Survival Modelling in Clinical Trial Forecasting: Integrating Kaplan-Meier and Ridge-Cox Proportional Hazards Approaches within the TrialIQ Framework

Author: Francis Osei
Email: Francis@bayezian.com

1. Introduction to Survival Modelling

When clinical trials are designed, one of the most pressing questions is not simply whether they will succeed, but *when* they might fail. Predicting this timing is crucial, especially given the high cost of early discontinuation and the lengthy timelines involved. Traditional approaches that categorise trials as either completed or failed can be informative, but they often miss the subtleties that come with time-sensitive outcomes. Not all trials reach an endpoint within the observation period, and some are still ongoing when data is analysed. This is where survival analysis becomes particularly useful.

Survival models are designed to handle precisely this kind of data. They take into account both the occurrence of an event and the time it takes to occur, while also accounting for censoring instances where the outcome is not yet observed. In the context of clinical trials, survival analysis allows us to ask not just whether a trial failed, but how long it lasted before it did. This gives us a richer and more realistic picture of trial dynamics.

In this project, we make use of two complementary survival modelling approaches. The first is the Kaplan-Meier estimator, which provides a straightforward, visual summary of how survival probabilities vary across groups. The second is the Cox proportional hazards model, which allows us to examine the effect of multiple protocol-level variables on trial duration. While the Kaplan-Meier curves help us observe general patterns, such as whether trials from industry sponsors tend to survive longer, the Cox model gives us a more detailed and statistically grounded understanding of which features are most associated with early trial termination.

Together, these methods form the backbone of our analytical framework. They are not only statistically appropriate for the nature of the data, but also deeply relevant to the practical concerns of trial designers, sponsors, and regulators who are looking to reduce risk and improve the likelihood of success from the very start.

2. Kaplan-Meier Estimation: Non-parametric Survival Insights

To begin our exploration of trial longevity, we turned to the Kaplan-Meier estimator, a classic yet remarkably intuitive method for summarising survival patterns over time. Unlike more complex models, it does not assume anything about the shape of the hazard function. This makes it a fitting choice for initial comparisons, particularly when we wish to assess how survival differs across key trial characteristics.

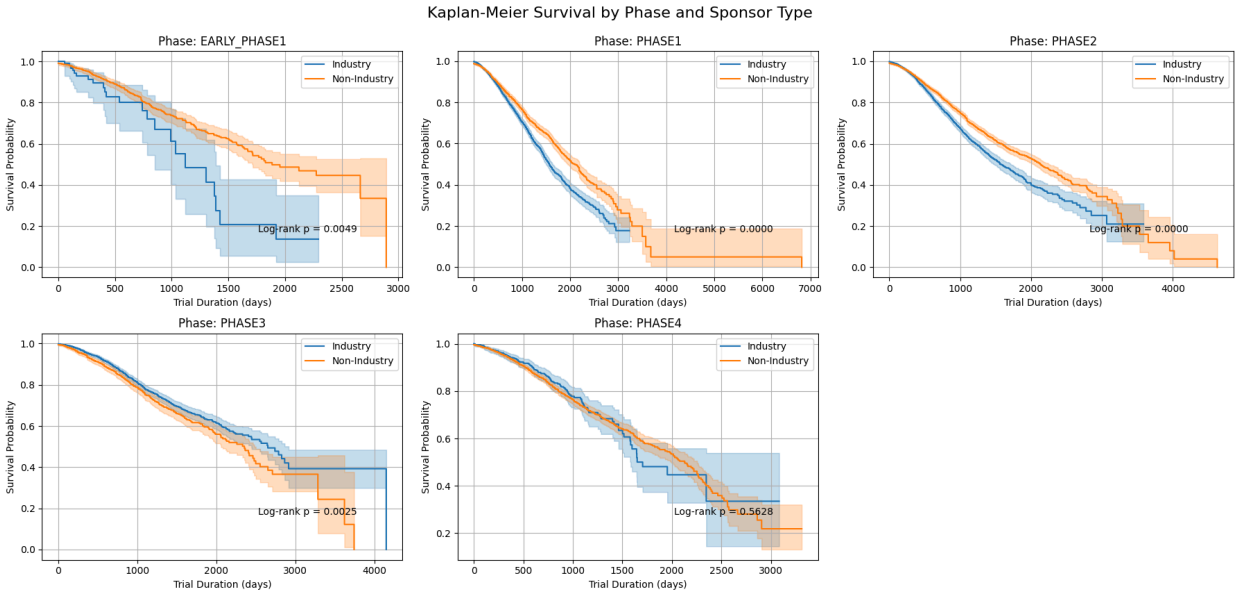
In this case, we focused on two such features: the phase of the trial and the type of sponsor involved. The results are shown in the figure below, which plots estimated survival probabilities for each trial phase, further broken down by whether the sponsor was from industry or not.

