Clinical Trials - Rough

RAUL, ABHISHEAK UNNITHAN

March 2025

0.1 Survival Data Start-Up

The type of data we are dealing with in this clinical trial is survival data. This type of data captures the amount of time that elapses before a particular event happens.

Survival data is trickier to handle than other data types for 2 main reasons. Firstly, survival data is very often skewed, so even though it is usually continuous, we can't just treat it as normally distributed. Secondly, with time-to-event data, we don't usually observe the full dataset.

0.2 Aside: Sample size calculations for time-to-event data

There are implications here for sample size calculations, which must take into account the duration of a trial; trials must monitor patients until a sufficient proportion has experienced the event (whatever it is). Sample size calculations for time-to-event data, therefore, have two components:

1. The power of the trial can first be expressed in terms of m, the number of complete observations.
2. A separate calculation is needed to estimate the number of participants needing to be recruited and the length of the trial to be sufficiently likely to achieve that value of m.

Both of these calculations rely on a number of modelling assumptions and on previous scientific/clinical data (if available).

We will think more about how this can be used in the next section **(Section 8?)** when we come to compare treatment effects.

0.3 Censored Times

If a trial monitors a sample of participants for some length of time, many will experience the event before the trial is finished. However, for some of the samples, we likely won't observe the event. This could be because it doesn't happen within the lifetime of the trial, or it could be because the participant exits the trial prematurely for some reason (eg. withdrawal or they stop attending follow-up appointments after a certain time). For these participants, we do know that they had not experienced the event up to some time t, but we don't know what happened next. All we know is that their time-to-event or survival time is greater than that time t. These partial observations are known as censored times, and in particular as right-censored times, because the event happens after the censored time.

0.4 Censored Times - My Writing

Before starting the trial analysis, we need to deal with the (right) censored times.

If we were to treat censored times as observations, i.e. as though the event had happened at time t, we would bias the results of the trial very seriously. The survival times reported would be systematically shorter than the true ones because some of the participants whose observations were censored may not experience the event by the end of the trial.
If we were to remove the censored times and only analyse the data in which the event was observed during the lifespan of the trial, we would be losing data and, therefore, valuable information. This approach may well also lead to bias, for example, if some subset of patients experienced the event quite soon into the trial, but the remainder had not (past the end of the trial). If our analysis ignores those who did not experience the event, we are likely to underestimate the general survival time.

Therefore, we need to include these censored times in our analysis. We can do this using Survival Curves or a Hazard Function.

The survival time (or time-to-event) t for a particular individual can be thought of as the value of a random variable T, which can take any non-negative value. We can think in terms of a probability

distribution over the range of T. If T has a probability distribution with an underlying probability density function $f(t)$, then the cumulative distribution function is given by

$$F(t) = P(T < t) = \int_0^t f(u)\, du,$$

and this gives us the probability that the survival time is less than t.

The survival function, $\mathrm{S}(t)$, is the probability that some individual, here a participant, survives longer than time t. Therefore, $\mathrm{S}(t) = 1 - F(t)$. The survival curve is given by plotting $S(t)$ against t. One summary that is often used is the median survival time, the time at which half of the participants have already experienced the event, and half haven't.

We can also use the survival function to derive the Hazard Function, $h(t)$. The Hazard function is the probability that an individual who has survived up to time t experiences the event just after time t. It can be written simply as:

$$h(t) = \frac{f(t)}{s(t)},$$

where $\mathrm{S}(t) = \Pr(T > t)$ and $f(\cdot)$ is the probability density of T.
The hazard function can take **any positive value** (unlike the survival function), and for this reason, $\log(h(t))$ is often used to transform it to the real line.

There are fundamentally 2 ways to deal with survival data: parametrically or non-parametrically. The non-parametric paradigm is prevalent in survival analysis. **We will consider some methods from both paradigms**.

0.5 Kaplan-Meier Estimator

The Kaplan-Meier estimator is a **non-parametric estimate** of $\mathrm{S}(t)$. The idea behind it is to divide the interval $[0, t]$ into many short consecutive intervals:

$$[0, s_1), [s_1, s_2), [s_2, s_3), \ldots, [s_{k-1}, s_k), [s_k, t],$$

where $s_k < s_{k+1}, s_0 = 0$ and $s_{k+1} = t$. We then estimate the probability of surviving past some time t by multiplying together the probabilities of surviving the successive intervals up to t. The probability of surviving a particular interval $[s_i, s_{i+1})$ is estimated by $1 - Q$, where:

$$Q = \frac{\text{Number who die in that interval}}{\text{Number at risk of death in that interval}}.$$

