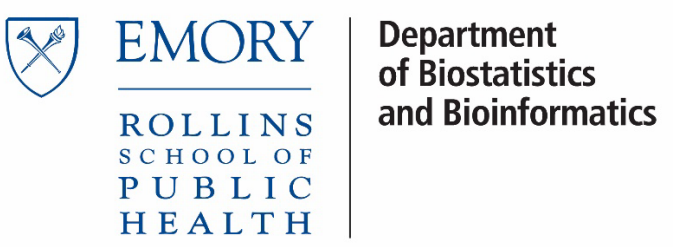


BIOS 522: Survival Analysis Methods

Reading 11:

Survival analysis in clinical trials

This week, we will discuss important key concepts in clinical trials. We will then focus on sample size and power calculations for clinical trials with survival outcomes. We will discuss clinical trial monitoring.

Part 1. Introduction to clinical trials

The importance of randomization

In **observational studies**, there may be important differences between individuals who receive the intervention and individuals who do not. Where these differences are also associated with differences in the survival outcome, there is **confounding** in the relationship between treatment and survival.

For example, in a developing country, children from the poorest communities may be the least likely to be vaccinated. These children may also be the most highly exposed to infectious diseases. Even if the vaccine has no effect on preventing infection, vaccinated individuals may have a lower rate of infection if they are also people who are generally at lower risk. We can adjust for measured covariates, such as a measure of household income, but there may be **unmeasured confounders** that cloud the effect.

Randomized clinical trials offer the clearest evidence regarding whether or not an intervention is effective. Because the intervention is randomly assigned, the groups will be well-balanced on average. Furthermore, there is no potential for measured or unmeasured confounding.

Clinical trial phases

For investigational products (e.g., treatments, vaccines, medical devices), clinical trial evaluation proceeds through several phases.

- **Pre-clinical**: animal and laboratory testing.
- **Phase 1**: small, first-in-human trials to assess safety.
- **Phase 2**: larger trials to assess safety and early evidence of efficacy.
- **Phase 3**: large trial to assess efficacy and safety to inform licensure.

Phase 3 trials can include hundreds, thousands, or tens of thousands of participants. For this reason, it is important that they are adequately designed in order to address the scientific question of interest. Often this question is whether or not the product works better than the best currently available alternative.

Clinical trial endpoints

The **primary endpoint** in a Phase 3 trial should directly measure the health outcome that is of greatest scientific interest. In trials with a time-to-event endpoint, the primary endpoint may be time until death, death from a specific cause, incidence of disease, a complication or specific adverse effect of interest, or symptomatic relief.

The selected primary endpoint may reflect a mixture of scientific and practical considerations. For example, while we may be most interested in a narrowly defined endpoint (death due to cancer), this could occur too rarely to be practical. We may select a broader or composite endpoint as a compromise.

Primary analysis

Most clinical trials are analysed by the **intention-to-treat (ITT)** principle. After randomization, clinical trial participants may be non-compliant with the intervention (e.g. not take all doses). There may be protocol deviations (e.g. individuals receive incorrect dosing). Participants may withdraw.

In an ITT analysis, participants are analysed according to their randomized *treatment assignment*, not the treatment actually received. It ignores any protocol deviations that occurs post-randomization. The purpose of the ITT analysis is to maintain the balance in prognostic factors generated from the original randomization.

For example, if a new treatment causes severe nausea such that patients are less likely to be compliant, this could diminish the potency of the drug, but we want to capture this effect in the analysis. This will be reflected in the ITT analysis but not in analysis restricted to compliant patients.

In contrast, the **per protocol** analysis includes only trial participants who completed the study without major protocol violations.

In vaccine trials, the per protocol analysis is also used to describe an analysis that excludes all participants who develop disease before all vaccine doses have been received. The modified time origin may be 7 or 14 days after the last dose. The purpose of this analysis is to measure the effect of the completed vaccine regimen, after an individual's immune system has been fully activated.

Part 2. Sample size and power

Hypothesis testing in clinical trials

If we have two survival distributions, $S_0(t)$ and $S_1(t)$, we might be interested in testing the null hypothesis that these functions are equal:

$$H_0: S_0(t) = S_1(t)$$

To test this null hypothesis, we commonly use the **log-rank test** to test if the two observed survival curves are equal.

