Department of Mathematical Sciences

Clinical Trials IV - (MATH4407)

Clinical Trials Report

Author:
Raul Unnithan

Supervisor:
Doctor Rachel Oughton

May 5, 2025

Declaration

This piece of work is a result of my work, and I have complied with the Department's guidance on multiple submissions and the use of AI tools. Material from the work of others not involved in the project has been acknowledged, quotations and paraphrases suitably indicated, and all uses of AI tools have been declared.

Contents

Chapter 1 Trial

1.1 Sample Size

There are two approaches to calculating sample size: formula-based and simulation-based methods. In this trial, simulation-based methods were used (see Section 2.1 for further justification).

A minimum sample size is desired because fewer participants mean reduced recruitment, treatment, monitoring, and data collection costs. It also leads to a faster completion as a smaller trial typically takes less time to recruit, run, and analyse. Also, it takes ethical considerations more into account. In trials involving terminally ill patients, it is especially important to minimise exposure to potentially ineffective treatments. Recruiting more participants than necessary could expose additional patients to unnecessary risk. Finally, trials with large sample sizes can face logistical challenges, especially when dealing with narrow inclusion criteria such as patients with a terminal illness over age 50.

However, we needed to make sure the sample size is not small because that can compromise the trial. That is, it might not pick up a true treatment effect if there is one. A weak study can give us unclear or wrong results, wasting resources and participants' time. Ethically, placing patients, especially those who have terminal illnesses, in an underpowered trial compromises the rationale for putting them at risk or withholding effective treatment from them. In simulation studies, this indicates the need to calculate the minimum number of subjects needed to achieve the desired power, not to use the minimum number.

To derive the minimum sample size required for this trial, power simulation was applied to the survival data using Cox Regression through the function `exp_sim_cox()`.

In this trial, it was assumed that survival time followed an exponential distribution in the control group. The code reflected this by assuming exponential survival for both groups, using specified hazard rates `rate_C` and `rate_T`. These were calculated based on the median survival times: for the control group, the median was 103 days, and for the treatment group, the median was assumed to be at least 35 days longer, i.e. 138 days. These medians corresponded to exponential rates, λ, via the formula:

$$\lambda = \frac{\log(2)}{\text{median}}.$$

Each participant was labelled as either control or treatment using `rep(c("C", "T"), each = N)`, consistent with the randomisation described in the trial. Although this contradicts a principle of simulation, which is to recreate as many features of the trial as possible, including the allocation, in the actual trial, the minimisation algorithm would be used, which guarantees (close to) exactly equal groups. Therefore, using a different allocation method in simulations was suitable here.

The next step was to deal with censoring. First, all participants were assumed to have died (`status = 1`) before censoring adjustments were applied.

The trial scenario mentioned that around 5% of participants were lost to follow-up. This loss was

implemented here in the simulation by randomly selecting 5% of participants and censoring them at a random time before their event (which is death here) occurred. Their status was updated to 0, and the survival time was drawn from a uniform distribution between 0 and their original death time.

Patients in the trial were also monitored for a fixed follow-up period of 9 months, equivalent to 274 days. Anyone whose simulated survival time exceeded this threshold was administratively censored, meaning the event was not observed within the follow-up window. Their time was shortened to 274 days, and their status was updated to 0 to indicate censoring.
A Cox proportional hazards (Cox PH) model was then used to assess whether the treatment arm (i.e. the control of the treatment group) was associated with a reduced hazard of death. No other baseline covariate was used because this is pre-trial, so none of this information was known. (see Subsection 2.1 for further justification of why the Cox PH model was used).
Note: using Cox PH in the simulation study also meant it was used in the results to ensure consistency across the trial.

The simulated trial aimed to detect a significant increase in survival time at a significance level of $\alpha = 0.05$. Here, the p-value for the treatment effect was extracted, and if it was less than 0.05, the null hypothesis (that the treatment has no effect) was rejected for that simulation.

After repeating the simulation `nsim` times, the estimated power was computed as the proportion of simulations where the treatment effect was statistically significant. This power reflects how likely this trial would be to detect a 35-day improvement in median survival, given the assumptions specified.