More precisely, if there are observed event times $t_1 < t_2 < \cdots < t_n$, and the number of deaths at time t_i is d_i out of a possible n_i. Then the Kaplan-Meier estimate of $S(t)$ is

$$\hat{S}(t) = \prod_{j:t_j \le t} \left(1 - \frac{d_j}{n_j}\right).$$

Notice that the number of people at risk at time t_{j+1}, denoted n_{j+1}, will be the number of people at risk at time t_j, which was n_j, minus any who died at t_j, which we write as d_j, and any who were censored after t_j in the interval $[t_j, t_{j+1})$. In this way, the Kaplan-Meier estimator incorporates information from individuals with censored survival times up to the point they were censored.

For a clinical trial, we want to **plot the survival curves separately for the different treatment groups**. This will give a first, visual idea of whether there might be a difference and also of the suitability of certain models (we'll talk about this later).

0.6 Parametric Approach

In a parametric approach, we assume that the survival time T follows some probability distribution, up to unknown parameters which we will estimate from the data. The simplest distribution for time-to-event data is the *exponential distribution*, which has density:

$$f(t) = \lambda e^{-\lambda t} \quad \text{for } t > 0,$$

survival function:

$$S(t) = 1 - \int_0^t \lambda e^{-\lambda u}\, du = e^{-\lambda t},$$

and mean survival time $\frac{1}{\lambda}$. The hazard function is, therefore:

$$h(t) = \frac{f(t)}{S(t)} = \lambda,$$

that is, the hazard is constant.

Given some dataset, we want to be able to find an estimate for λ (or the parameters of our distribution of choice).

0.6.1 Maximum Likelihood Estimator

The first parametric approach we can use is the maximum likelihood estimator.

Suppose our dataset has n times $t_1, t_2, \ldots, t_n$ fully observed and $n - m$ are censored. We can create a set of indicators $\delta_1, \delta_2, \ldots, \delta_n$, where $\delta_i = 1$ if observation i is fully observed, and $\delta_i = 0$ if it is censored.

Usually, the likelihood function is computed by multiplying the density function evaluated at each observed time (with the parameter(s)) by the survival function for the censored times. For the censored times (those for which $\delta_i = 0$), we only know that the event time is greater than t_i. For these observations, it is the survival function, $P(T > t_i)$, that contributes to the likelihood.

Therefore, for any probability distribution, we have:

$$L = \prod_{i=1}^{n} \big[f(t_i)\big]^{\delta_i} \, \big[S(t_i)\big]^{(1-\delta_i)}. \quad (7.1)$$

If we assume $T \sim \text{Exp}(\lambda)$, then the log-likelihood is:

$$\ell(\text{data}) = \sum_{i=1}^{n} \Big[\delta_i \big(\log \lambda - \lambda t_i\big) - (1 - \delta_i)\, \lambda t_i\Big].$$

From this, we can find the MLE:

$$\hat{\lambda} = \frac{\sum_{i=1}^{n} \delta_i}{\sum_{i=1}^{n} t_i}.$$

We can fit an exponential distribution to the data simply by **estimating the MLE**.

7.2.3 The Weibull distribution

Having only one parameter, the exponential distribution is not very flexible and often doesn't fit data at all well. A related but more suitable distribution is the *Weibull distribution.*

The probability density function of a Weibull random variable is

$$f(t \mid \lambda, \gamma) = \begin{cases} \lambda\gamma t^{\gamma-1} \exp\big[-\lambda t^{\gamma}\big] & \text{for } t \geq 0, \\ 0 & \text{otherwise.} \end{cases}$$

Here, γ is the shape parameter, and λ is the scale parameter. If $\gamma = 1$, then it reduces to an exponential distribution. **This is not needed here** because from the trial scenario, it is thought that for patients in the control group, a reasonable approximation to the time to mortality is given by the exponential distribution.

0.7 Comparing survival curves

Really, what we would like to be able to do is to compare two survival curves (showing, for example, the results from different treatments) so that we can say whether one is significantly different from the other. In most cases, this boils down to constructing a hypothesis test along the lines of:

$$H_0 : \text{the treatments are the same}$$

$$H_1 : \text{the treatments are different.}$$

There are various ways to do this, and we will look at some now.