We calculate the **log-rank test statistic** Z that sums over the differences between the groups across all failure times. We compare this test to a standard normal distribution in order to calculate a **p-value**. Recall that the p-value is the probability of observing data as or more extreme than what was observed, assuming that the null hypothesis is true.

Based on the p-value, we decide whether to **reject** or **not reject** H_0. We compare the p-value to our pre-specified **significance level** α.

If $p < \alpha$, reject H_0

If $p \geq \alpha$, do not reject H_0

The most commonly used significance level is $\alpha = 0.05$.

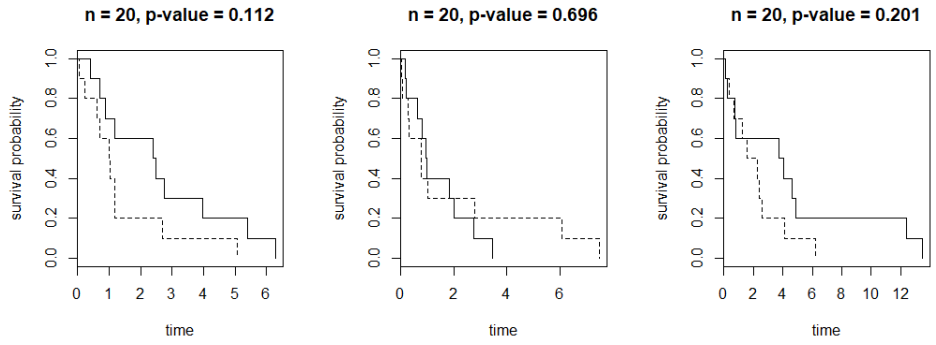
Recall that rejecting the null hypothesis when it is true is called a **type I error**. When the null hypothesis is true, the probability of rejecting the null and declaring significance by pure random chance is set so that it is no larger than α (e.g. 0.05). It is very important to limit or **control** the type I error rate. If a treatment has no benefit on survival, we want our study to correctly conclude no effect (true negative). If an ineffective product is licensed, it is expensive, and it could cause toxicity in patients who take it without providing any benefit. It also becomes difficult to later sort out if the treatment actually works, because it might no longer be ethical to implement a placebo-controlled trial.

On the other hand, if there is a true beneficial effect of an intervention, we want to make sure we are able to detect it. False negatives (**type II errors**) occur. If

we pick a sample size that is too small, we may not be able to detect a statistically significant difference in survival, even though a clinically meaningful effect exists. The end result is that we may miss out on a potentially life-saving intervention.

Example: Imagine a clinical trial in which participants are randomized in a 1:1 ratio to a new intervention or placebo control. For participants receiving placebo, imagine that their survival follows an exponential distribution with rate parameter 0.6 $days^{-1}$. For participants receiving the intervention, their survival follows an exponential with rate parameter 0.3 $days^{-1}$. The true hazard ratio for participants receiving the new intervention versus placebo is 0.3/0.6=0.5. In our setting, this reduction in hazard is clinically meaningful.

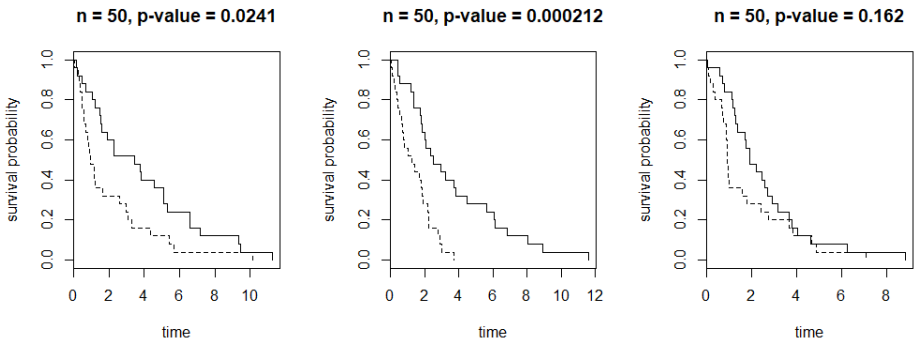
We can use simulations to generate *three hypothetical trials* from the same scenario. The differences between them are just due to random chance. First, imagine that each trial has sample size $n = 20$ (10 participants in each arm). The Kaplan-Meier curves and log-rank test p-values are reported. The dashed line is the placebo arm, and the solid line is the intervention arm.



Even though we know that the true effect of the intervention is to reduce the hazard by 50%, the sample size is too small to detect the effect. None of the log-rank tests are statistically significant (p>0.05 for each trial). These are type II errors (false negatives).