What becomes immediately clear is that early-phase trials face greater uncertainty. For instance, in Early Phase I, industry-sponsored trials demonstrate consistently higher survival probabilities compared to those run by non-industry groups. This difference is both visual and statistically significant, with a log-rank p-value of 0.0049, indicating that the two curves are unlikely to have arisen by chance alone.

A similar trend continues into Phase I and Phase II, where industry-led studies appear to progress for longer periods before failure or discontinuation. The separation between the curves here is even more pronounced, and the statistical evidence is correspondingly stronger, with p-values near zero. These findings are consistent with expectations. Industry sponsors often have more resources, clearer regulatory pathways, and the benefit of prior experience, all of which can contribute to better trial endurance.

By contrast, as we move towards Phase III and Phase IV, the distinction between sponsor types becomes less marked. In Phase IV, in particular, the two survival curves are nearly indistinguishable, and the log-rank test confirms that the difference is not statistically meaningful. One interpretation is that by the time trials reach these advanced stages, they have passed through numerous filters of feasibility and funding, regardless of who is sponsoring them.

The Kaplan-Meier analysis provides a valuable first glimpse into how certain design attributes may shape trial outcomes. However, it only considers one or two variables at a time. To understand how these features interact and which ones matter most when considered together, we now turn to a multivariable approach, the Cox proportional hazards model.



3. Cox Proportional Hazards Model: Development and Tuning

While the Kaplan-Meier method offered a helpful snapshot of survival trends across sponsor types and trial phases, it does not allow us to evaluate several features at once. In reality, trial duration is influenced by a combination of design decisions, eligibility complexity, regulatory pathways, and operational factors. To capture this interplay, we applied the Cox proportional hazards model, a semi-parametric approach that estimates the effect of covariates on the hazard, or risk, of failure over time.

The modelling process began with structured data from ClinicalTrials.gov, consisting of over 32,000 trials. We carefully cleaned and prepared the dataset, removing constant variables, handling missing values, and grouping rare categories to avoid sparsity. Continuous variables, such as site count or age thresholds, were left in their natural scale, as they had already been normalised in earlier preprocessing.

