Project Proposal to Evaluate Methods to Accommodate Non-Proportional Hazards in the Analysis of Clinical Trials

Martin Hewing

19/04/2019

1 Introduction

A clinical trial can be defined as "any form of planned experiment that involves patients and is designed to elucidate the most appropriate treatment of future patients under a given medical condition". [1]

The purpose of the majority of such trials is to measure the time from when the patient is randomly introduced into the clinical study and given the treatment (experimental treatment versus standard treatment) to when some critical event occurs. Many trials have a time to event outcome. This event is often the survival time of the patient, for example the time to death for cancer patients, or time from initial diagnosis for HIV to progress to AIDS. Conversely, many have other types of outcome, for example measurements of some quantity such as blood pressure, glucose level or number of asthma attacks. However, it is important to note that "survival analysis" is the study of this process, the final outcome may not be actual patient survival time. [2]

The analysis and comparison of survival curves is often carried out using Cox's regression model, [3] The Cox model makes use of the proportional hazards assumption. When comparing a new treatment versus a standard treatment, there is the assumption that the ratio of the hazards for both groups is constant over time. However, a question of fundamental importance has been raised, in regard to the interpretability of the hazard ratio when non-proportional hazards are present. If the true hazard ratio is non-constant over time, it will lead to the supposition that the proportional hazards assumption has been violated. The ramifications of this violation are that the actual parameter estimation via the Cox regression model procedure might not give a measure of the differences between groups that is meaningful, in so much that it is not the true mean of the hazard ratio over time. [4]

In contemporary literature a number of different techniques within survival analysis have been put forward in an attempt to overcome issues in regard to non-proportional hazards between different treatment groups and its effect on interpretability of results in the Cox regression model. This project will analyse a compendium of differing methods via simulation studies with the ultimate aim of selecting the most efficient method(s) and critically evaluating the strengths and weaknesses of this/these technique(s).

1.1 The Cox Proportional Hazards model

This paper will now proceed to give a concise outline of censoring, the hazard function, the Cox model and some of the most important characteristics of this particular approach.

Censoring happens when some information about the survival times of individuals is available, but it is not known precisely. In general there are three separate answers to why censoring is likely to occur. Firstly, the individual does not get the event before the end of the study. Secondly, in the course of the study period, an individual is lost to follow-up, and thirdly, because an individual withdraws from the study, but for a reason other than due to getting the event of interest.

Additionally, there are three types of censoring that can happen. Firstly, data that is right-censored. In this case the true time survived is the same as or more than the recorded survival time. Secondly, data that is left-censored. In this case the true time survived is the same as or less than the recorded survival time. Thirdly, data that is interval-censored. In this case the true time survived is within a time interval which is known to the researcher. This project will focus on data that is right censored.

The hazard function(also known as the conditional failure rate) is given by:

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t} \tag{1}$$

An explanation of the various components of the hazard function are as follows: $h(t)$ is equal to the limit as Δt (gives instantaneous potential) gets closer to zero of the probability that an individual gets the event in the time interval $[t, t + \Delta t]$ given the individual has already survived up to time t divided by a small change in t, noting

that, $h(t)$ is not a probability as P has been divided by Δt.

Formally $h(t)$ can be defined as a function that "gives the instantaneous potential per unit time for the event to occur, given that the individual has survived up to time t". Alternatively, $h(t)$ can be thought of as like a speedometer. This device gives drivers information on their current velocity, for example a car travelling at 60 miles per hour, has the potential to cover 60 miles in the next hour. However, because the vehicle might accelerate or decelerate in that time period, at that moment when 60 MPH was observed, the speed reading does not show how many miles the car will actually cover, it just gives drivers velocity or instantaneous potential at that moment (this is assuming that it is given the driver has already driven some distance).

In a similar fashion the hazard function calculates the potential to get the event being studied at that instant, given survival up to that point in time. Furthermore, it is important to note that for specific t time values the hazard function has two properties: firstly, $h(t)$ is not limited to an upper bound value. Secondly, $h(t)$ is never negative, it is more than or equal to zero.

The Cox proportional hazards regression model formula is given by:

$$h(t, \mathbf{X}) = h_0(t) e^{\sum_{i=1}^{p} \beta_i X_i} \tag{2}$$

$$\mathbf{X} = (X_1, X_2, \ldots, X_p)$$

The left hand side of the above expression is the time t hazard that an individual faces given a collection of explanatory variables $\mathbf{X}$ whereas, the right hand side of (2) is the product of the baseline hazard function given by $h_0(t)$ and an exponential expression which is raised to the $\beta_i X_i$ summed over the p predictor X variables.

Additionally in regard to the proportional hazards assumption it is important to note that in the formula given in (2) the baseline hazard is a function of t but not the $X's$. Conversely, the exponential expression does not involve the $t's$, but does involve the $X's$. Therefore, the $X's$ are "time-independent" and can be defined as "any variable whose value for a given individual does not change over time". Moreover, two important properties of the Cox proportional hazards model regression model formula are that firstly when all the $X's$ are zero then the formula equals the baseline hazard function as an exponential raised to the zero equals one, and the exponential in the Cox formula ensures that the fitted hazard is not negative.

Secondly, the Cox model is semi-parametric because $h_0(t)$ the baseline hazard function is unspecified, whereas , parametric models such as the Weibull or Exponential have a fixed functional form, so the Cox model is appropriate in many data situations where functional form of the baseline hazard is unknown as it can often accurately approximate the correct parametric form.

The estimates of the β's (parameters) in the Cox model: $h(t, \mathbf{X}) = h_0(t) e^{\sum_{i=1}^{p} \beta_i X_i}$ are computed using a maximum likelihood approach to give the ML estimates $\hat{\beta}_i$.