In fact, at this sample size, we will return a false negative 66% of the time. Only 34% of the time will we (correctly) declare that there is a significant difference between the two groups.

Now imagine increasing the sample size to $n = 50$ (25 participants in each arm). We generate three more hypothetical trials with the same hazard functions but a larger sample size. The Kaplan-Meier curves and log-rank test p-values are reported.



We see that two of the three hypothetical trials demonstrate a significant effect of treatment. These are true positives. But the third, by chance, is non-significant. This is a type II error.

At this sample size, we will return a false negative 31% of the time. Only 69% of the time will we detect a true effect.

As we increase the sample size further, the chance of a type II error decreases, and the chance of a true positive increases. We refer to the probability of a true positive as the **power** of our study.

Total sample size	**Power**	**Type II error**
$n = 20$	34%	66%
$n = 30$	48%	52%
$n = 40$	59%	41%
$n = 50$	69%	31%
$n = 60$	77%	23%
$n = 70$	83%	17%
$n = 80$	87%	13%
$n = 90$	91%	9%
$n = 100$	93%	7%

Assuming that the true hazard ratio is 0.5, we need a sample size of at least $n = 70$ to detect a significant effect 80% of the time. We need a sample size of at least $n = 90$ to detect a significant effect 90% of the time.

The power depends on the underlying true hazard ratio. It is easier to detect a large effect (e.g. hazard ratio of 0.3, which is further from the null value of 1), meaning we will have higher power at the same sample size. Conversely, it is harder to detect a small effect (e.g. hazard ratio of 0.7, closer to the null value of 1), meaning we will have lower power at the same sample size.

The table below summarizes power at the same sample sizes assuming three different hazard ratios. The strength of the effect increases as we move from left to right, and so does the power at the same sample size

Total sample size	**Power** $\boldsymbol{HR = 0.7}$	**Power** $\boldsymbol{HR = 0.5}$	**Power** $\boldsymbol{HR = 0.3}$
$n = 20$	12%	34%	77%
$n = 30$	16%	48%	91%
$n = 40$	20%	59%	97%
$n = 50$	24%	69%	99%
$n = 60$	28%	77%	>99%
$n = 70$	32%	83%	>99%
$n = 80$	36%	87%	>99%
$n = 90$	39%	91%	>99%
$n = 100$	43%	93%	>99%

■

Trial planning

It is very important that studies are designed to be able to detect a clinically meaningful effect where one exists. It is wasteful to implement trials that are too small to detect a difference. It is also unethical, when one considers that trials involve human volunteers/participants.

This motivates the need for **sample size and power calculations**. When planning a trial with a time-to-event outcome, we can calculate the sample size we need to achieve a particular power using the log-rank test. Power is the probability of a true positive given that a clinically meaningful effect exists. The definition of a clinically meaningful effect will vary from setting to setting.

Power depends on several key elements, including:

- The type I error α
- The effect size, e.g. the hazard ratio θ

Interestingly, in time-to-event trials, the power does not directly depend on the sample size. Instead, it depends on the **expected number of events/failures $\boldsymbol{d}$ during the study period**. Of course, a natural way to increase the expected number of events/failures is to increase the sample size n, but the calculations involve d and not n. Another way to increase the expected number of events/failures is to follow each participant longer. Both of these modifications increase the overall cost of the trial.

The typical process for designing a clinical trial:

1) Select the primary endpoint and analysis
2) Determine the desired power
3) Identify the smallest effect size that is clinically meaningful

4) Calculate the required number of events based on this power and effect size
5) Establish the incidence rate in the control population
6) Determine the *number of participants to enroll* and the *length of time they should be followed*
7) Adjust for losses to follow-up, etc.

Let's review each step in turn…

Step 1) Select the primary endpoint and analysis

The choice of primary endpoint (e.g. cause-specific mortality) is context-specific and will depend on the health outcome of greatest scientific interest.

For time-to-event outcomes, the primary analysis is usually by a log-rank test comparing the survival functions. Our objective is to test:

$$H_0: S_1(t) = S_0(t)$$

Suppose that the hazard functions in the two groups are $\lambda_0(t)$ and $\lambda_1(t)$, with constant hazard ratio:

$$\theta = \frac{\lambda_1(t)}{\lambda_0(t)}$$

We are interested in testing $H_0: \theta = 1$ or equivalently that $H_0: \log(\theta) = 0$.