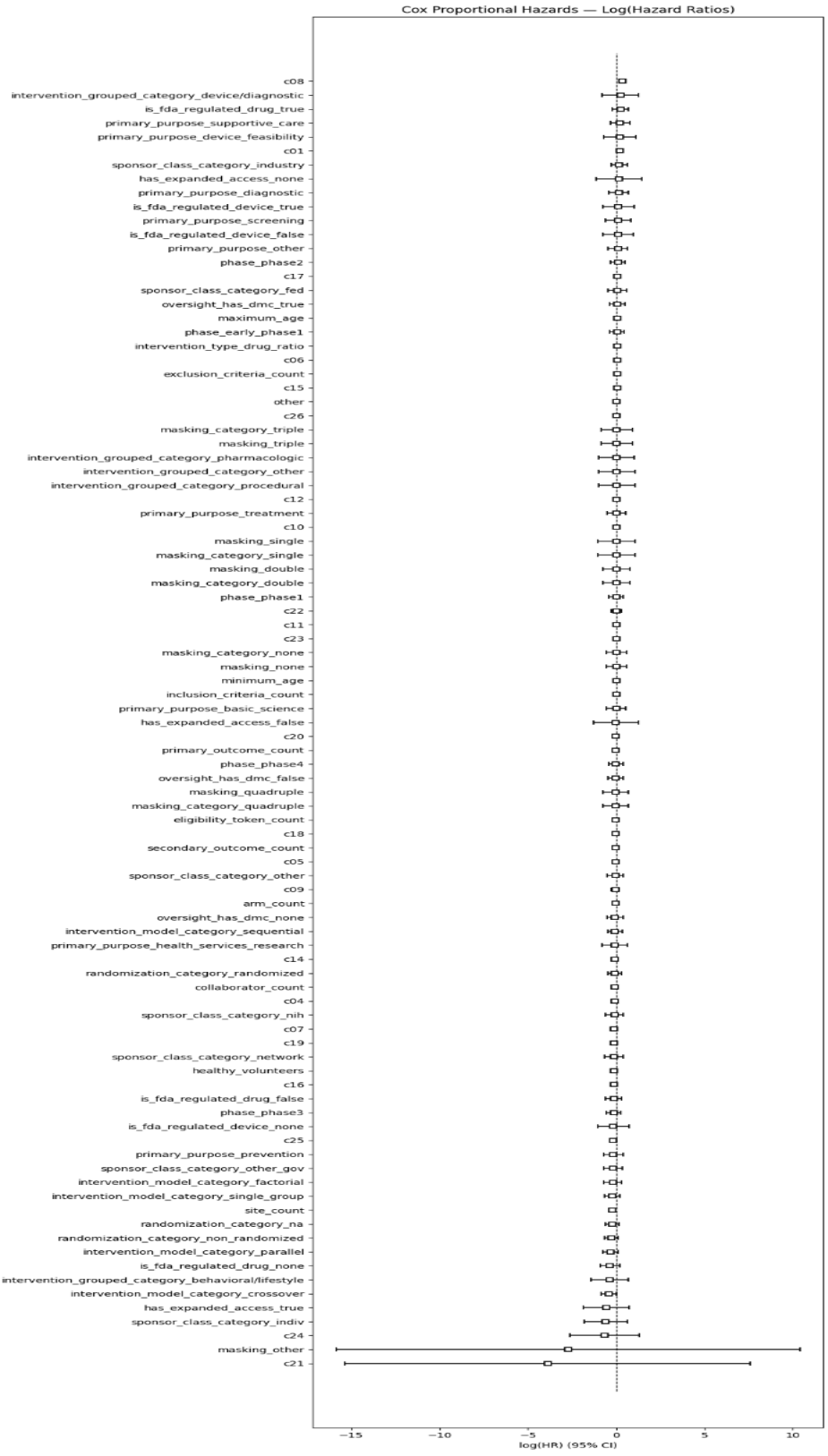
A crucial step in fitting a Cox model is selecting an appropriate level of regularisation. We performed five-fold cross-validation over a range of penaliser values to guard against overfitting. The optimal penaliser yielded a validation C-index of approximately 0.63, indicating a reasonable ability to rank trials by their risk of early failure.

Once trained, the model was used to estimate log hazard ratios for each covariate, as shown in the forest plot below.

Each row in the plot represents a trial design variable, with the square marking the estimated log hazard ratio and the horizontal line indicating the 95% confidence interval. Variables to the right of the zero line are associated with a higher risk of failure (shorter survival), while those to the left suggest a protective effect.

For example, trials involving device/diagnostic interventions, or those classed under C08 (likely therapeutic area-specific), appear to carry a markedly elevated hazard. Conversely, certain masking strategies, such as double or triple blinding, are associated with improved survival, which may reflect stronger methodological rigour. Other interesting findings include a reduced hazard among trials with federal sponsors and a greater risk where regulatory oversight is absent or unclear.

Taken together, these results provide a more nuanced picture of what drives trial failure, going well beyond phase and sponsor type alone. In the following section, we examine how well this model performs and how its risk estimates distribute across different subgroups.