Another important concept is the hazard ratio. This ratio allows comparison of two individuals that are differentiated by their predictor set ($X's$). Computing the hazard ratio as follows:

$$\widehat{HR} = \frac{\hat{h}\,(t, \mathbf{X}^*)}{\hat{h}(t, \mathbf{X})} \tag{3}$$

$$\mathbf{X} = (X_1, X_2, \ldots, X_p)$$

The hazard ratio is the division of $h\left(t, \mathbf{X}^*\right)$ by $h(t, \mathbf{X})$. Noting that the individual with the greater hazard is in the numerator (usually the placebo group) so that the hazard ratio is a value greater than one and $\hat{h}\left(t, \mathbf{X}^*\right) \geq \hat{h}(t, \mathbf{X})$, which aids in interpretability.

$$\widehat{HR} = \frac{\hat{h}_0(t) e^{\sum_{e=1}^{p} \hat{\beta}_i X_i^*}}{\hat{h}_0(t) e^{\sum_{e^i=1}^{p} \hat{\beta}_i X_i}} = e^{\sum_{i=1}^{p} \hat{\beta}_i (X_i^* - X_i)} \tag{4}$$

To arrive at the final formula in (4): Substitute the Cox model formula into the hazard ratio for the respective $\mathbf{X}^*$ and $\mathbf{X}$ values to express it in terms of the regression coefficients. Noting that the expression does not depend on t.

The proportional hazards assumption meaning can now be explained, using the hazard ratio.

$$\hat{h}\left(t, X^*\right) = \hat{\theta}\hat{h}(t, X) \tag{5}$$

By letting $\hat{\theta}$ equal to $\exp\left[\sum_{i=1}^{p} \hat{\beta}_i \left(X_i^* - X_i\right)\right]$ it follows that $\frac{\hat{h}(t, \mathbf{X}^*)}{\hat{h}(t, \mathbf{X})} = \hat{\theta}$ and so $\hat{h}\left(t, X^*\right) = \hat{\theta}\hat{h}(t, X)$ where the constant of proportionality $\hat{\theta}$ is not dependent on t. Therefore, one subjects hazard function is in proportion to another subjects hazard. [5]

2 Review of the Literature Concerning the Cox Proportional Hazards Regression Model under Conditions of Non-Proportional Hazards, Highlighting Proposed Techniques to Overcome the Violation of the Proportional Hazards Assumption

The major aim of this project is to assess the robustness the Cox model in regard to non-proportional hazards in an evolving clinical trials environment, and to analyse whether different techniques might be superior in a given situation in terms of efficiency, interpretability or any other factor that could potentially improve the ability to create cost effective life-saving medicines. Furthermore, in regard to objectives, this paper strives to provide robust inferences from analysing simulated data that is interpretable and will potentially advance subject knowledge in this area.

This proposal will now outline the processes undertaken to source the relevant literature in regard to Cox regression in the face of non-proportional hazards and methods to overcome these issues. The searches will attempt to locate papers that use methods that have been specifically developed and/or proposed for handling non-proportional hazards in the context of clinical trials. Additionally, examining other survival analysis models that do not have a proportional hazards assumption, and also publications featuring clinical trials that included non-proportional hazards. This information will be utilised in an attempt to achieve the aims and objectives of this paper.

The majority of searches carried out in the process of compiling this paper have been carried out via the "Search the literature" section of the Mathematical Sciences homepage on the University of Bath library portal. In the portal there are a number of relevant databases. However, it was decided to focus on those that specialise in mathematics and/or medicine.

The first platform searched was "PubMed" due to this database specialising in medicine. The criteria of the search was to find research papers with key words in the title or abstract, and this objective was carried out via an advanced search by title and abstract for a number of different in terms sorted by "Best Match" in the PubMed

portal starting with the simplest terms to more complex in an attempt to filter out results that are not relevant.

The first term "proportional hazards" returned 106158 results, "non-proportion hazards" returned 54126 results, whereas, "Cox non-proportional hazards" returned 68 results and from this the majority of the papers have been sourced after reading through the abstract, paying attention to citations and trying to find out which researchers might be the most prominent in the field, which techniques and evaluation seemed to make most logical sense in regard to furthering the aims of this project. Searches using the method of experimenting with different combinations of search terms returned. i Google Scholar using "Cox non-proportional hazards " returned 3,980 results, while "alternative Cox non-proportional hazards " returned 2,520, further attempting to refine the search by stipulating a ten year date range further reduced the results returned to 1,710 and from this visually trying to select relevant papers, which is highly problematic from an efficiency point of view. Searches of MathSciNet, Web of Science returned negligible results. The papers that were chosen, have been selected on merit due to the relative impact they have had on this subject area (via critical evaluation in other papers) and the recommendation of this project's supervisor.

First paper to be reviewed: Comparison of Treatment Effects Measured by the Hazard Ratio and by the Ratio of Restricted Mean Survival Times in Oncology Randomized Controlled Trials: Ludovic Trinquart, Justine Jacot, Sarah C. Conner, and Raphael Porch.

[6]

As previously stated a critically important concept within survival analysis is that of censoring. Due to the issues presented by censoring. The consequences are that survival time cannot be viewed as a regular continuous outcome, be summarised using summary statistics such as the mean or standard deviation, visualised in the usual manner using scatterplots etc. or analysed with modelling techniques such as linear regression. However, a major drawback of working in the probability domain is that it can be difficult to interpret and visualise probabilistic relationships such as the hazard ratio.