Step 2) Determine the desired power

The desired power is usually set at 80% or 90% by convention. Higher power is always preferred, but may require a prohibitively expensive trial. Power lower than 80% may be considered too risky.

We denote our type II error probability as β, thus our power is $1 - \beta$. For power of 80%, $\beta = 0.20$ and $1 - \beta = 0.80$.

Step 3) Identify the smallest effect size that is clinically meaningful

The selection of a clinically meaningful effect size is context-specific. It depends on many factors, including the effect of the best available alternative, the safety profile of the intervention, the risk/benefit ratio in the target population, and other practical considerations (e.g. ease of use, cost, availability).

For example, for a vaccine trial, we usually desire vaccine efficacy of 70%, which translates to a hazard ratio of 0.3. Vaccines are typically provided to

healthy individuals, and so the vaccine must be highly effective to balance out any potential risks.

For a therapeutic to treat a fatal illness, a lower effect size might be acceptable, such as a hazard ratio of 0.9 (10% reduction). There is a different risk/benefit ratio for very sick individuals.

A larger sample size is required to detect a smaller effect. In our previous example, we needed a sample size of about $n = 90$ to have 90% power to detect a hazard ratio of 0.5, while we needed a sample size of about $n = 30$ to have 90% power to detect a hazard ratio of 0.3. Often the assumed effect size in a trial is a compromise between scientific goals and practical considerations.

While in our calculations we express the desired effect size in terms of the hazard ratio θ, we may start off thinking about the effect size on a different scale. We may want to improve 5-year survival by 25%. Or we may want to double median survival. If we make the simplistic assumption that our data follow an exponential distribution, we can convert these types of effects into clinically meaningful hazard ratios

Example: Suppose we want to detect a 50% improvement in median survival from 12 months to 18 months. If we assume our data follow an exponential distribution with rate λ_i, recall that the median for an exponential is:

$$t_{0.50.i} = \frac{\log 2}{\lambda_i}$$

Therefore, the rate λ_i is a function of the median:

$$\lambda_i = \frac{\log 2}{t_{0.50,i}}$$

For the control group with median survival of 12 months, the rate is:

$$\lambda_0 = \frac{\log 2}{12} = 0.0578 \text{ months}^{-1}$$

For the intervention group with median survival of 18 months, the rate is:

$$\lambda_1 = \frac{\log 2}{18} = 0.0385 \text{ months}^{-1}$$

Thus, the hazard ratio corresponding to this effect is:

$$\theta = \frac{\lambda_1}{\lambda_0} = \frac{\frac{\log 2}{18}}{\frac{\log 2}{12}} = \frac{12}{18} = \frac{1}{1.5} = 0.667$$

Notably, we would return the same hazard ratio of 0.667 for any 50% improvement in median survival, regardless of the median survival in the control group.

■

Step 4) Calculate the required number of events

Given our assumed hazard ratio θ, the total number of events d required to achieve our desired power $1-\beta$ by a two-sided level α log-rank test is approximately:

$$d = \frac{4\left(z_{1-\alpha/2} + z_{1-\beta}\right)^2}{[\log(\theta)]^2}$$

In the numerator, $z_{1-\alpha/2} = 1.96$ for a two-sided test with $\alpha = 0.05$. $z_{1-\beta} = 0.842$ when power is 80%, or $z_{1-\beta} = 1.282$ when power is 90%.

Example: To detect a hazard ratio of 0.7 with 90% power at a 2-sided significance level of $\alpha = 0.05$, we need the following number of events:

$$\begin{aligned} d &= \frac{4\left(z_{1-\alpha/2} + z_{1-\beta}\right)^2}{[\log(\theta)]^2} \\ &= \frac{4(1.96 + 1.282)^2}{[\log(0.7)]^2} \\ &\approx \frac{42}{0.1271} = 331 \end{aligned}$$

■

The following table summarizes the number of events required for various hazard ratios. By convention, we always round up in sample size calculations.

Hazard ratio θ	80% power	90% power
0.7	247	331
0.6	121	162
0.5	66	88
0.4	38	51
0.3	22	29

Step 5) Establish the incidence rate in the control population

In order to determine how many participants we must enroll to achieve the required number of events, we must establish the event rate in the control

population. For reasons of simplicity, it is common to assume that the control group times follow an exponential distribution with rate λ_0.