When simulating, ideally, the goal is to carry out as many simulations as possible to gauge the most accurate power for each sample size. This leads to a drawback of the simulation approach, which is that the number of simulations we can run is bound by the processing speed of the technology we have available.

In spite of this limitation, in this trial, the function `exp_sim_cox` was simulated on 100 loops, each with 10000 simulations, `nsim=10000`. After testing different sample sizes, this led to $N = 320$ being the minimum sample size that outputted a power of at least 90%. Therefore, 320 participants were needed for the control and the treatment group, and hence, 640 participants for the whole trial.

1.2 Missing Data Check

Prior to allocation, we conducted a missing data check, using the `any` function, to assess the completeness of participant information. This function returned `FALSE`, meaning no missing values were identified in the dataset at this stage, so no imputation procedures were necessary. Since participants entered the trial sequentially and were allocated immediately upon recruitment, participants entered the trial with complete data, and we proceeded directly to binning the relevant continuous variables.

1.3 Allocation

The only continuous baseline covariate in this trial was `Age`, and all participants were 50 years old and above, as specified by the trial scenario. The clinicians were also particularly interested to learn about the effect of the intervention in those aged 70 and over.

A key feature when binning is that the number of observations should spread well across each bin because this avoids bins with too few observations, leading to unreliable estimates.

Combining these two considerations meant that a suitable grouping for age was to use 2 bins in a new variable `AgeGroup`. One bin was for participants who were aged `"50-70"`, and the other was for

those who were "70+". This gave the following spread across each bin:

Age Group:	50-70	70+
Count:	328	312

Table 1.1: Age Group Distribution

Table 1.1 shows a roughly well-spread, as desired, so we were ready for allocation.
Minimisation was used over other methods for allocation over the other methods (see Subsection 2.2 for further justification). Minimisation aims to minimise differences between treatment groups. It also balances individual outcome-related factors rather than their interactions, and follows a set algorithm.

The first patient is allocated using simple randomisation. Next, patients are recruited sequentially and need to be assigned to a trial arm, where for each new patient, the level of each factor is listed. In this trial, the 2 factors were: `AgeGroup` (with levels: `"50-70"` and `"70+"`) and `DiseaseLevel` (with levels: `"Moderate"` and `"Severe"`).

Next, a sum is calculated to quantify the imbalance across all factors. For each factor level, the difference between the number of patients allocated to each treatment arm is computed. These differences are then summed across all factor levels as:

$$(n_A + x_{++} - n_B + x_{++}) + (n_A + y_{++} - n_B + y_{++}),$$

where n_A and n_B represent the number of patients in treatment arms A and B, and x, y represent the levels of the factors.

Based on the imbalance sum, the new patient is allocated to a treatment arm with a certain probability. If the sum is negative, the patient is allocated to arm A with probability P, where $P > 0.5$. If the sum is positive, the patient is allocated to arm B with probability P. If the sum is zero, the patient is allocated to arm A with a probability of P, where $\frac{1}{2} < P < 1$ to retain some randomness. The latter was used in this trial, with a probability `P=0.8`, for a strong balance across covariates while retaining enough randomness to prevent selection bias.

The minimisation algorithm here accounted for the two covariates: `DiseaseLevel` and `AgeGroup`. These were stored in the covariate matrix `covmat`. Each covariate was given equal weight, represented by `covwt`, ensuring no single covariate dominated the allocation process. There were two treatment arms in this trial, group C and group T, denoted as `trtseq` = (0,1). Finally, the allocation ratio was set at 1:1, `c(1,1)`, meaning that, ideally, patients would be distributed equally across both arms.

For all subsequent patients, the function `Minirand` was used to determine treatment allocation. This function used the covariate matrix, weighting scheme and treatment ratio to assign the new patient in a way that minimised imbalance across the stratification factors. The final treatment allocations were then stored for trial results.

The total imbalance across treatment arms after applying minimisation was `0`, indicating that minimisation successfully achieved covariate balance.

Disease Level	**Control**	**Treatment**
Moderate	207	207
Severe	113	113

Table 1.2: Group counts by Disease Level.

Age Group	**Control**	**Treatment**
50–70	164	164
70+	156	156

Table 1.3: Group counts by Age Group.