This gives motivation for the use of restricted mean survival time (RMST). This technique gives practitioners the ability to calculate the mean survival time up to a time point t^*. Formally, for the time-to-event random variable T, the equation for the true RMST is given by

$$\mu\left(t^*\right) = E\left[\min\left(T, t^*\right)\right] = \int_0^{t^*} S(t)\mathrm{d}t \tag{6}$$

Where the restricted mean at time t^* equals the expectation of a minimised function comprising of the random variable T and t^*, which is equal to the integral of the survival function at t between the limits of 0 and t^* integrated with respect to t. An intuitive explanation of this concept is that RMST is equal to the area beneath the survival curve up to a point t^* and can be thought of as the *'t* - year life expectancy'*, for example a recipient of a treatment could be informed that his/her life expectancy with disease A on treatment B over the next 12 months is 6 months or alternatively treatment C increases your life expectancy during the next 24 months by 8 months, when compared to treatment D. [7]

A paper examining the validity of using RMST versus HR is Trinquart *et al.* The authors conducted an empirical study looking at a sample of published phase III "parallel-group" randomised oncology trials from the last six months of 2014 from five separate journals.

The study attempted to ascertain whether the hazard ratio (HR) or the difference of restricted mean survival time (RMST) provided superior results in regard to overcoming some of the drawbacks (incorrect assessments of various treatment effects that are caused by interpreting HR as relative risks and that HR averages are dependent on the duration of follow-up) of using the HR to quantify survival benefit (longer life expectancy and/or improved

quality of life).

The method employed by the authors was to reconstruct "individual patient data" (IPD) for each trial of the time to event outcome by comparing HR to RMST. Firstly the researchers conducted a complex search of a number of journals including three general medical ("New England Journal of Medicine", "Lancet", "Journal of the American Medical Association") and two that specialise in oncology ("Journal of Clinical Oncology", "Lancet Oncology"), seeking those journals with reports on non-inferiority and superiority "parallel-group" trials of a randomised nature, whilst discarding all trial types apart from phase III. Secondly, the authors analysed the available trial data from the five journals to extract the non-adjusted HR and 95% confidence intervals, whilst noting if there was an assessment of any violation of the proportional hazards assumption. From the published Kaplan-Meier survival curves the IPD for each trial was then reconstructed. Furthermore, specialist software was used to compute the time and survival coordinates based on the aforementioned Kaplan-Meier curves. Moreover, the quantity of patients at risk and a total of the number of events was calculated. This information was used to solve inverted Kaplan-Meier equations by using an iterative numerical algorithm. Thirdly, the researchers having estimated the treatment effect in regard to the HR and RMST by the above "IPD" method, proceeded to implement a log-rank test to test for a null treatment effect, additionally using the "Grambsch-Therneau" method to test for the presence of hazards that are non-proportional. A Cox PH model was then used to estimate the HR and variance. If the HR was below a value of one, experimental intervention was favoured. An assessment of the RMST at the time horizon t^* (t^* : pre-specified as the "minimum of the largest observed event time of the two groups") was made in both the control and the experimental group. Furthermore, a determination of the RMST (both the difference and ratio) was calculated from the the survival function via the Kaplan-Meier estimate and the variances were computed using the delta method.

The experimental treatment was favoured if the difference in RMST was more than zero, or if a ratio of RMST was more than one. Moreover, a null-treatment effect was tested for using a comparison of the RMST difference (and ratio) with the standard error of the RMST which follows a standard normal distribution. Fourthly, comparing the HR and RMST in regard to the treatment effects, the authors note that the ratio of RMST and HR are two separate methods that quantify the difference in two survival curves, and, therefore, their interpretation is different. However, it is noted that both estimates (which can be thought of as the size of the respective treatment effects) are on an equal relative scale (so they can be compared).

The method used to assess the relative differences between HR and RMST was to compute which measure was further from its null systematically. Specifically, two techniques were made use of: i. plotting the HR's versus the ratios and differences of the RMST in the respective trials, furthermore, an inverse transformation of the HR enabled the researchers to show that if the RMST ratio and/or HR value greater than one would show that the experimental treatment was better. Moreover, using this information it was possible to ascertain the number of occasions that the HR gave a bigger treatment effect than the ratio of RMST in regard to the experimental treatment. ii. A different rule which used the ratio of the RMST to the HR to provide an estimate of differences in treatment effect in each trial: in this rule the HR should be twice as large or below half the RMST ratio. This method provides an estimate of the variance, which was computed via the bootstrap technique. The output (ratio between RMST and HR) can be quantified as a value more that one, shows that the HR approach is superior.

In conclusion, this paper found empirically that RMST conservatively measures the treatment effect versus a measurement via a HR approach. This was the same for overall survival and different time-to-event outcomes, regardless of whether non-proportionality of hazards was present, which relates to the analysis at the start of this section in regard to misinterpretation of HR's and the risk that it poses. Furthermore, difference in RMST gives clinicians the ability to quantify the absolute value of survival differences, which in-turn allows professionals to score the clinical benefit of this effect. Moreover, whether "minimally effective treatment" has been achieved can also be assessed by comparing the treatment under study to the direct clinical benefit threshold which has been predefined beforehand. Additionally, as the researchers found that treatment effects when quantified using HR's and not RMST, regardless of the proportional hazards assumption, appear to be of greater clinical benefit in a systematic way (when in-fact they are not), therefore, the authors recommend that RMST treatment effects measurements should be reported as standard in all trials, where the outcome is "time-to-event".

The researchers also point out that RMST has another advantage over HR's regardless of the proportional hazards assumption in regard to efficiency. When there is a small number of events it causes the HR confidence interval to become wide, whereas, RMST does not suffer from this issue, this would be important in non-inferiority trials.

The authors conclude by discussing the limitations of their study. Importantly, it is recognised that this paper was based on IPD's that were reconstructed, and therefore, some drop in the standard of the IPD's is expected when compared to the original trial IPD's. Furthermore, the integration technique used to estimate the RMST via the Kaplan-Meier curves, is prone to instability when there is a small number of patients at risk. However, this is most likely to happen in the tail of the distribution, and does not often fall in the time horizon that is being measured by RMST.