Example: If the median survival in the control group is 12 months, the rate is:

$$\lambda_0 = \frac{\log 2}{12} = 0.0578 \text{ months}^{-1}$$

■

Example: If the five-year failure survival probability is 80%, we can convert this to a rate. Recall that for an exponential distribution:

$$S(t) = \exp(-\lambda_0 t)$$

Therefore, if the survival probability at time $t = 5$ years is 80%:

$$S(5) = \exp(-\lambda_0 5) = 0.80$$

Solving for λ_0:

$$\begin{aligned} \exp(-\lambda_0 5) &= 0.80 \\ -\lambda_0 5 &= \log(0.80) \\ \lambda_0 &= -\frac{\log(0.80)}{5} = 0.0446 \text{ years}^{-1} \end{aligned}$$

■

Step 6) Determine the number of participants to enroll and the length of time they should be followed

We can then use the elements above to calculate our sample size, i.e. how many participants to recruit. Recall that our goal is to design a study that observes the required number of events d, as this is the number of events needed to achieve our power. We know the rate in the control group λ_0 and the rate in the intervention group:

$$\lambda_1 = \theta \lambda_0$$

Assuming the data follow an exponential distribution, for a trial of length t_{max}, the probability that a participant in the control group has the event during the trial is calculated using the fact that $F(t) = 1 - S(t) = 1 - e^{-\lambda t}$:

$$p_0 = 1 - e^{-\lambda_0 t_{max}}$$

If there are n participants in the trial, $n/2$ of which are in the control arm, the expected number of events in the control arm is:

$$d_0 = \frac{n}{2}(p_0)$$

A similar calculation can be done for the intervention arm, with expected number of events:

$$d_1 = \frac{n}{2}(p_1) = \frac{n}{2}\left(1 - e^{-\lambda_1 t_{max}}\right)$$

To increase the number of events $d = d_0 + d_1$, there are two key strategies:

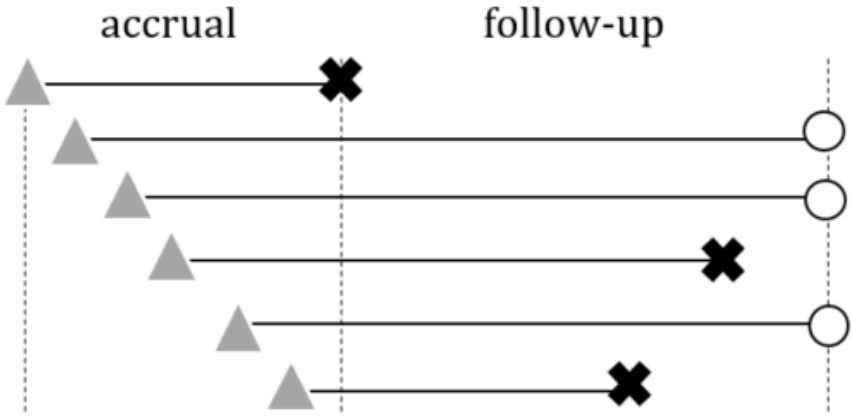
- Increase the sample size n
- Increase the duration of follow-up t_{max}

The preferred design will depend on the available resources, the number of eligible trial participants, and the expected duration of the trial.

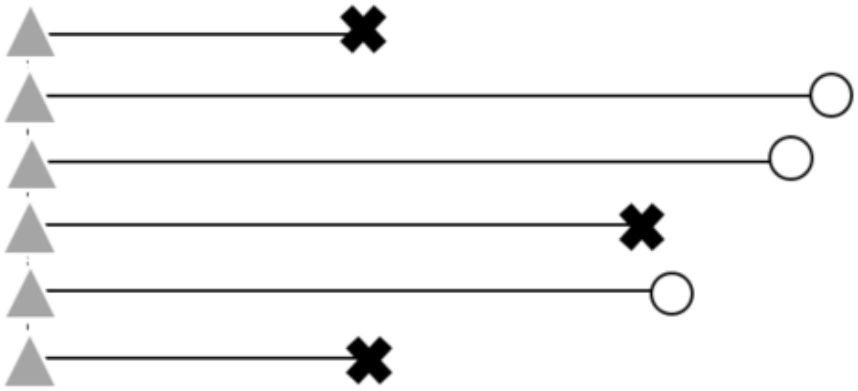
If it is important to quickly arrive at a scientific conclusion, a large trial may be conducted over a short period of time.