Tables 1.2 and 1.3 show the distribution of participants across treatment arms for the two stratifying variables: `DiseaseLevel` and `AgeGroup`. In both cases, the number of participants is exactly equal between the control and treatment groups within each level. This confirms that minimisation successfully achieved covariate balance, ensuring that any observed treatment effects are not confounded by imbalances in these key baseline characteristics.

1.4 Results

We first needed to deal with the (right)-censored times. If we treated 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.

Building on its use in the simulation study, the primary method used for analysing the results was Cox Regression. Cox Regression incorporates censored data through partial likelihood, which compares the relative hazard of events across individuals still at risk, allowing censored observations to contribute information up to their censoring time.

The Kaplan-Meier (KM) estimator was also used, but in a supplementary role to visualise survival curves. While it does not adjust for covariates, it incorporates censored data and provides a non-parametric estimate of the survival function, offering an initial understanding of differences between groups.

To reflect the clinicians' specific interest in patients aged 70 and over, a binary variable `Age70` was created from the baseline covariate `Age`, assigning a value of 1 to participants aged 70 or above and 0 otherwise. As age was recorded at baseline, prior to treatment assignment and outcome observation, this derived variable is valid for use throughout the analysis (so it would not introduce post hoc bias). It was used in stratified KM plots and proportional hazards (PH) assumption checks, and could be included in Cox regression models either as a covariate or as a stratification factor.

1.4.1 Kaplan-Meier Estimator Plots

The KM survival plot is stratified by just the treatment arm and `Age70`. This allows for a descriptive assessment of survival patterns across treatment arms for those aged 70 and over, which the clinicians are particularly interested to learn about.

The reason this is only a descriptive assessment is that this analysis was not pre-specified, and no formal statistical interaction was tested. Therefore, the plot is intended as merely exploratory.
Note: Disease severity was not included in the KM plots to preserve visual clarity and focus on age group and treatment arm comparisons, which were most relevant to the clinicians' stated interest. However, the effect of disease severity was accounted for in the Cox Regression analysis.

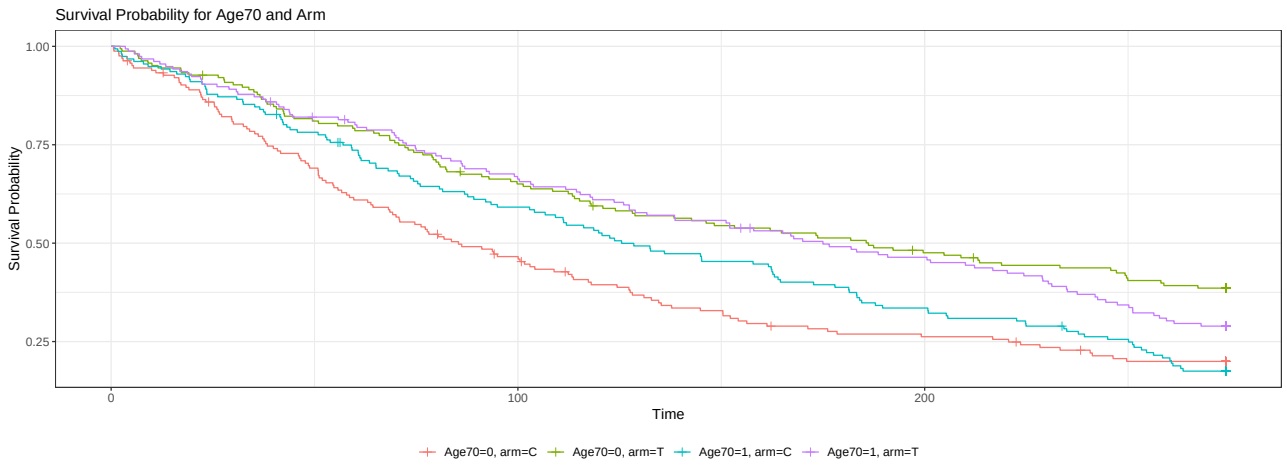

Figure 1.1: Kaplan-Meier survival curves stratified by `Age70` and `arm`.