This proposal will refer to the work of Royston *et al.* [8] in an attempt to evaluate the research of Trinquart *et al.* [6] and how the results of the reviewed paper might impact this project. The main argument in favour of RMST is that this approach takes into account all the survival distribution up to the t^* value, rather than just a point in time. Furthermore, the RMST is straightforward to interpret and so could have an immediate impact at the clinical level in providing professionals with as much accurate and relevant information when compared to HR's. Moreover, the fact that there is no proportional hazards assumption is an advantage as there is less scope for imprecise results . However, some limitations of RMST have been raised that can also be applied to this research, for example if the t^* value has been incorrectly specified, then RMST can give results that are not appropriate, calling into question reproducibility by researchers without prior clinical background knowledge. As there is no mention of the risk of misspecification by the authors, it follows that the RMST approach is likely to be advantageous over using HR's in the presence of non-proportional hazards.

Second paper to be reviewed: Estimation of treatment effects in weighted log-rank tests Ray S. Lin, Larry F. León.

[9]

The analysis of the work of Lin *et al.* will begin with a concise explanation of the log rank test. This procedure is used to test the statistical equivalence of two survival curves for groups of two (or greater). This chi-square test (large sample) gives researchers the ability to get a sense of the overall differences between two survival curves so they can be compared. However, it does not provide evidence that survival curves of the true population are different. For two groups the log-rank test is as follows:

$$H_0 : \mathbf{S}_1(t) = \mathbf{S}_0(t) \tag{7}$$

$$H_1 : \mathbf{S}_1(t) \neq \mathbf{S}_0(t) \tag{8}$$

For all t, where $\mathbf{S}_1(t)$ *and* $\mathbf{S}_0(t)$ *are the two true survival functions.*

To conduct the test a comparison of the Log-rank statistic is made with the chi-square value. If the log-rank statistic is greater than the critical value we reject the null hypothesis and conclude that the survival curves are different. [10]

This proposal will now proceed to analyse the study conducted by Lin *et al.*. Firstly, the authors make note that the log-rank test is often used in the analysis of survival endpoint values. The authors further note that this

particular test gives the most statistical power in the presence of proportional hazards when compared to other types of tests. However, under non-proportional hazards the log-rank test suffers from a reduction in statistical power, additionally the regular Cox model gives estimates that are biased. Furthermore, research has found that a general form of the log-rank test known as the weighted log-rank test can lead to an increase in statistical power when non-proportional hazards are present. For example if a large proportion of subjects in a study were to discontinue treatments early, the estimate of the treatment effect is reduced and so there may be a fall in statistical power as those subjects might not continue get any benefit. However, if additional weight is given to earlier time periods in the study when subject discontinuation was less, then the benefit of treatment can be more accurately measured and vice versa. Moreover, the researchers note that it has been shown by Schoenfield [11] that if the weights are assigned in a proportional manner to the ratio size of the log hazard this leads to a test with the most possible power. Additionally, when the weights have been pre-selected (based on prior information such as treatment characteristics, study design type and the clinical situation) then type-one errors are preserved. Furthermore, the researchers state that although weighted log-rank tests have been used in past studies, there continues to be issues with how to select an appropriate weight function in regard to the relevant clinical setting and the subsequent interpretation of the results of the test in regard to treatment benefit.

Given the above analysis the authors propose a time varying based Cox model that will provide an estimate of the treatment effect, which will be complementary to the weighted log rank test. Therefore, the weight function assumptions are explicitly stated in regard to the relative sizes of the treatment benefits through time, and thus are examinable and can be verified in the context of the clinical setting. In addition the score test of the time varying based Cox model is equal to the weighted log-rank test, and so the model estimate gives a value of the treatment effect time profile. The researchers also note that as stated previously the model assumptions can be evaluated in regard to previous knowledge. Furthermore, an assessment of model fit can be made by analysing residual patterns. Moreover, the time profile (treatment effect) estimate can be used to make a judgement on the usefulness of the clinical benefit of the treatments.

Now follows a discussion of the technical details of methods undertaken by the authors. The researchers aim is to use their amended Cox model and weighted log-rank test in an attempt to provide evidence that this approach is potentially superior in the presence of non-proportional hazards compared to the standard log-rank test.

A group of n subjects are assigned randomly into the control arm or treatment arm of a clinical trial that has an endpoint that is a "time to event". Furthermore, if the i-th subject is randomly assigned to the control arm denoted by $X_i = 0$ and conversely if assigned to the treatment arm denoted by $X_1, X_2, \ldots X_n, X_i = 1$. Additionally, let the censoring times or event be denoted by $T_1, T_2, \ldots, T_n$ and the status be denoted by $\delta_1, \delta_2, \ldots, \delta_n$ (censoring is given by $\delta_i = 0$ and the event is given by $\delta_i = 1$ respectively). Moreover, let the J ordered event times be given by $T_{(1)} \leq T_{(2)} \leq \ldots \leq T_{(J)}$. The weighted log-rank statistic is given by:

$$Z = \frac{\sum_{j=1}^{J} w\left(T_{(j)}\right)\left(O_j - E_j\right)}{\sqrt{\sum_{j=1}^{J} w\left(T_{(j)}\right)^2 V_j}} \tag{9}$$

In the above formula the number of expected ((Noting that the expected are calculated assuming the null hypothesis (of no different in survival curves) is true.)) and observed time $T_{(j)}$ treatment arm events is denoted by E_j and O_j respectively. The E_j variance is given by V_j and the non-negative weight function of time is given by $w(.)$. The authors additionally note that there is no change if $w(.)$ is normalised or multiplied by k (a constant scalar).