For interventions targeting populations with rare diseases, it may be necessary to follow a smaller population for a longer time.

In reality, not all trial participants will enter the study at time 0. Instead, the **accrual** will occur in a **staggered** manner over a period of time.



Thus, when aligned with respect to their time of study entry/time of randomization, trial participants will have differing lengths of follow-up.



Sample size calculation formulae are available that account for the additional follow-up time during this accrual period.

Step 7) Adjust for losses to follow-up, etc.

Especially for long trials, it is important to plan for losses to follow-up and other sources of missing information (e.g. laboratory failure). A simple approach is to inflate the overall sample size.

Example: We calculate that we need 100 participants in our trial, but estimated loss to follow-up over the course of the trial is 20%. Thus, our final sample size will only be 80% of the initial sample size. If we enroll 100 participants initially, only 80 will remain at the end of the trial, and our trial will be underpowered.

If we had enrolled 125 participants, $0.8(125) = 100$ participants will remain at the end of the trial, and we will be adequately powered.

A simple way to calculate the inflated sample size $n_{initial}$ is to divide our desired sample size n_{needed} by the fraction remaining at the end of the study:

$$n_{initial} = n_{needed}/0.8$$

Continuing with our previous example:

$$n_{initial} = \frac{100}{0.8} = 125$$

Thus, if we expect 20% loss to follow-up, we should enroll 125 participants.

■

Part 3. Data monitoring

In clinical trials, it is often desirable to conduct **interim analyses** of study data while data collection is ongoing. The purpose is two-fold:

- Ethical: if one treatment is substantially worse than another, then it is unethical to continue to give the inferior treatment to patients.
- Timely reporting: If the hypothesis of interest has been clearly established halfway through the study, then science and the public may benefit from early reporting.

Unplanned interim analyses (i.e. multiple testing) can seriously inflate the type I error of the trial. For example:

- If 1 test is performed at $\alpha = 0.05$, type I error for the trial is 5%.
- If 2 tests are performed at $\alpha = 0.05$, type I error for the trial is 8.3%.
- If 3 tests are performed at $\alpha = 0.05$, type I error for the trial is 10.7%.

If interim analyses are to be performed, it is important to carefully plan these in advance, and to adjust all tests appropriately so that the type I error is maintained below the pre-set limit.

What can we do to protect against this type I error inflation?

Pocock approach

We could pick a smaller significance level, say α', to use at each interim analysis so that the overall type I error stays at level α. This is the **Pocock approach**.

For example:

- If 2 tests are performed at $\alpha' = 0.0294$, type I error for the trial is 5%.
- If 3 tests are performed at $\alpha' = 0.0221$, type I error for the trial is 5%.

A problem with the Pocock method is that even the very last analysis is performed at level α'. This tends to be very conservative at the final analysis.

O'Brien and Fleming approach

A preferable approach would be to vary the α' levels used for each of the interim analyses, but try to keep the very last one "close" to the desired overall significance level (e.g. 0.05). The **O'Brien-Fleming approach** achieves this.

For example:

- If 2 tests are performed, the first is conducted at $\alpha' = 0.0054$ and the second is conducted at $\alpha' = 0.0492$. The type I error for the trial is 5%.

- If 3 tests are performed, the first is conducted at $\alpha' = 0.0006$, the second is conducted at $\alpha' = 0.0151$, and the third is conducted at $\alpha' = 0.0471$. The type I error for the trial is 5%.

In large randomized Phase 3 trials, we often have 1-3 equally spaced interim looks before the final analysis. For time-to-event data, *the timing of the looks is determined by the number of events*. If 50 events are required for full power, then a single look at the halfway point would occur after 25 events (total across both groups) have occurred. Thus, investigators monitor *accrued events* instead of *sample size* to plan the timing of their analyses.

These type of procedures in which you conduct interim analyses and decide whether to stop the trial early are known as **group sequential** or **sequential monitoring analyses**.

Some designs further allow for early stopping of the trial if it is highly unlikely that the intervention has a significant effect. For example, halfway through the trial, the groups have identical results. Early stopping for lack of effect is known as stopping for **futility**.

Part 4. Looking ahead

To close out the course, we will learn about competing risks. Competing risks analyses allow us to consider multiple events, where one event prevents the occurrence of another event. For example, where we are interested in studying time until cardiovascular death, a non-cardiovascular death is a type of competing risk. These require a different framework for analysis.