Figure 1.1 displays that the treatment group generally showed higher survival probabilities than the control group within both age strata (i.e. the treatment arm lies above the control arm throughout follow-up). Hence, from this consistent separation and ordering, Figure 1.1 suggests that the treatment may offer survival benefits across both `Age70` subgroups, including patients aged 70 and over.
Note: some survival curve overlap across the two `Age70` subgroups in the same `arm` is expected due to natural variability and inherent age-related differences in mortality; for instance, younger patients typically face a lower risk of death.

The KM curve also showed steady, stepwise declines across all subgroups, with no extended plateaus, indicating a relatively consistent occurrence of events throughout the follow-up period. Each step corresponds to one or more events, and the absence of long plateaus suggests no prolonged intervals without observed deaths.

In addition, the survival curves within each `Age70` subgroup for each arm do not cross, indicating the PH assumption (of constant hazard ratios over time) is valid, justifying the use of Cox Regression.

The KM curves were also examined on a log-log scale, for `Age70` and each arm, to assess the PH assumption further informally.

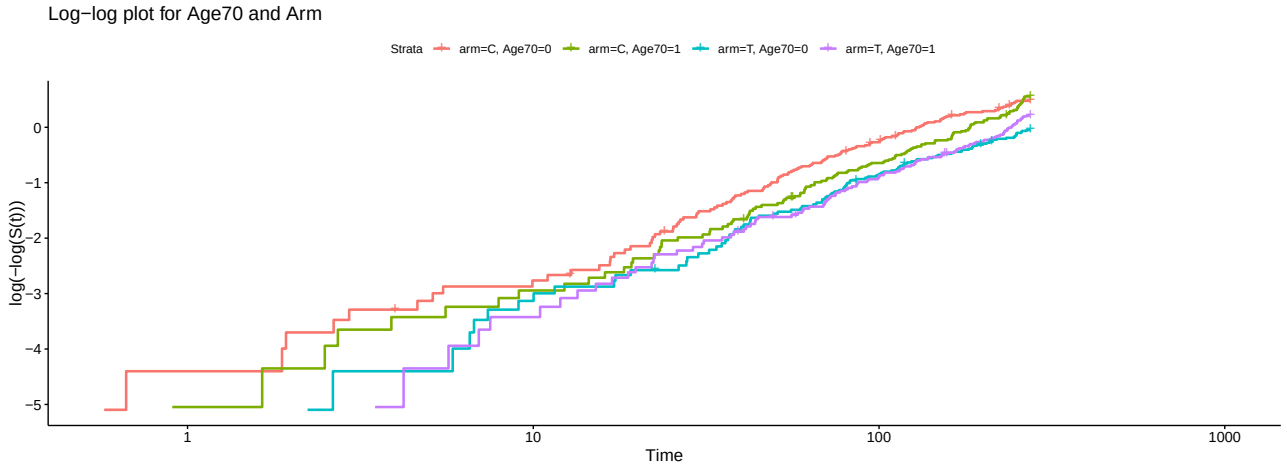

Figure 1.2: Log-log plot for assessing the proportional hazards assumption.

These lines appear roughly parallel. Therefore, the Cox model's PH assumption may be valid (reinforcing the survival probability plot in Figure 1.1.

The brief touchpoint between the treatment at Time ≈ 25 days and control curves (for `Age70=1`) does not indicate a violation of the PH assumption. The log-log plot looks at overall trends, and the curves remain broadly parallel and preserve their relative ordering, which supports the Cox model's validity.

Censoring marks appeared intermittently across both Figures 1.1 and 1.2, without clustering in any specific group or period. This pattern aligns with the trial's expectation that approximately 5% of participants would be lost to follow-up and suggests that most censoring was administrative (due to survival beyond the 9-month follow-up) rather than informative.

1.4.2 Cox Regression

Although the non-parametric paradigm is prevalent in survival analysis, this trial used Cox Regression, a semi-parametric method, to analyse results (see Subsection 2.1 for further reasoning).

`DiseaseLevel`, `Age` and `arm` were used as covariates in the Cox PH model. Before proceeding with subsequent results, the PH assumption had to be checked for the model, as the prior KM estimator plots only gave informal indicators of this. For Cox PH, this was done via the Global Schonfeld Residual Test. A 5% significance level was used in comparison, as this is a common benchmark.

