# Random Survival Forests for Clinical Trial Risk Prediction: A Flexible, Non-linear Approach to Time-to-Failure Modelling

Author: Francis Osei
Email: Francis@bayezian.com

# 1. Introduction and Motivation

Predicting whether a clinical trial will succeed or fail is a complex task, not least because the factors influencing trial outcomes are both numerous and highly interdependent. Traditional survival models, such as the Cox proportional hazards model, offer a strong statistical foundation but rely on linear assumptions and may struggle to capture subtle interactions among design variables. Parametric models, while more structured, require strict distributional assumptions that rarely hold in real-world data.

To address these challenges, we turned to Random Survival Forests, a non-parametric ensemble approach that brings the flexibility of decision trees into the domain of survival analysis. Unlike its linear counterparts, the RSF model can naturally model non-linear relationships and interactions without the need to specify them in advance. This makes it especially well suited for protocol-level metadata, where variable formats vary widely and relationships are often unknown or non-intuitive.

We aimed to see whether RSF could offer meaningful improvements in predictive accuracy, both in absolute terms and across key trial subgroups. At the same time, we were interested in whether the model could offer a useful lens into the underlying drivers of trial failure, particularly by estimating variable importance and generating individual-level risk scores.

In the sections that follow, we describe how the model was built and evaluated, how it performed on unseen data, and what insights it provided about the structure and dynamics of clinical trial risk.

# 2. Model Architecture and Training

The Random Survival Forest model used in this study was based on an ensemble of decision trees, each trained on bootstrapped samples of the data, with survival curves estimated through node-level aggregation. This method allowed the model to capture complex interactions and non-linear relationships between trial design features and survival time, without assuming any specific hazard structure or proportionality constraint.

To ensure the model was appropriately tuned, we conducted a grid search using cross-validation. A range of hyperparameter combinations was tested, focusing on the number of features considered at each split, the minimum number of samples required to create a split, and the minimum number of observations per leaf. The evaluation metric throughout was the concordance index, chosen for its ability to measure risk discrimination in censored survival settings.

Across the parameter grid, several configurations yielded very similar performance, with average concordance values clustering around the mid-0.66 range. Ultimately, the model with the best

cross-validated performance used half the available features at each split, required at least twenty samples in a leaf, and allowed splits with as few as five samples. This combination offered a balance between stability and granularity, helping the model generalise well without overfitting.

The final model was then trained on the full dataset using these selected parameters. Internal evaluation showed a concordance index above 0.81, indicating strong internal discrimination. Importantly, no feature scaling was required, and categorical variables were handled via one-hot encoding after grouping rare levels. This straightforward pre-processing made the model relatively simple to apply across a wide range of trial records.

Having established a strong in-sample performance, we then turned to evaluate how well the model generalised to unseen data, and whether its predictive power remained consistent across key subgroups of interest.

# 3. Out-of-Sample Performance and Fairness Analysis

To assess how well the Random Survival Forest model generalised beyond the training data, we evaluated its predictions on a separate test set comprising over eight thousand clinical trials. The key performance metric was the concordance index, which reflects the model's ability to correctly rank trials by their risk of early failure while accounting for censoring.

The model achieved a concordance index of approximately 0.69 on the test set. This marks a notable improvement over previous models explored in this study, including the Cox proportional hazards and parametric survival approaches. While not perfect, this level of discrimination suggests the RSF model is capable of capturing meaningful patterns in trial design that correlate with time-to-failure.

However, overall performance is only part of the story. To ensure the model behaves reasonably across different types of trials, we also examined its fairness across subgroups. This involved computing concordance scores separately within categories such as sponsor type, regulatory status, and trial design structure.

Results varied across groups but were largely consistent. For example, industry-sponsored trials had a C-index around 0.68, while non-industry trials were slightly higher. Trials flagged as FDA-regulated showed marginally lower predictive accuracy, but still within a reasonable range. Some subgroups, such as trials involving expanded access or behavioural interventions, produced more variable results, though these often involved small sample sizes and should be interpreted with care.

When grouping trials by continuous variables such as site count, maximum age, or eligibility complexity, the model also showed robust performance. Across low, mid, and high strata, concordance values remained stable and showed no evidence of systematic bias in one direction or another.

These findings suggest that the model generalises well and does not disproportionately favour any particular type of trial. This is essential for practical use, particularly if the tool is to be trusted across different therapeutic areas, sponsor classes, and trial phases.

In the next section, we take a closer look at how the model generates risk estimates and explore which features appear most influential in shaping its predictions.

# 4. Risk Stratification and Variable Influence

Beyond overall performance metrics, one of the most valuable aspects of the Random Survival Forest model is its ability to provide insight into the relative importance of different trial features. By examining the model's risk predictions and their distribution across subgroups, we can begin to understand not just *how well* the model performs, but *why* it arrives at the predictions it does.