0.7.1 Parametric: likelihood ratio test

For a parametric analysis, our null hypothesis that the two treatments are the same can be reduced to a test of whether the parameter(s) for each group are the same. We can do this using a likelihood ratio test. We have already calculated the log-likelihood for the exponential distribution in the MLE section.

$$\ell(\lambda) = m\log\lambda - \sum_{i=1}^{n} t_i$$

$$\hat{\lambda} = \frac{m}{\sum_{i=1}^{n} t_i}.$$

Working with the exponential distribution, we can model the survival function as:

$$S(t) = e^{-\lambda_C t} \quad \text{for participants in group } C,$$

$$S(t) = e^{-\lambda_T t} \quad \text{for participants in group } T,$$

and the null hypothesis boils down to

$$H_0 : \lambda_C = \lambda_T = \lambda.$$

We can then adapt the log-likelihood we found in the MLE section in light of the separate groups, and we can then perform a maximum likelihood test by finding:

$$\lambda_{LR} = 2\Big(m_C \log\big(\tfrac{m_C}{t_C^+}\big) \; + \; m_T \log\big(\tfrac{m_T}{t_T^+}\big) \; - \; m \log\big(\tfrac{m}{t^+}\big)\Big).$$

and we refer this value to a χ_1^2 distribution.

We can also find a confidence interval for the difference between λ_T and λ_C by using the asymptotic variances of the MLEs, which are $\frac{\lambda_T^2}{m_T}$ and $\frac{\lambda_C^2}{m_C}$. Therefore, the limits of a $100(1-\alpha)\%$ confidence interval for $(\lambda_T - \lambda_C)$ are given by

$$\frac{m_T}{t_T^+} \; - \; \frac{m_C}{t_C^+} \; \pm \; z_{\alpha/2}\sqrt{\frac{m_T}{(t_T^+)^2} \; + \; \frac{m_C}{(t_C^+)^2}}.$$

One feature of the exponential model that is convenient is that the hazard function is constant. The hazard ratio often summarises comparisons between treatment groups in survival trials: the ratio of the hazard functions for the two groups. In general, this is a function of t, but for two exponential hazard functions, it is simply the ratio of the λ values.

We could also perform LR tests with the fitted Weibull distributions (**remember we don't need to do this here...**), but instead, we will continue through some more commonly used methods.

0.7.2 Non-parametric: the log-rank test

The log-rank test is performed by creating a series of tables and combining the information to find a test statistic.

We work through each time t_j at which an event is observed (by which we mean a death or equivalent, not a censoring) in either of the groups.

For notation, we will say that at time t_j:

- n_j patients are 'at risk' of the event,
- d_j events are observed (often the 'event' is death, so we will sometimes say this).

For groups C and T, we would, therefore, have a table representing the state of things at time t_j, with this general form:

Group	**No. surviving**	**No. events**	**No. at risk**
Treatment	$n_{Tj} - d_{Tj}$	d_{Cj}	n_{Cj}
Control	$n_{Cj} - d_{Cj}$	d_{Tj}	n_{Tj}
Total	$n_j - d_j$	d_j	n_j

Harder, Less Commonly Used Method:

Under H_0, we expect the deaths (or events) to be distributed proportionally between groups C and T, and so the expected number of events in group X (C or T) at time t_j is

$$e_{Xj} = n_{Xj} \times \frac{d_j}{n_j}.$$

This means that

$$e_{Cj} + e_{Tj} = d_{Cj} + d_{Tj} = d_j.$$

If we take the margins of the table (by which we mean n_j, d_j, n_{Cj} and n_{Tj}) as fixed, then d_{Cj} has a **hypergeometric distribution**, which has its mean and variance.

In the notation of our table at time t_j, we have

$$\mathbb{E}(d_{Cj}) = e_{Cj} = n_{Cj} \times \frac{d_j}{n_j}$$

$$\text{var}(d_{Cj}) = v_{Cj} = \frac{d_j n_{Cj} n_{Tj} (n_j - d_j)}{n_j^2 (n_j - 1)}$$

With the marginal totals fixed, the value of d_{Cj} fixes the other three elements of the table, so considering this one variable is enough.

Under H_0, the numbers dying at successive times are independent, so

$$U = \sum_j (d_{Cj} - e_{Cj})$$

will (asymptotically) have a normal distribution, with

$$U \sim N\left(0, \sum_j v_{Cj}\right).$$

We label $V = \sum_j v_{Cj}$, and in the log-rank test we refer $\frac{U^2}{V}$ to χ^2_1.

A somewhat simpler, and more commonly used method:

A somewhat simpler and more commonly used version of the log-rank test uses the fact that under H_0, the expected number of events (e.g., deaths) in group X is:

$$E_X = \sum_j e_{Xj},$$

and the observed number is:

$$O_X = \sum_j d_{Xj}.$$

The standard χ^2 test formula can then be applied, and the test statistic is:

$$\frac{(O_C - E_C)^2}{E_C} + \frac{(O_T - E_T)^2}{E_T}.$$

It turns out that this test statistic is always smaller than $\frac{U^2}{V}$, so this test is slightly more conservative (i.e., it has a larger p-value).