Cox Proportional Hazards — Log(Hazard Ratios)
c08
intervention_grouped_category_device/diagnostic
is_fda_regulated_drug_true
primary_purpose_supportive_care
primary_purpose_device_feasibility
c01
sponsor_class_category_industry
has_expanded_access_none
primary_purpose_diagnostic
is_fda_regulated_device_true
primary_purpose_screening
is_fda_regulated_device_false
primary_purpose_other
phase_phase2
c17
sponsor_class_category_fed
oversight_has_dmc_true
maximum_age
phase_early_phase1
intervention_type_drug_ratio
c06
exclusion_criteria_count
c15
other
c26
masking_category_triple
masking_triple
intervention_grouped_category_pharmacologic
intervention_grouped_category_other
intervention_grouped_category_procedural
c12
primary_purpose_treatment
c10
masking_single
masking_category_single
masking_double
masking_category_double
phase_phase1
c22
c11
c23
masking_category_none
masking_none
minimum_age
inclusion_criteria_count
primary_purpose_basic_science
has_expanded_access_false
c20
primary_outcome_count
phase_phase4
oversight_has_dmc_false
masking_quadruple
masking_category_quadruple
eligibility_token_count
c18
secondary_outcome_count
c05
sponsor_class_category_other
c09
arm_count
oversight_has_dmc_none
intervention_model_category_sequential
primary_purpose_health_services_research
c14
randomization_category_randomized
collaborator_count
c04
sponsor_class_category_nih
c07
c19
sponsor_class_category_network
healthy_volunteers
c16
is_fda_regulated_drug_false
phase_phase3
is_fda_regulated_device_none
c25
primary_purpose_prevention
sponsor_class_category_other_gov
intervention_model_category_factorial
intervention_model_category_single_group
site_count
randomization_category_na
randomization_category_non_randomized
intervention_model_category_parallel
is_fda_regulated_drug_none
intervention_grouped_category_behavioral/lifestyle
intervention_model_category_crossover
has_expanded_access_true
sponsor_class_category_indiv
c24
masking_other
c21
−15
−10
−5
0
5
10
log(HR) (95% CI)

4. Model Evaluation and Performance

Once the Cox model was trained, we evaluated its ability to distinguish between trials likely to fail early and those with greater longevity. The most widely used measure for this is the concordance index, or C-index, which reflects how well the model ranks pairs of trials by their relative risk. A score of 1 would indicate perfect discrimination, while 0.5 suggests performance no better than chance.

On the training data, the model achieved a C-index of 0.6256, with a partial log-likelihood of –67,995.99 and an AIC of 136,177.97. These figures are in line with expectations for real-world survival models based on registry data, where outcomes are often influenced by unmeasured or post-enrolment variables beyond the scope of protocol design. On the held-out test set, performance remained consistent, with a global C-index of 0.6335 across over 8,000s.

Beyond overall discrimination, it was essential to assess how fairly the model performed across different subgroups. We explored this by calculating C-indices within categories such as sponsor class, regulatory involvement, trial design features, and continuous variables like site count and exclusion criteria complexity, grouped into low, mid, and high bins.

The results revealed subtle yet meaningful disparities. For instance, trials sponsored by industry performed slightly better under the model, with a C-index of 0.6339 compared to 0.6288 for non-industry trials. Similarly, trials involving FDA-regulated drugs had slightly lower predictive performance (0.5999) than those that were not (0.6266). Some design types, such as parallel group trials, were associated with markedly higher C-indices (0.6569) than crossover models (0.6075), suggesting the model handled the former more reliably.

Among continuous variables, the site count and secondary outcome count stood out. Trials with a higher number of sites or additional outcomes tended to be modelled with better accuracy. This may be because larger, more complex trials offer richer design signals that the model can pick up on.

These fairness checks are not simply academic exercises. They help ensure that the model does not unduly favour or penalise specific types of trials and can be trusted as part of a decision-support system. While no model is perfectly balanced, understanding its blind spots allows for more responsible use and, where necessary, adjustment or stratification in downstream applications.

In the next section, we shift from global performance to a more individual-level perspective, examining how partial risk scores distribute across key predictors and what they reveal about the model's internal logic.

5. Interpretation of Key Risk Drivers

Having established the model's overall performance and subgroup reliability, the next step was to explore *why* it makes the predictions it does. The Cox model offers a clear advantage here, as its coefficients correspond directly to the log of the hazard ratio. This allows us to interpret the relationship between each covariate and the likelihood of early trial failure with some clarity.

To begin with, we examined the magnitude and direction of the model's coefficients. These were visualised in the forest plot shown earlier, where variables associated with increased failure risk appear on the right-hand side, and those that seemed to offer a protective effect appear on the left. Several findings stood out immediately.

Trials with no form of masking, or with only single-blinding, tended to show higher hazards. This may reflect weaker trial design or insufficient control for bias, both of which can lead to early discontinuation. In contrast, double-blind and triple-blind studies were associated with longer survival, possibly because they indicate greater methodological rigour or more established operational planning.