Table 1.4: Global and Individual Schoenfeld Residual Tests

Covariate	χ^2	**df**	**p-value**
DiseaseLevel	2.0017	1	0.15712
Age	14.7331	1	0.00012
arm	0.0613	1	0.80438
Global	17.9833	3	0.00044

Table 1.4 shows the results of the Global and Individual Schoenfeld Residual Tests for the PH assumption. The global test was significant, indicating strong evidence that the PH assumption may be violated in the model overall. There was also indicating strong evidence that the PH assumption may be violated in the model for the covariate `Age`. At the same time, no such violations were detected for `DiseaseLevel` or `arm`. Therefore, global significance was likely due to the time-varying effect of `Age`.

Figure 1.3 shows the Schoenfeld Residual plot for the `Age` covariate from the fitted Cox PH model. `Age` has been proven to be significant already (from Table 1.4 but this plot gives a further visualisation of and tests the PH assumption by assessing whether residuals vary systematically over time.

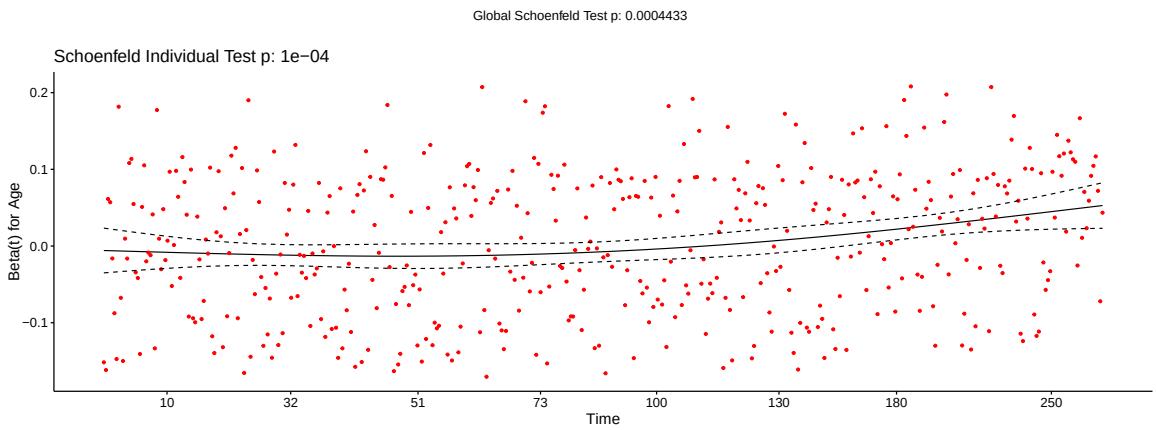

Figure 1.3: Schoenfeld Residual plot for `Age`

This plot shows a clear upward trend in the estimated coefficient over time, suggesting that the effect of `Age` on the hazard is not constant. The curve departs from zero and increases throughout the follow-up period, with pointwise confidence bands diverging as well. This visual pattern indicates a potential violation of the PH assumption, agreeing with the assessment from Table 1.4.

Therefore, to address this violation, the continuous covariate `Age` was removed from the model, and the binary variable `Age70` was introduced. This variable reflects the clinicians' specific interest in patients over 70 and was incorporated using `strata(Age70)` in the Cox Regression model. Stratification allows the baseline hazard to vary between age groups without estimating a separate coefficient, thereby relaxing the PH assumption for `Age70` while preserving valid inference for the remaining covariates. Here were the Global and Individual Schoenfeld Residual Tests for this model:

Table 1.5: New Global and Individual Schoenfeld Residual Tests

Covariate	χ^2	**df**	**p-value**
DiseaseLevel	2.935	1	0.087
arm	0.242	1	0.623
Global	3.192	2	0.203

The p-value of the Global Schoenfeld Residual Test is above our significance level of 0.05, suggesting that the PH assumption is reasonably satisfied in the stratified model. Now, we could examine the estimated hazard ratio from the treatment effect in this model.

For this new Cox PH model, the estimated hazard ratio for the treatment group relative to the control group was 0.646, with a 95% confidence interval of [0.537, 0.777]. This indicated that the treatment is associated with a 35.4% reduction in the hazard of death.