Making use of the above information the authors then state the hazard function for a Cox model (proportional hazards) in regard to the i-th subject as:

$$\lambda\left(t; X_i\right) = \lambda_0(t) e^{w(t)\beta X_i} \tag{10}$$

The authors explain the treatment coefficient is given by β and the baseline hazard function is $\lambda_0(t)$. They then proceed to show that the weighted log-rank test is equal to the Cox model (10) score test. The approach is further developed as the authors proceed to incorporate the concept of adjustment factor effect into the analysis of the

above model, which is explained by the following formula:

$$A(t) = \frac{w(t)}{\max(w(t))} \tag{11}$$

The researchers then state that the weighted log-rank test, weight function at time t is $w(t)$ and $A(t)$ is the adjustment factor weight. As $w(.)$ will produce non-negative values, it follows that $A(t)$ will also be non-negative and will have a maximum value of one at certain point(s) in time. Given the preceding analysis the Cox model hazard function can now be stated as:

$$\lambda(t; X) = \lambda_0 e^{A(t)\beta X} \tag{12}$$

The authors note that there is no change in the Z (weighted log-rank statistic) value from scaling by $\max(w(t))$ and therefore, the weighted log-rank test that incorporates the weight function $w(t)$ is still equal to the score test of the model given in (9). Furthermore, the hazard function given in (12) can be expressed as a coefficient (constant) that has a covariate which varies with time and can be viewed as:

$$X^*(t) = A(t)X \tag{13}$$

The above formula is a representation of the treatment assignment that has an adjustment factor weighting. Furthermore, an updated Cox model that takes into account the preceding analysis is:

$$\lambda(t; X) = \lambda_0 e^{\beta X^*(t)} \tag{14}$$

From this an estimate of the β coefficient can be obtained. Additionally, $\hat{\beta}$ estimates from models that have covariates that vary with time have been found to be unbiased in this paper. Furthermore, as $A(t)$ is less than or equal to one, the coefficient β is a representation of the maximum effect in the time course. Moreover, the greatest weights (time t where $A(t) = 1$) are assigned to the points of time where the subjects experience the maximum effect (time t where $w(t) = \max(w(t))$) in the weighted log-rank test that corresponds to this analytical framework. In addition, conditional on using the correct model, the weighted log-rank test (or equally the model score test) gives the highest power and is optimal as per Schoenfield [11] which provides the theoretical underpinning for this result.
Proceeding from the above analysis the hazard ratio for the treatment and control arms for a function of time HR(t) can be derived as follows:

$$HR(t) = \frac{\lambda_0 e^{A(t)\beta \times 1}}{\lambda_0 e^{A(t)\beta \times 0}} = e^{\beta A(t)} = \left[HR^F\right]^{A(t)} \tag{15}$$

Where the full effect (maximum effect) is represented by $\text{HR}^F = e^\beta$ and in the model $e^{\beta A(t)}$ the treatment coefficient β is the time varying effect that has the effect treatment factor $A(t)$ as its weight.

In the literature a multitude of different approaches have been put forward in regard to choosing an appropriate weight function each with various strengths and weaknesses. For example, Harrington *et al.* [12] propose a family of weight function $G^{\rho,\gamma}$ that can give a representation of various different shapes of function based on an observation of survival as follows:

$$w(t) = S(t)^\rho (1 - S(t))^\gamma \tag{16}$$

Where $S(t)$ is the pooled population survival function; γ and ρ are the weight and shape parameters respectively.

When $\gamma = \rho = 0$, the weighted equals the standard log-rank test and when $\gamma = 0$ and $\rho = 1$, the weighted equals the Prentice-Wilcoxon test (a test where the time point t weight can be assigned based on time t survival). Furthermore, there is more allocation of weight to the middle points of time rather than the two end points when $\gamma = 1$ and $\rho = 1$, additionally when $\gamma = 1$ and $\rho = 0$ there is more weight allocation to time points that are later. When the $G^{1,1}$ weight function is used there is no prior effect from the treatment effect on the first event, and then

effect subsequently rises over the time period, reaching the maximum effect at approximately the median survival time, then the effect falls over time. The following function expresses the preceding analysis mathematically for the choice where rho and gamma are both 1:

$$\mathrm{HR}(t) = \left[\mathrm{HR}^F\right]^{S(t)(1-S(t))} \tag{17}$$

The researchers make use of simulation studies to test three models. i. Cox model (standard) and log-rank test. ii. A Cox model and weighted log-rank test. iii. A model of short and long term effects proposed by Yang *et al.* [13]. Using the three approaches/models proposed above, there is consideration of two further settings that considered different scenarios for how the treatment effect changes over time. Firstly; there is a delayed effect of the treatment and secondly; effect of long-term treatment falls when discontinuation of treatment is substantial. Furthermore, a 10,000 run simulation was conducted for various scenarios to calculate a number of different statistical measures such as an estimate of the hazard ratio, statistical power, the standard error and more. In the first setting there is the assumption that there is a minimal effect gained from the treatment at the beginning (HR = 0.9) and then the treatment will impose its maximum effect $\left(\mathrm{HR}^F = 0.68\right)$ after a delay τ_D of a specific period. In the second setting, the maximum effect of the treatment HR^F is imposed at the start and then falls slowly as more subjects discontinue the treatment. There is a loss of effect as all the subjects have finished their treatments. If $\gamma(t)$ is the proportion of subjects that are still receiving the treatment by time t, thus the proportion of subjects that have had the treatment discontinued is given by $1-\gamma(t)$. Furthermore, if subjects have a complete loss of the effect of the treatment, that takes place straightaway after stopping the treatment, it follows that the time t hazard ratio is given by $\mathrm{HR}(t) = \mathrm{HR}^F\gamma(t)+(1-\gamma(t))$. It should be noted that the treatment possibly could have an effect that continues beyond the end of the treatment. The authors found by letting the time period of the additional treatment effect be given by τ_P, it follows that $\mathrm{HR}(t) = \mathrm{HR}^F\,\gamma\left(t-\tau_P\right) + \left(1-\gamma\left(t-\tau_P\right)\right)$ for all $t \geq \tau_P$ and $\mathrm{HR}(t) = \mathrm{HR}^F$ for all time $t < \tau_P$.