Notice that for both of these test statistics, the actual difference between observed and expected is used, not the absolute difference. Therefore, **if the differences change in sign over time**, the values are likely to cancel out (at least to some extent), and **the log-rank test is not appropriate**.

0.7.3 Semi-parametric: the proportional hazards model

As with continuous and binary outcome variables, what we would really like to be able to do is to adjust our model for baseline covariates. It seems intuitively reasonable to suppose that factors like age, sex, disease status, etc. might affect someone's chances of survival (or whatever event we're concerned with).

The conventional way to do this is using a proportional hazards model, where we assume that

$$h_T(t) = \psi h_C(t).$$

General Proportions Hazard Model

For any $t > 0$ and some constant $\psi > 0$, we call ψ the *relative hazard* or *hazard ratio*. If $\psi < 1$, then the hazard at time t under treatment T is smaller than under control C. If $\psi > 1$, then the hazard at time t is greater in group T than in group C. The important point is that ψ doesn't depend on t. The hazard for a particular patient might be greater than for another, due to things like their age, disease history, treatment group and so on, but the extent of this difference doesn't change over time.

We can adopt the concept of a *baseline hazard function* $h_0(t)$, where for someone in group C (for now), their hazard at time t is $h_0(t)$, and for someone in group T it is $\psi h_0(t)$. Since we must have $\psi > 0$, it makes sense to set

$$\psi = e^\beta,$$

so that $\beta = \log \psi$ and $\psi > 0 \ \forall \ \beta \in \mathbb{R}$. Note that $\beta > 0 \iff \psi > 1$.

We can now (re-)introduce our usual indicator variable G_i, where

$$G_i = \begin{cases} 0 & \text{if participant } i \text{ is in group } C \\ 1 & \text{if participant } i \text{ is in group } T \end{cases}$$

and model the hazard function for participant i as

$$h_i(t) = \exp[\tau G_i] h_0(t).$$

This is the proportional hazards model for the comparison of two groups. Now, the relative hazard is a function of the participant's characteristics. Naturally, we can extend it to include other baseline covariates, as we have with linear models in ANCOVA and logistic regression.

8.3.1 General proportional hazards model

Extending the model to include baseline covariates $X_1, \ldots, X_p$, we have

$$\psi(\mathbf{x}_i) = \exp(\tau G_i + \beta_1 x_{1i} + \ldots + \beta_p x_{pi}) = \exp(\mathbf{x}_i^T \boldsymbol{\beta}),$$

where we collect τ into $\boldsymbol{\beta}$ and G into $\mathbf{x}$, and the hazard function for participant i is

$$h_i(t) = \psi(\mathbf{x}_i)\, h_0(t).$$

The linear component $\mathbf{x}_i^T \boldsymbol{\beta}$ is often called the **risk score** or **prognostic index** for participant i.

The general form of the model is, therefore:

$$h_i(t) = \exp[\mathbf{x}_i^T \boldsymbol{\beta}]\, h_0(t), \tag{8.2}$$

and we can rewrite it as:

$$\log\left(\frac{h_1(t)}{h_0(t)}\right) = \mathbf{x}_i^T \boldsymbol{\beta}.$$

Notice that there is no constant in the linear term – if there was, it could just be absorbed into the baseline hazard function.

There are ways of fitting this model that rely on specifying the hazard function using parametric methods, but the method we will study (and **the most widely used**) is the one **developed Cox Regression**...

Interpreting Hazards Model

Since our primary interest is in comparing the effect of some new treatment with that of the control, we must understand what the coefficients mean and, in particular, how they relate to the treatment effect. Let's do that (as usual) by considering two participants who are identical in all baseline covariates, one in group C and one in group T. We have

$$h_i^C(t) = \exp(\beta_1 x_{1i} + \ldots + \beta_p x_{pi}) h_0(t) \qquad \text{in group C}$$

$$h_i^T(t) = \exp(\tau + \beta_1 x_{1i} + \ldots + \beta_p x_{pi}) h_0(t) \qquad \text{in group T.}$$

From this, we can find the **hazard ratio** at time t for the two treatments to be

$$\frac{h_i^T(t)}{h_i^C(t)} = \exp(\tau)$$

and τ is the log of the hazard ratio for the two treatments, adjusting for the other covariates. A value of $\tau = 0$ implies a hazard ratio of one and no evidence of difference between the treatments.