Ultimately, the treatment effect was highly statistically significant ($p = 3.59 \times 10^{-6}$), suggesting there is sufficient evidence to say there is a survival benefit for patients receiving the treatment, after accounting for stratification by age group.

Chapter 2 Trial Considerations

2.1 Sample Size

Simulation-Based Method

The simulation approach has multiple advantages over the prior formula-based methods looked at. The first of these is transparency: if the data-generating mechanism is made clear, then the assumptions behind the trial are also clear, and anyone can replicate the simulation. Reproducibility is a big issue in clinical trials. Second, the simulation approach brings flexibility because it can simulate arbitrarily complex or unusual trials. In contrast, the prior formula-based methods are limited to very specific circumstances. Simulation-based methods also enable the exploration of the implications of the decisions we'll have to make, such as how the allocation is made. Finally, it enables trial practice. This process requires performing the planned analysis at the planning stage, albeit likely in a simplified form, thus raising any potential issues early enough to adapt the plan of analysis.

The first and third advantages could also be true, though not automatically, of a well-planned trial that used conventional sample size formulae. Still, the second is an advantage unique to simulation. Therefore, simulation was used to generate the number of participants for this trial on survival data.

Evaluation Method

There are fundamentally 3 ways to analyse survival data: using parametric, semi-parametric or non-parametric tests. The log-rank test is a non-parametric method for comparing overall survival distributions between groups. It is a simple and commonly used method which involves using the observed and expected number of events to derive a test statistic. However, this method is non-parametric, and 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. Therefore, this should be considered when analysing results. The Likelihood Ratio Test (LRT) is a parametric test, and it can be adjusted for the exponential distribution, as an example, solely in the context of comparing overall group parameters, specifically the rate parameters λ_C and λ_T for control and treatment groups, respectively. Although parametric regression models can, in general, be extended to incorporate baseline covariates, the Cox Regression model was used because it can handle them more easily.

2.2 Allocation

With the data complete and bins assigned, the next step was arm allocation. The simplest method for this is simple random allocation, where each participant is independently assigned to a trial arm with equal probability. This ensures a truly random sample and avoids the need for centralised randomisation. However, it can lead to unequal group sizes in small trials, reducing statistical power. It also relies on the seed used in the sample function, which can lead to unpredictable imbalances.

Random Permuted Blocks help maintain balance but may introduce predictability if the block size is fixed and known, creating a risk of selection bias. Using random block lengths mitigates this by making allocations less predictable, though patterns can still emerge in some cases.

Biased Coin Design dynamically adjusts allocation probabilities to favour the under-represented group, improving balance but increasing the risk of predictability if the current imbalance is known. Urn models take a similar adaptive approach, heavily correcting early imbalances and slowly converging toward random allocation as the trial progresses. However, this strong correction early on can lead to highly variable allocation probabilities, which may be unpredictable and difficult to justify statistically.

The final allocation methods considered were stratifying and minimisation. These are both more complex than the other methods proposed. Stratified sampling divides each factor into levels and assigns treatment separately for each combination of factor levels (strata), often using permuted block designs. Strata are determined by multiplying the number of levels in each factor:

$$\begin{aligned}\text{Total Strata} &= (\text{Levels of } \texttt{AgeGroup}) \times (\text{Levels of } \texttt{DiseaseLevel}) \\ &= 2 \times 3 = 6 \text{ strata.}\end{aligned}$$

Since there are 640 participants (320 per arm), and stratification is performed across 6 strata, this yields approximately 53 participants per stratum, or around 26 to 27 per treatment arm within each stratum. While this sample size is mostly sufficient to avoid major imbalances, simple stratified randomisation alone could still lead to chance uneven distributions, particularly in smaller strata. Therefore, minimisation was used to mitigate this risk and ensure a better balance of outcome-relevant covariates across treatment arms.

2.3 Extra Idea to Consider

Switching the primary outcome in a survival trial from time-to-event data to a binary outcome, such as whether a patient is alive at some predetermined point (for example, after exactly six months), would have significant implications for the trial's design, analysis, and practicalities.