The results of the simulations in the first setting when the treatment effect was delayed are as follows: In regard to statistical power, the power in the three approaches/models all fall as τ_D rises. However, it is noted that the weighted log-rank test has greater power than the other two tests (score of the regular Cox model and Yang tests) when there is correct specification of the weight function. In terms of estimates of standard error the proposed model provides unbiased estimates, whereas, both the regular model and the Yang tests give biased estimates of the beta due to underestimating the standard error and difficulty separating short/long term effects when τ_D is not equal to zero but close to that value respectively. Furthermore, in general under all three approaches/models there is a preservation of $type-I$ errors with the two Cox models (regular and proposed) performing better than the Yang tests. When a Grambsch–Therneau test (alpha = 5%) is used to test for non-PH it was found that the the regular Cox model violates the PH assumption more than the proposed model and this was exacerbated by longer delay. The explanation is that the the *A(t)* covariate which can vary with time gives the proposed Cox model the flexibility to mitigate this issue. Additionally, when a sensitivity analysis was performed, the weighted log-rank test exhibited a loss of power when there was a misspecification in the τ_D value. Although, the power of the weighted log-rank test was still greater than the regular log-rank test. Further analysis points towards robust and stable results even when there is misspecification of τ_D and $A(t)$ in the weighted log-rank test model.

In regard to setting two when there was a long-term reduction in the treatment effect. When τ_P is shorter, there is a fall in the statistical power of both the regular and weighted models, with the proposed model consistently outperforming the regular model. Furthermore, in the proposed model there was less violation of the assumption of proportional hazards and superior statistical performance in relation to the various metrics considered. Whereas, in the regular Cox model was shown to have a bias toward the null hypothesis. In terms of the Yang tests, the performance was below the weighted log-rank test model in regard to statistical power. Moreover, the Yang test had a standard error value that was higher than the other two approaches/models. In addition, in both settings (with and without misspecification), the results are on the whole stable and robust, even though there could be an underestimation of treatment effect.

The authors concede that there are a multitude of different methods available to provide treatment effect estimates in the presence of non-proportional hazards. It would wise to consider the possibility that the Yang test

could have chosen from various different techniques (such as adaptive lasso, cubic spines) with prior knowledge that it will not provide superior results to the weighted log-rank test and that other methods could be more suitable. The researchers do acknowledge that methods that flexible might be susceptible to over-fitting and thus perform poorly on unseen data. However, it is unclear why the Yang test was chosen rather than a competing approach, such as the Wilcoxon test.

A clear strength of the authors approach is that they have identified the issue of when the treatment effect is delayed, for example in oncology trials. The approach proposed by the authors in an attempt to overcome these time varying effects clearly has the potential to bring real benefits to patients. On the other hand a weakness of the study is that real data was not used and the researchers relied completely on simulation, and this could have created results that do not reflect real life outcomes in terms of unexpected data structure.

Third paper to be reviewed: The use of restricted mean survival time to estimate the treatment effect in randomized clinical trials when the proportional hazards assumption is in doubt Patrick Royston and Mahesh K. B. Parmar.

[8]

The authors begin by stating that the hazard ratio (invalid if the proportional hazards assumption is not met) is used in the majority of time-to-event right censored randomised control trials (RCT's), to evaluate whether a regular or prototype treatment is superior.

Many RCT's have a target HR and the authors state that the choice of HR level has been influenced by the seminal papers of Freedman [14] and Schoenfield [15]. In cancer RCT's for example the actual true HR is often around 0.75, between 80% to 90% power in a two sided test at a 5% significance level. A log-rank test is then used to assess the null hypothesis that the HR is equal to one (noting that a likelihood test is a viable alternative due to asymptotic equivalence under the proportional hazards assumption). Furthermore, the log-rank test can be considered somewhat robust to non-PH as it continues to have statistical power even under non-PH when evaluating the control versus the treatment. The log-rank test is used in the comparison of distribution functions, (no assumption of the shape of the distribution is made) while taking censoring into account. The authors further note that at the intersection of survival curves (Kaplan-Meier curves are often used to visualise the control and treatment arm survival curves) there is a loss of statistical power.

The researchers proceed to explain that in a time-to-event RCT where a measure of primary outcome is being evaluated, in general two metrics are usually presented; firstly a null hypothesis test and secondly a treatment effect summary. Furthermore, they state it would be prudent to include a test of non-proportional hazards, to backup this statement the authors point towards the work of Mok *et al.* [16] as their rationale. This study was criticised for reporting results as significant when non-proportional hazards were present, leading to questions of validity.

In an attempt to answer the question to what is an appropriate summary statistic (if one exists) in the presence of non-PH, the authors explore an approach that allows an interpretation of the individual HR as an average of the HR over the duration of follow-up time. They cite a study where a weighted Cox regression estimate was calculated by Schemper *et al.* [17] by evaluating a number of average HR's, coming to the conclusion that the average HR (AHR) proposed by Kalbfleisch *et al.* [18] was most appropriate.

$$AHR = \frac{\int [h_1(t)/h(t)]\, w(t) f(t) \mathrm{d}t}{\int [h_0(t)/h(t)]\, w(t) f(t) \mathrm{d}t} \tag{18}$$

The two group hazard functions are given by $h_0(t)$ and $h_1(t)$ respectively (Noting that $h(t) = h_0(t) + h_1(t)$). Additionally $w(t)$ is a user chosen weight function. Furthermore, $f(t)$ is a density function, although there is ambiguity to which distribution it belongs to.