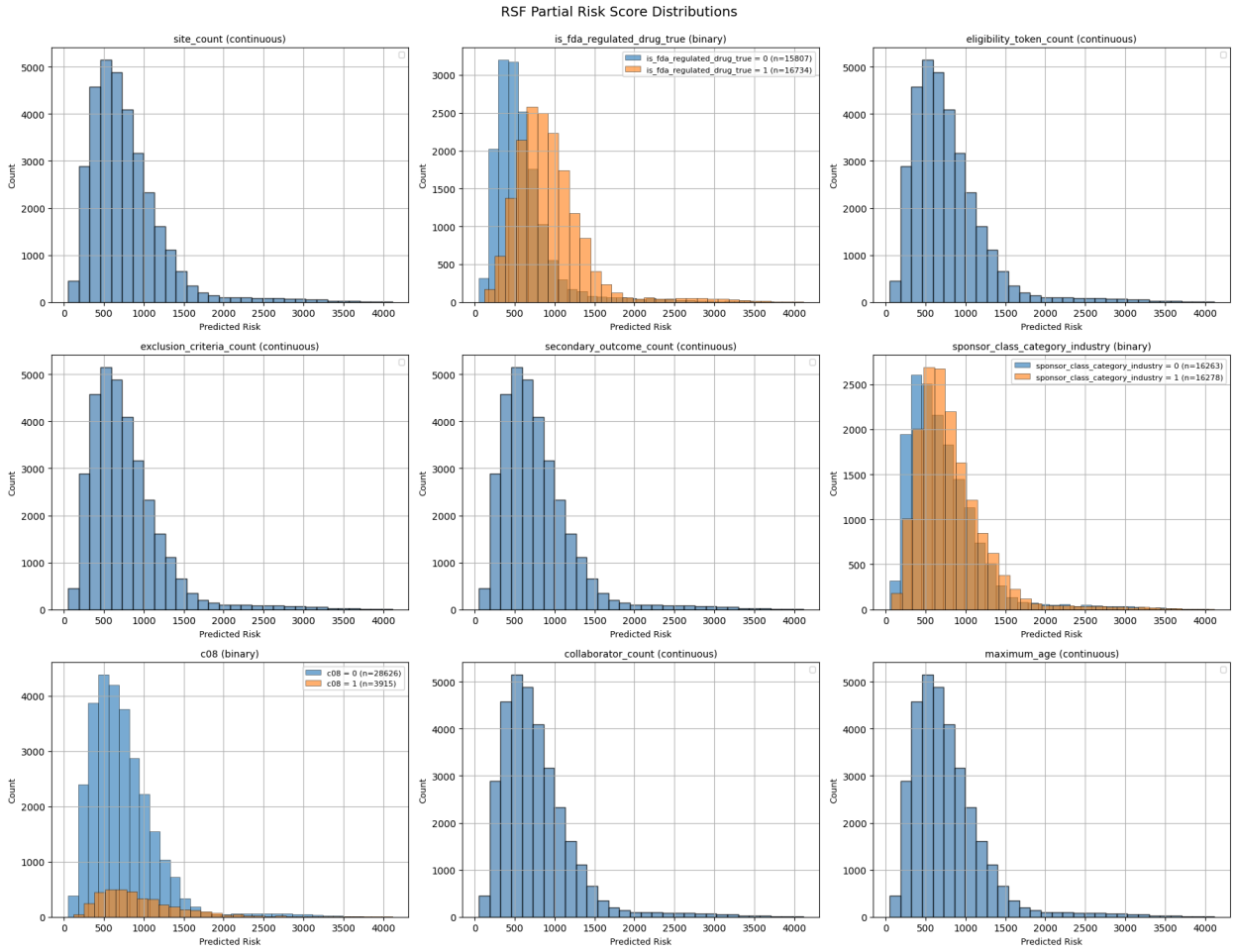
To explore this, we began by estimating partial risk scores for each trial in the training data. These scores represent the model's relative assessment of how likely each trial is to fail early, given its specific design characteristics. The figure below shows how these scores are distributed across several key predictors, including sponsor type, intervention model, and eligibility complexity.

As shown, trials sponsored by industry organisations tended to cluster towards lower predicted risk, while those with behavioural or crossover interventions displayed wider distributions with more high-risk outliers. This aligns with earlier findings from both the Cox and Kaplan-Meier models, but with greater nuance and depth. In particular, the RSF model captured more variation within each group, highlighting that not all industry trials are safe and not all academic ones are fragile.

We also applied a permutation importance procedure to identify which features had the most influence on the model's predictions. The most influential variables included sponsor classification, FDA regulation status, the number of exclusion criteria, and the type of masking used in the trial. Interestingly, these align closely with the most impactful variables identified by the Cox model, which adds a layer of reassurance that the RSF model is capturing meaningful clinical patterns rather than noise.

These outputs are not just academic. They offer practical guidance for trial designers by highlighting where design choices appear to carry the greatest risk. For example, reducing the number of exclusion criteria or considering more robust blinding may, in aggregate, lead to lower predicted risk and improve trial feasibility.

In the final section, we bring together the results from all models and reflect on the strengths and limitations of the RSF approach in the broader context of trial forecasting.




# 5. Summary of Findings and Applicability

The Random Survival Forest model demonstrated that flexible, non-linear methods can offer real advantages in predicting clinical trial failure using structured design data. Among all models explored in this project, RSF delivered the strongest test performance, with a concordance index approaching 0.69. This suggests it was better able to capture the complex, multifactorial nature of trial risk than its linear or parametric counterparts.

Crucially, the model's predictions held up across a variety of subgroups. It performed reliably across industry and non-industry sponsors, across trials with and without regulatory oversight, and a broad range of intervention types. This consistency is essential if the model is to be applied in real-world decision-making, where robustness and fairness are just as important as accuracy.

The RSF model also contributed valuable insight into which protocol features carry the most weight. Variables such as sponsor class, exclusion criteria complexity, and masking design emerged repeatedly as key drivers of survival risk. This not only supports earlier statistical findings but also helps inform practical trial design by pointing to areas where risk might be mitigated.

However, the model is not without its limitations. Like all data-driven tools, it is constrained by the quality and scope of the input data. Features not captured in the registry, such as operational readiness, geographic enrolment challenges, or funding uncertainty, will not be reflected in the predictions. As such, the RSF model should be viewed as a complement to expert judgement, not a replacement.

Within the broader framework, the RSF model offers a valuable middle ground. It is more flexible and accurate than classical approaches, but still interpretable enough to be used in risk triage, feasibility assessments, and early design reviews. With further development and integration, it could support a more evidence-informed protocol planning process, helping sponsors anticipate problems before they arise.