0.7.4 Cox Regression

The beauty of Cox regression is that it avoids specifying a form for the baseline hazard, $h_0(t)$, altogether. We don't need to estimate this hazard to make inferences about the hazard ratio:

$$\frac{h_i(t)}{h_0(t)}.$$

Moving on, if we set

$$\delta_i = \begin{cases} 0 & \text{if individual } i \text{ is censored} \\ 1 & \text{if individual } i \text{ is observed} \end{cases}$$

then, we can write Equation (8.3) as

$$L(\boldsymbol{\beta} \mid \text{data}) = \prod_{i=1}^{n} \left(\frac{\exp[\mathbf{x}_i^T \boldsymbol{\beta}]}{\sum_{l \in R(t_i)} \exp[\mathbf{x}_l^T \boldsymbol{\beta}]} \right)^{\delta_i},$$

where $R(t_i)$ is the risk set at time t_i.

From this, we can find the log-likelihood

$$\ell(\boldsymbol{\beta} \mid \text{data}) = \sum_{i=1}^{n} \delta_i \left[\mathbf{x}_i^T \boldsymbol{\beta} - \log \sum_{l \in R(t_i)} \exp(\mathbf{x}_l^T \boldsymbol{\beta}) \right].$$

The MLE $\hat{\boldsymbol{\beta}}$ is found using numerical methods, often Newton-Raphson.

How can we tell if a proportional hazards model is appropriate?

We can't easily visualise the hazard function for a dataset and instead would plot the survival curve. **So, can we tell if the proportional hazards assumption is met by looking at the survival curve?**

It turns out that if two hazard functions are proportional, their survival functions won't cross one another.

The important thing is that the survival curves do not cross. This is an informal conclusion, and lines not crossing is a necessary condition but not a sufficient one. It may also be that the survival curves cross when a particular [influential] covariate is factored out but not when it isn't.

0.7.5 Other Cox Regression Diagnostics

Having fit a Cox proportional hazards model, it's important to check that it is an appropriate fit to the data. We've seen already that the survival curves mustn't cross, but there are other, more sophisticated methods we can use to assess the model.

It is important to examine the proportional hazards assumption for every covariate we include in the model (including the group/ arm variable), and how we do this depends on whether the covariate is continuous or categorical.

Continuous Variables

There are partial residuals, known as Schoenfeld residuals, that can be used to assess whether the proportional hazards assumption is appropriate for a continuous variable.

We can think of $X_i = (X_{i1}, \ldots, X_{ip})'$, the set of covariates for a participant who experiences the event at time t_i, as a random variable:

$$E(X_{ij} \mid R_i) = \frac{\sum_{k \in R_i} X_{kj} \exp(\boldsymbol{\beta}' X_k)}{\sum_{k \in R_i} \exp(\boldsymbol{\beta}' X_k)},$$

where R_i are the indices of those at risk at time t_i. You can think of this as the average of the $X_{\cdot j}$ values of those at risk at time t_i, weighted by their relative hazard. We can write

$$\hat{E}(X_{ij} \mid R_i)$$

to denote this quantity with the MLE $\hat{\boldsymbol{\beta}}$ substituted for $\boldsymbol{\beta}$.

The partial residual at time t_i is therefore the vector

$$\hat{r} = (\hat{r}_{i1}, \ldots, \hat{r}_{ip}),$$

where

$$\hat{r}_{ik} = X_{ik} - \hat{E}(X_{ik} \mid R_i).$$

If we plot the Schoenfeld residuals against time, we should see a random scatter around zero (the kind of plot we look for when assessing residuals against fitted values of a linear regression model).

There is a proposed statistical test using the Schoenfeld residuals, in which the null hypothesis is that the proportional hazards assumption holds. This can be implemented by the function `cox.zph` in the survival package.

Categorical Variables

Recall from Equation (8.4) that

$$S(t) = \exp(-H(t)),$$

where $H(t)$ is the cumulative hazard function

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

From this we find that

$$\log(S(t)) = -H(t).$$

If we have two groups A and B for which the proportional hazards assumption is satisfied, then for some constant ψ

$$H_A(t) = \psi H_B(t).$$

We can combine these two equations to find

$$\log[-\log(S_A(t))] = \log(H_A(t)) = \log\psi + \log(H_B(t)) = \log\psi + \log[-\log(S_B(t))].$$

Under the Cox Regression model, $\log(H(t))$ is **linear in the covariates**, and so we will have [roughly] parallel lines. We can plot this in R using `ggsurvplot` and setting `fun="cloglog"`.