area') lc(white%10)"
local alook_trt "sort fc(`trt_color_area') lc(white%10)"
local alook_trteff "sort pstyle(ci) fc(`trteff_color_area') lc(white%10)"

local llook_con_gp "sort lc(`con_color_line') lp(`con_pattern') lw(thick) c(stepstair)"
local llook_trt_gp "sort lc(`trt_color_line') lp(`trt_pattern') lw(thick) c(stepstair)"
local llook_trteff "sort lc(`trteff_color_line') lp(`trteff_pattern') lw(thick)"


local ylabelopts "angle(horizontal) grid labsize(medium)"
local xlabelopts "labsize(medlarge)"
local xscale_opts "lwidth(medthick)"
local yscale_opts "lwidth(medthick)"

local ind_gr_size "ysize(4) xsize(4)"
local fy_val "30"
local comb_gr_size "ysize(8) xsize(6)"

*S(t)
*text sizes are too big for individual graphs but work when combined into a panel
tw rarea skmub0 skmlb0 _t,    `alook_con'             ///
|| rarea skmub1 skmlb1 _t,    `alook_trt'             ///
|| line skm0 _t,              `llook_con_gp'          ///
|| line skm1 _t,              `llook_trt_gp'          ///
xaxis(1 2 3)                                                      ///
ytitle("Survival proportion", size(medsmall) )                    ///
ylabel(0.0(0.2)1.0, format(%3.1f) `ylabelopts' )                  ///
yscale(range(0 1.0) `yscale_opts' )                               ///
xtitle("", size(medsmall) axis(1))                                ///
xtitle("", axis(2))                                               ///
xtitle("", axis(3))                                               ///
xscale(range(0 30) `xscale_opts' axis(1) )                        ///
xscale(range(0 30) lstyle(none) axis(2) )                         ///
xscale(range(0 30) lstyle(none) axis(3) )                         ///
xlabel(0(6)30, `xlabel_opts' axis(1))                             ///
xlabel(0(6)30, nolabels axis(3))                                  ///
xlabel(-2.2 `" "{bf:Chemo}" "At-risk" "Censored" "Event" "{bf:Chemo+mAb}" "At-risk"
"Censored" "Event" "' `opt', notick custom norescale labsize(medsmall) axis(2)
labjustification(right))                                    ///
legend(     label(3 "Chemo") label(4 "Chemo + mAB")         ///
            order(3 4)                                      ///
            position(7) ring(0) cols(1)                     ///
            region(lstyle(none) ) symxsize(*0.45)  )        ///
plotregion( color(white) fcolor(white) margin(small) )      ///
graphregion(color(white) fcolor(white) margin(l+10 b-8))    ///
ysize(4) xsize(4)                                           ///
name(km_plot, replace) draw

*text sizes are too big for individual graphs but work when combined into a panel
*hazard ratio over time
local st 0 /* left truncate at 5th centile if needed */
tw rarea hrrptd_uci hrrptd_lci _t if _t>`st', `alook_trteff'                 ///
|| line hrrptd _t if _t>`st', `llook_trteff'                                 ///
text(0.25 3 "Average HR=0.83, 95%CI 0.72,0.96" "p=0.01", size(medlarge) placement(se)
margin(zero)) ///
ytitle("Hazard Ratio", size(medsmall) )                                      ///
ylabel(0.2 0.5 1 2 8, format(%3.1f) `ylabelopts')                            ///
yscale(range(0.05 10) log  `yscale_opts' )                                   ///
yline(1, lpattern(dash) lwidth(medthick) lcolor(black))                ///
yline(`c_hr', lpattern(solid) lwidth(thick) lcolor(gs4%60))            ///
yline(`c_lb', lpattern(shortdash) lwidth(thick) lcolor(gs4%60))        ///
yline(`c_ub', lpattern(shortdash) lwidth(thick) lcolor(gs4%60))        ///
xtitle("Time since randomisation (months)")                            ///
xlabel(0(6)30, `xlabelopts')                                           ///
xscale(range(0 30)  `xscale_opts'  )                                   ///
fysize(`fy_val')                                                       ///
legend(off)                                                            ///
plotregion( color(white) fcolor(white) margin(small) )                 ///
graphregion(color(white) fcolor(white) margin(l+10 ))                  ///
`ind_gr_size'                                                          ///
name(hrt_plot, replace) draw

graph combine km_plot hrt_plot, cols(1) colfirst                       ///
graphregion(color(white) fcolor(white) margin(tiny))                   ///
`comb_gr_size’ nocopies iscale(*1)
exit
```