Sponsor type also remained a key driver. Industry-sponsored trials were more likely to survive longer, even after accounting for other variables. This reinforces earlier Kaplan-Meier results and may point to better resource planning or more thorough feasibility assessments at the design stage.

Among the categorical predictors, the presence of crossover intervention models and the absence of regulatory oversight appeared to increase the hazard, while parallel-group designs were more stable. Notably, trials involving C08-coded therapeutic areas (which often include respiratory and immunological diseases) showed a higher risk, a signal that warrants further exploration within specific medical domains.

To complement this, we selected a subset of the most influential predictors, both categorical and continuous, and visualised how the model's partial hazard scores were distributed across their values. The resulting figure is presented below.

This plot provides a more granular look at how individual predictors shape risk. For binary variables such as sponsor type or intervention model, we see a clear separation in the distributions of partial hazard. Trials lacking FDA regulation or featuring expanded access programmes tend to cluster at higher risk levels. For continuous variables like eligibility token count or maximum age, we observe a more gradual shift, with higher values often associated with broader and more ambitious trials and, as it turns out, more risk.

These patterns not only validate the model's internal logic but also offer immediate, actionable insight. Trial designers might consider revisiting eligibility criteria, reassessing intervention structure, or engaging regulatory bodies more thoroughly during early planning stages.

In the next section, we take a step back and compare these findings with our earlier non-parametric analysis. This helps us understand where the two approaches align and where they offer different perspectives on trial viability.

6. Partial Risk Stratification

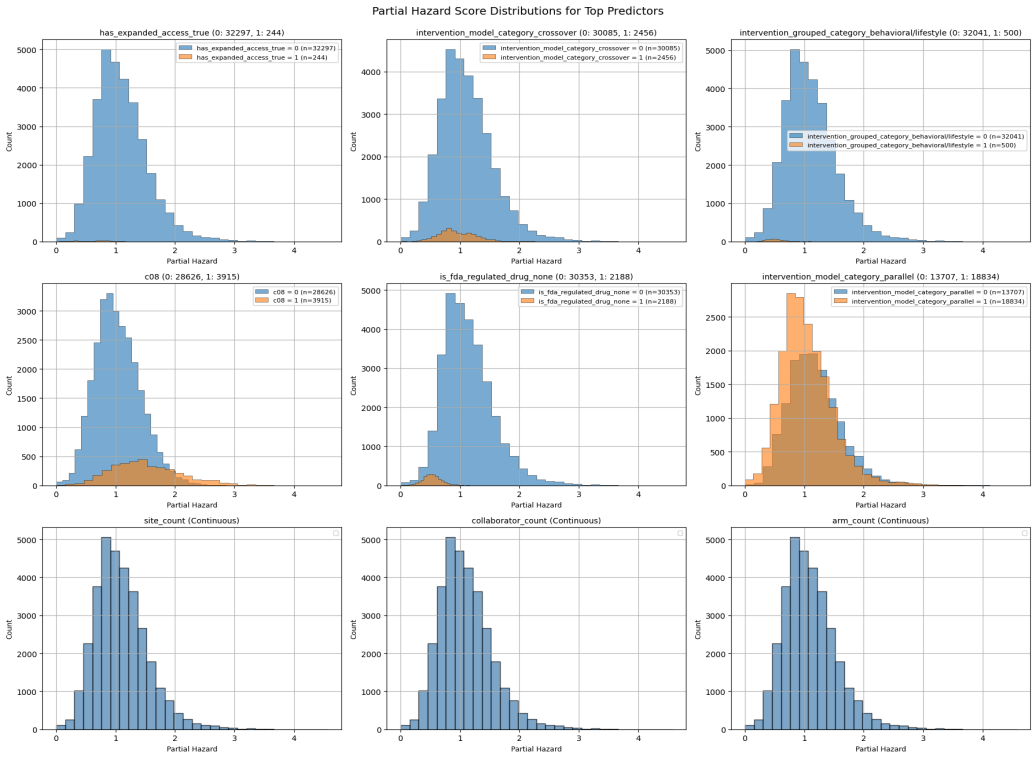
While model coefficients tell us the average effect of each variable, they do not reveal how risk is distributed across individual trials. To fill this gap, we examined the model's *partial hazard scores* and the underlying risk estimates for each trial, given its design features, and stratified them by key predictors. This helped us move beyond abstract coefficients to something more tangible: how risk plays out across real examples in the dataset.