Therefore an alternative approach to providing an appropriate summary statistic in the presence of non-PH could be preferred. The researchers continue to explain that they propose to empirically test an RMST approach used by Zucker [19] as the primary measure when non-PH are present, and the secondary measure when the assumption is satisfied.

The authors go on to state the methodology that they will follow in an attempt to show that the RMST approach is viable versus alternative competing techniques.

Firstly they give a technical explanation of the RMST (Identical to (6) in this proposal), Secondly, they discuss using "pseudo-observations" (the researchers use the term pseudovalues) put forward by Andersen *et al.* [20] in an attempt to examine how covariates affect RMST. The authors then continue to explain that pseudo-values are RMST jackknife estimates which are calculated from the sample survival curve Kaplan-Meier estimate (advantage of providing distribution free RMST estimates), and the entire sample RMST t^* value estimate is their mean value which is unbiased. Therefore, the authors were able to model the effects of covariates on RMST with the pseudovalues as the response value in a generalised linear model (noting that the standard errors must use the robust "sandwich" estimator). Thirdly, the researchers analyse the Cox model in regard to their RMST approach when the possibility of non-proportional hazards is likely and the regular Cox model would be biased. To overcome the issues with bias they recommend each treatment group (Kaplan-Meier estimate) should be integrated separately.

The survival estimates can then be thought of as coming from a stratified treatment group Cox model. However, the researchers point out that this approach has issues, for example the un-stability of $\widehat{s}_j(t)$.

Fourthly, the researchers discuss survival models that are parametric and flexible. The authors refer to Royston *et al.* [21] as an example of a study that made use of a flexible approach to accommodate various different baseline distributions. The log cumulative hazard function with the covariate vector $\mathbf{x}$ is given by:

$$\ln H(t;\mathbf{x}) = \ln H_0(t) + \mathbf{x}'\boldsymbol{\beta} = s(\ln t) + \mathbf{x}'\boldsymbol{\beta} \tag{19}$$

The log cumulative hazard function (baseline) $\ln H_0(t) = s(\ln t)$ is computed as a log time cubic spline (restricted), where the spline $s(\ln t)$ is a linear combination of regression parameters γ and basis functions given by $s(\ln t) = \gamma_0 + \gamma_1 \ln t + \gamma_2 v_1(\ln t) + \cdots + \gamma_{K+1} v_K(\ln t)$. The authors explain that all of the $K+1$ basis functions apart from the initial *ln t* are dependent on an "*interior knot*". They continue to state these are pairs of cubic polynomial segments (contiguous) that are joined in log-time and that the basis functions are created so that they are joined at the knots of the polynomial segments.

Some of the advantages of using splines are ease of use in the presence of non-proportional hazards by manipulating the spline function, and as already stated being flexible in regard to different baseline distribution functions.

The authors then outline how they estimate the RMST at time $\mathbf{x}$ by predicting the cumulative hazard function (log) from (19), then turning it into a survival function and finally integrating that survival function over $(0, t^*)$. Noting that the survival function is totally smooth and specified due to being fully parametric.

Now that the various techniques that the researchers plan to compare have been outlined (i. Pseudovalues, ii. Stratified treatment group Cox model, iii. Flexible parametric approach), the study proceeds to discuss three selected datasets that have varying degrees of non-proportional hazards. i. An advanced kidney cancer trial (RE01) where no evidence of non-PH has been identified (via a Grambsch–Therneau test at the 5% significance level). ii. An advanced ovarian cancer trial (GOG111) with evidence of non-PH via a Grambsch–Therneau test with $P = 0.006$. iii. A Lung cancer trial (IPASS) with the presence of non-PH (via significant a Grambsch–Therneau test) and crossing survival curves.

In trial i, (An advanced kidney cancer trial (RE01)) as there is no evidence of non-proportional hazards the treatment effect is legitimately estimated by the HR. However, the authors highlight the fact that the RMST estimates (flexible parametric and pseudo-values) at $t^* = 4$y via the flexible parametric method provide similar results to

the Stratified treatment group Cox model and are significant at the 1% significance level, therefore all the methods give unbiased results as expected.

Continuing to trial ii, (advanced ovarian cancer trial (GOG111)) it can be seen that the arm of the experimental treatment shows a marked improvement in survival. However, because the Grambsch–Therneau test of $P = 0.006$ identified the presence of non-PH (if the data censored at year five the test becomes insignificant) it calls into question whether the HR is a legitimate measure of survival when estimated via the Stratified treatment group Cox model. The researchers state that the RMST estimates at $t^* = 7$y could give reliable estimates even if non-PH are a concern.

In trial iii, (A Lung cancer trial (IPASS)) the authors state that crossing survival curves (progression free survival) point towards the presence of highly significant levels of non-PH, and therefore cast doubt on the validity of using the Stratified treatment group Cox model in this situation. However, both the RMST approaches give significant results and would be preferred over methods that are susceptible to non-PH. Noting that the authors do discuss the possibility over both overfitting and under-fitting the data in regard to RMST in the respective trials, surmising that overfitting is usually not as much as a problem as under-fitting (it may bias the restricted mean), and that adjusting the degrees of freedom in the flexible model is the best solution to this issue.

The above analysis is then used to come to a reasoned conclusion about how to design future clinical trials so that they can potentially be robust to the influence of non-PH. The researchers, outline a four stage plan that could be followed. Firstly, use a log-rank test (somewhat robust to non-PH) to carry out an appropriate hypothesis test on treatment effect and come to reasoned conclusions. Secondly, irrespective of the significance of the log-rank test, a Grambsch–Therneau test (Scaled Schoenfeld residual approach) should be carried out to ascertain if non-PH are present (visualisations can help). Thirdly, when non-PH are not found, the HR is the most relevant treatment effect primary summary (including confidence interval). Fourthly, if non-PH have been detected then the most appropriate estimate of the treatment effect primary measure is the t^* difference in RMST (and CI) via either the pseudovalue or flexible parametric technique. Noting that in most cases the most sensible t* value is motivated by clinical requirements.