The most notable design implication is the loss of timing information. Time-to-event analysis not only considers the occurrence of an event but also the time of occurrence by making use of data on study participants with censoring, {i.e., those lost to follow-up or those alive when the study ends (in this trial, the study ended after 9 months)}. On the other hand, a binary outcome at one time point ignores such distinctions, grouping all events which happen before the six-month point as the same, irrespective of whether they occur early or late in the follow-up period. Loss of timing information also generally reduces statistical power because the timing and sequence of events between participants, which would typically inform the estimation of quantities like the event rate in an exponential survival model, are not fully used in the estimation of binary outcomes.

To compensate for this loss in power, it may also be necessary to increase the sample size because the trial becomes less sensitive to detecting smaller treatment effects that can still have significant implications, such as increased longevity or improved quality of life. This consideration is especially applicable when the treatment effect causes late mortality rather than complete prevention during the interval of interest. For such a case, the application of a binary endpoint may fail to capture benefits that would be seen with a time-to-event analysis. Therefore, this makes the selection of the six-month cut-point specifically important because when a large number of events are clearly before or after this specific point in time, the binary result cannot represent the genuine treatment effect. This

necessitates relying on prior clinical experience to select an appropriate time point for assessment, introducing the risk that the chosen timing may not adequately capture the treatment effect.

How the actual sample size is calculated would also change because binary outcome designs rely on expected survival proportions at the chosen time point, rather than hazard rates or event distributions over time, as in time-to-event analyses. This shift makes binary designs more dependent on accurate prior estimates of fixed-time survival rates, which can be harder to obtain, particularly when prior data are limited or survival varies substantially over time. Hence, this may reduce planning flexibility.

Also, a fixed-time-point design requires all participants to be followed for the same fixed duration. By contrast, event-driven time-to-event trials can terminate when a pre-specified number of events have occurred, improving trial efficiency and making better use of trial resources, especially when event rates are uncertain. Moreover, there are ethical implications. Time-to-event methods allow for intermediate analyses based on accumulating event data, which can be a basis for stopping early in instances of benefit or harm (say, for example, the treatment has dire effects on those taking it). In contrast, binary outcome designs lack this flexibility, since survival status is only assessed at the end of the follow-up period, possibly delaying pivotal decisions about treatment efficacy or safety.

Analytically, binary outcomes redirect attention away from survival techniques, which have greater flexibility in modelling {including non-parametric (such as KM estimation), semi-parametric (such as Cox Regression), and parametric methods (such as the LRT)}, toward methods based on proportions, such as chi-squared tests or logistic regression if covariate adjustment is required. These methods based on proportions provide summary measures such as risk differences or odds ratios that reflect treatment effects at a specific time point, rather than over the full duration of follow-up. As such, binary analysis can mask delayed or time-varying treatment effects.

Additionally, binary outcomes prevent the estimation of the full survival function, which is often a key output in survival trials. This limits the ability to report useful metrics such as median survival time or survival probabilities across different time points.

Analysis of censored data differs a lot as well. For time-to-event analysis, censored subjects carry some information. For binary analysis, patients lost to follow-up before the pre-specified time point can be excluded altogether, and bias may arise, particularly if those patients were more or less likely to experience the event compared to those who remained in the study. This also obscures the timing of treatment benefit, making it difficult to determine how early or late a treatment begins to have an effect. In addition, because censoring is not explicitly modelled in binary analysis, informative censoring (for example, a patient dropping out because their condition worsens and they choose palliative care) becomes harder to detect and adjust for.

Practically, a binary outcome may seem easier as it needs only one outcome measure for each participant. But this implies that a participant's survival status at the time point of interest must be correctly ascertained, which may be logistically demanding. The method also lacks flexibility because deaths occurring shortly after the cut-off are excluded from the primary analysis, regardless of their clinical significance, whereas deaths occurring shortly prior to the cut-off are included completely. This can mislead the interpretation, particularly if results are clustered near the threshold.

However, a practical benefit from binary outcomes is that results derived are more simply explained to non-statistics literate audiences. For instance, explaining that "80% of patients were alive at six months in the treatment group compared to 70% in the control group" is easy to understand for a non-technical audience. Nonetheless, this ease is at the cost of not always being in a position to convey a complete image of how survival varies with time.