The figure below presents a series of histograms, each showing the distribution of partial hazard scores for selected variables. These were chosen based on their overall contribution to the model and their practical relevance to trial design.

Take a sponsor class, for example. Industry-led trials consistently occupy the lower end of the risk spectrum, with a relatively narrow spread. By contrast, non-industry trials show a broader and right-shifted distribution, suggesting greater variability and a larger proportion at high risk. This reflects not only resource differences but perhaps also inconsistent trial planning across academic and non-profit institutions.

A similar pattern emerges when comparing trials with and without FDA regulation. Those lacking regulatory involvement tend to cluster at higher risk, possibly due to weaker oversight, less stringent protocol review, or lower barriers to entry in the design process.

What makes these distributions particularly insightful is that they reveal overlap. Not every non-industry trial is high-risk, and not every industry trial is guaranteed to survive. There are pockets of high-performing designs in every category, which reinforces the importance of looking beyond labels. Instead of assuming risk based on the trial phase or the sponsor alone, the partial hazard offers a more individualised measure that trial designers, funders, and regulators can use to make better-informed decisions. In short, this stratified analysis shows that survival risk is not just a function of *what kind of trial* is being run, but *how it is designed*. In the next section, we turn to the relationship between our two modelling approaches, Kaplan-Meier and Cox, to explore where they complement each other and where they diverge.



7. Comparative Insights: Kaplan-Meier versus Cox

Having explored both the Kaplan-Meier and Cox proportional hazards approaches in depth, it is useful to reflect on how these two methods relate to one another. Although both belong to the broader field of survival analysis, they serve different purposes and offer complementary strengths.

The Kaplan-Meier estimator is fundamentally descriptive. It allows us to visualise survival probabilities over time across different groups without making strong assumptions. This made it especially valuable early in the analysis, where we observed clear differences in trial duration between industry and non-industry sponsors, particularly in early phases. These findings provided an initial hypothesis: that sponsor type and trial phase meaningfully affect survival and perhaps interact in complex ways.

The Cox model then enabled us to test and refine these ideas in a multivariable setting. It confirmed that sponsor type was indeed associated with survival, but also revealed that other design features, such as masking, intervention model, regulatory involvement, and complexity of eligibility criteria, were often more influential. In doing so, the Cox model helped disentangle correlations that the Kaplan-Meier curves could not address on their own.

Importantly, there were also areas where the two approaches diverged. For instance, the Kaplan-Meier plots showed diminishing differences between sponsor types in Phase IV trials. The Cox model, however, suggested that even within late-phase studies, certain regulatory and operational features continued to drive risk, although the sponsor label alone became less predictive. This suggests that while sponsor class is a useful proxy early on, it loses some of its explanatory power once a trial has already progressed past major feasibility hurdles.

In practical terms, this comparison underscores why both models matter. Kaplan-Meier curves are ideal for highlighting broad trends and engaging stakeholders unfamiliar with statistical modelling. The Cox model, meanwhile, provides the nuance and rigour needed to make design-level decisions based on individual trial characteristics. When used together, they form a robust and flexible analytical toolkit — one that can inform both exploratory analysis and more targeted design optimisation.

In the following section, we consider the limitations of these approaches, including the assumptions they rely upon and the types of bias they may introduce. This is essential context for understanding how far we can take the insights generated here and where caution is warranted.

8. Limitations and Model Assumptions

As with any modelling exercise, the insights drawn from this study must be considered alongside the assumptions that underpin them. While both the Kaplan-Meier and Cox models are widely used and statistically robust, they are not without limitations, some inherent to the methods themselves, others related to the data at hand.

The Cox model, in particular, relies on the proportional hazards assumption. This means it assumes that the ratio of hazards between any two trials is constant over time. In practice, this is not always realistic.

For instance, certain design factors may exert their influence early on but become less relevant as the trial progresses. Although no major violations were detected during model fitting, the assumption remains a simplification and may not hold across all subgroups or trial phases.

Another important consideration is censoring. Our models treated trials without a recorded failure as right-censored, that is, still ongoing or yet to reach an outcome. This is a sensible and standard approach, but it assumes that censoring is non-informative. If trials are more likely to be censored because, for example, they were paused for funding reasons or reporting delays, then the risk of bias increases. Unfortunately, such subtleties are rarely visible in registry data and are difficult to adjust for statistically.