In terms of strengths and weaknesses of the preceding recommendations in regard to using RMST, the authors do comment that some of the major advantages of using this approach are the interpretability, lack of PH assumption (i.e gives robust accurate results in various scenario's). However, limitations are that the t^* value must be carefully chosen so not to lead to results that are misleading. Therefore, clinical expertise must be applied to the choice of t^*.

In addition to the authors evaluation, the opinion of this proposal in regard to strengths of the above paper are that they used two different RMST approaches which allowed the reader to compare the results of both Pseudovalues and Flexible parametric to the standard Cox model approach, and this gave more confidence in their findings as both performed well with non-PH. Furthermore, they used real data in their analysis which could help highlight the need for RMST to be used more often (simulation studies can feel somewhat abstract, but increases in survival from real world data are potentially more believable somehow). In terms of limitations, it might be difficult to tune the models that the authors have proposed without specialist knowledge of the clinical treatment area, leading to issues of reproducing their findings with another dataset.

3 Work plan

In this section a concise work plan will be put forward. The aim of this study to empirically test the models that have been proposed in the current literature to try and break and/or further confirm their findings with different data and/or changes in assumptions. It is hoped that this will allow for the robustness, and strengths/weaknesses to be evaluated, with the ultimate aim of forwarding understanding in this area of research so that professions have accurate information that can be used to increase patient benefit. This will be carried out using simulated data.

i. The three papers reviewed in this proposal were selected because they seem to present robust solutions to

the issue of non-PH. However, the literature has many other studies that could be of potential interest to provide different research approaches that may give unbiased, accurate results when non-PH are present, therefore, this study will continue to examine and review appropriate literature, in an attempt to forward the study objectives.

ii. In regard to simulated data, this study will attempt to emulate the process(es) used to simulate the data in the various models that are being evaluated and in an attempt to try and create simulated datasets that are representative of the real world this project will look at a multitude of different survival curves (e.g crossing and those that are initially together and then separate).

The simulation technique that will be primarily used will be Monte-Carlo, furthermore the data will be fitted using a variety of methods including Weighted log-rank and RMST and assessed via examination of their relative bias, summary statistics, coverage and 95% confidence intervals and goodness of fit metrics such as mean square error.

iii. Evaluation of the findings of the above simulations and the output in regard to the current literature as stated above.

Timeline: In weeks one and two this project will continue to examine current literature in an attempt to locate additional approaches that could be evaluated in the presence of non-PH to complement those already found. In weeks three to seven, commence simulating the respective datasets with a focus on creating data that reflect different survival curve trajectories (e.g crossing and those that are initially together and then separate). In weeks eight to nine the data will be fitted and assessed. In weeks ten to twelve the results will be evaluated in regard to the current literature.

References

[1] Pocock S J. *Clinical Trials, A practical Approach*. Wiley, 1983.

[2] Machin D et al. *Medical Statistics*. Wiley, 2007.

[3] Cox D.R. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34 No. 2:187–220, 1972.

[4] Uno H et al. Moving beyond the hazard ratio in quantifying the between-group difference in survival analysis. *Journal of Clinical Oncology*, 2380-5:32, 2014.

[5] Klein D et al. *Survival Analysis A Self-Learning Text*. Springer pp 37-47, 2005.

[6] 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*, 1813-1819:34, 2016.

[7] Royston P. Restricted mean survival time: calculation and some applications in trials and prognostic studies, url = https://www2.le.ac.uk/members/pl4/workshop2011-1/royston-stockholm-10nov2011b.pdf, 2011.

[8] Royston P. The use of restricted mean survival time to estimate the treatment effect in randomized clinical trials when the proportional hazards assumptionis in doubt. *Statistics in Medicine*, Volume 30, Issue 19, 2011.

[9] Lin et al. Estimation of treatment effects in weighted log-rank tests. *Contemporary Clinical Trials Communications*, 8:147–155, 2017.

[10] Klein D et al. *Survival Analysis A Self-Learning Text*. Springer pp 70-72, 2005.

[11] D. Schoenfeld. The asymptotic properties of nonparametric tests for comparing survival distributions. *Biometrika*, 68:219–316, 1981.

[12] Harrington et al. A class of rank test procedures for censored survival data. *Biometrika*, 68:553–566, 1982.

[13] S. Yang et al. Improved logrank-type tests for survival data using adaptive weights. *Biometrics*, 66:30–38, 2010.

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

[15] Schoenfeld DA. Sample-size formula for the proportional-hazards regression model. *Biometrics*, 39:499–503, 1983.

[16] Mok et al. Gefitinibor carboplatin–paclitaxel in pulmonary adenocarcinoma. *New England Journal of Medicine*, 361:947–957, 2009.

[17] Schemper et al. The estimation of average hazard ratios by weighted cox regression. *Statistics in Medicine*, 28:2473–2489, 2009.

[18] Kalbfleisch et al. Estimation of the average hazard ratio. *Biometrika*, 68:105–112, 1981.

[19] Zucker DM. Restricted mean life with covariates: modification and extension of a useful survival analysis method. *Journal of the American Statistical Association*, 93:702–709, 1998.

[20] Andersen et al. Regression analysis of restricted mean survival time based on pseudo-observations. *Lifetime Data Analysis*, 10:335–350, 2004.

[21] Royston et al. Flexible proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. *Statistics in Medicine*, 21:335–350, 2002.