The data itself also imposes some constraints. Although ClinicalTrials.gov provides a rich source of protocol-level features, not all variables are complete or consistent. Some trial descriptors are missing altogether, while others are prone to ambiguity or inconsistent formatting. We addressed this through systematic cleaning and rare-category grouping, but these steps inevitably introduce a degree of smoothing and abstraction.

Finally, it is worth noting that while our model performed reasonably well in ranking trial risk, its predictive strength remains moderate. A C-index of around 0.63, while meaningful, implies that a substantial portion of variance remains unexplained. This is to be expected, as many factors influencing trial success, such as site-level execution, real-time patient enrolment challenges, or sponsor behaviour after launch, lie beyond the scope of what can be captured from pre-enrolment metadata.

In short, the models presented here are best viewed as decision-support tools rather than definitive outcome predictors. They offer a structured way to surface design-related risks and prioritise further review, not to replace judgement or real-world feasibility checks. In the next section, we explore precisely how these outputs might be applied in a real-world context to support more effective clinical trial design.

9. Implications for Clinical Trial Design Optimisation

One of the central aims of this work was not simply to build statistical models, but to use them in the service of better decisions. The real value lies in how these findings can inform clinical trial design, particularly in the early stages before recruitment begins and before avoidable risks are locked into the protocol.

The outputs from both the Kaplan-Meier and Cox analyses offer several avenues for practical use. For example, if a proposed trial mirrors the characteristics of those that historically failed early, such as limited masking, unclear regulatory status, or a high number of exclusion criteria, then these signals could be flagged during the design review process. In this way, the model becomes an evidence-based second opinion, nudging sponsors and design teams to revisit certain decisions before the trial is launched.

Partial hazard scores, in particular, offer a way to stratify risk at the individual trial level. Rather than relying solely on phase or sponsor type, designers can obtain a more precise estimate of where their trial sits compared to others. This is especially helpful for academic or investigator-led trials where institutional knowledge may be less extensive and design resources more constrained.

Moreover, the model's fairness evaluations serve an additional purpose. They help ensure that trials in underrepresented categories, such as those with behavioural interventions or expanded access components, are not unfairly penalised or overlooked by overly general models. Understanding where prediction is weaker allows teams to apply judgment more selectively and avoid over-reliance on algorithmic output.

Taken together, these tools do not replace domain expertise or operational planning, but they do provide an evidence-led foundation for asking better questions. Could this trial benefit from stronger blinding? Are the eligibility criteria more complex than necessary? Has the design been adequately stress tested against known risk factors?

By surfacing these considerations early, the modelling framework supports a more proactive and reflective approach to trial planning. This aligns well with emerging trends in regulatory science, feasibility assessment, and adaptive protocol development.

In the final section, we bring together these strands to reflect on the broader potential of this approach and the next steps for future work.

10. Conclusion

This study set out with a simple but ambitious aim: to understand why clinical trials fail, and whether it is possible to anticipate that failure before a single patient is enrolled. By combining non-parametric survival curves with a multivariable Cox model, we have built a framework that not only highlights broad risk patterns but also allows for more detailed insight into the design choices that shape trial success.

The Kaplan-Meier analysis provided a clear and accessible view of survival probabilities across trial phases and sponsor types. It confirmed widely held beliefs about the relative strength of industry sponsorship and the vulnerability of early-phase studies. Yet it also showed the limits of this view, revealing that survival cannot be explained by a single feature alone.

The Cox model added much-needed depth. It quantified the impact of specific design elements, from masking and regulatory involvement to trial complexity and intervention structure. By generating partial risk scores and identifying key drivers, the model moves beyond general trends to offer guidance at the level of individual trials. These insights are not just academically interesting. They are directly applicable to protocol review, feasibility assessment, and early risk flagging.

Importantly, the study has also been careful to acknowledge what the models cannot do. They do not account for everything. They do not predict outcomes with certainty. And they should never be used in isolation. But as part of a wider decision-making process, they have real value.

In sum, this has been more than an exercise in modelling. It has been a demonstration of how data, thoughtfully applied, can support better judgment and more resilient trial design. And that, ultimately, is the most promising finding of all.