Exploring Parametric Survival Models for Trial Duration Forecasting: A Comparative Evaluation Using Weibull, Log-logistic, and Log-normal Approaches

Author: Francis Osei
Email: Francis@bayezian.com

1. Introduction and Rationale

In attempting to forecast the duration of clinical trials and their risk of early failure, much attention is typically given to flexible, non-parametric approaches such as the Kaplan-Meier estimator or the Cox proportional hazards model. These methods require fewer assumptions and often perform well across diverse datasets. However, there is a longstanding interest in parametric models for survival analysis, particularly because they offer fully specified survival functions and can be more efficient when their assumptions are met.

In this part of the project, we explored whether such parametric approaches could offer competitive performance in modelling time-to-failure for clinical trials, using structured metadata from ClinicalTrials.gov. Specifically, we focused on three widely used models: the Weibull, log-logistic, and log-normal accelerated failure time (AFT) models.

Parametric models, unlike the semi-parametric Cox model, assume a specific functional form for the distribution of survival times. In principle, this should allow for more precise estimation and extrapolation, particularly in cases where the hazard function changes over time. For example, the Weibull model accommodates both increasing and decreasing hazards, while the log-logistic and log-normal models allow for more complex curvature.

Our aim here was to assess whether these stronger assumptions could translate into better predictive accuracy, or at the very least, provide alternative insights into the structure of survival risk in clinical trials. As the following sections will show, the results were mixed and somewhat sobering, offering a valuable lesson in the limits of model complexity when faced with noisy or incomplete design data.

2. Model Development and Comparison

The dataset used for model development consisted of structured trial records, with time-to-event information expressed in days and an event indicator representing trial failure. Records missing either duration or event status were excluded, and trials with zero duration were filtered out to avoid distortion. After this initial processing, we trained three classes of accelerated failure time models, Weibull, log-logistic, and log-normal, using penalised likelihood estimation to prevent overfitting.

For each model, a grid of penalisation strengths was explored, and the optimal configuration was selected based on the concordance index and Akaike Information Criterion (AIC). The training pipeline was standardised across models to ensure fair comparison, with care taken to align preprocessing and ensure feature consistency.

The results, summarised in the table below, pointed to modest and broadly similar performance across the three distributions. The log-logistic model slightly edged ahead in terms of fit, achieving the lowest AIC, while the Weibull and log-normal models offered near-identical results. However, none of the models delivered a substantial improvement over the Cox model reported earlier.

In terms of interpretability, the log scale of the parametric coefficients made it possible to identify features with consistently strong influence across models. These included sponsor class, regulatory engagement, and various design features such as site count and masking. A stratified survival curve grid was generated to visualise how predicted survival curves shifted across values of these top predictors. While visually appealing, these curves showed limited separation, reinforcing the view that the models were struggling to capture sharp risk gradients.

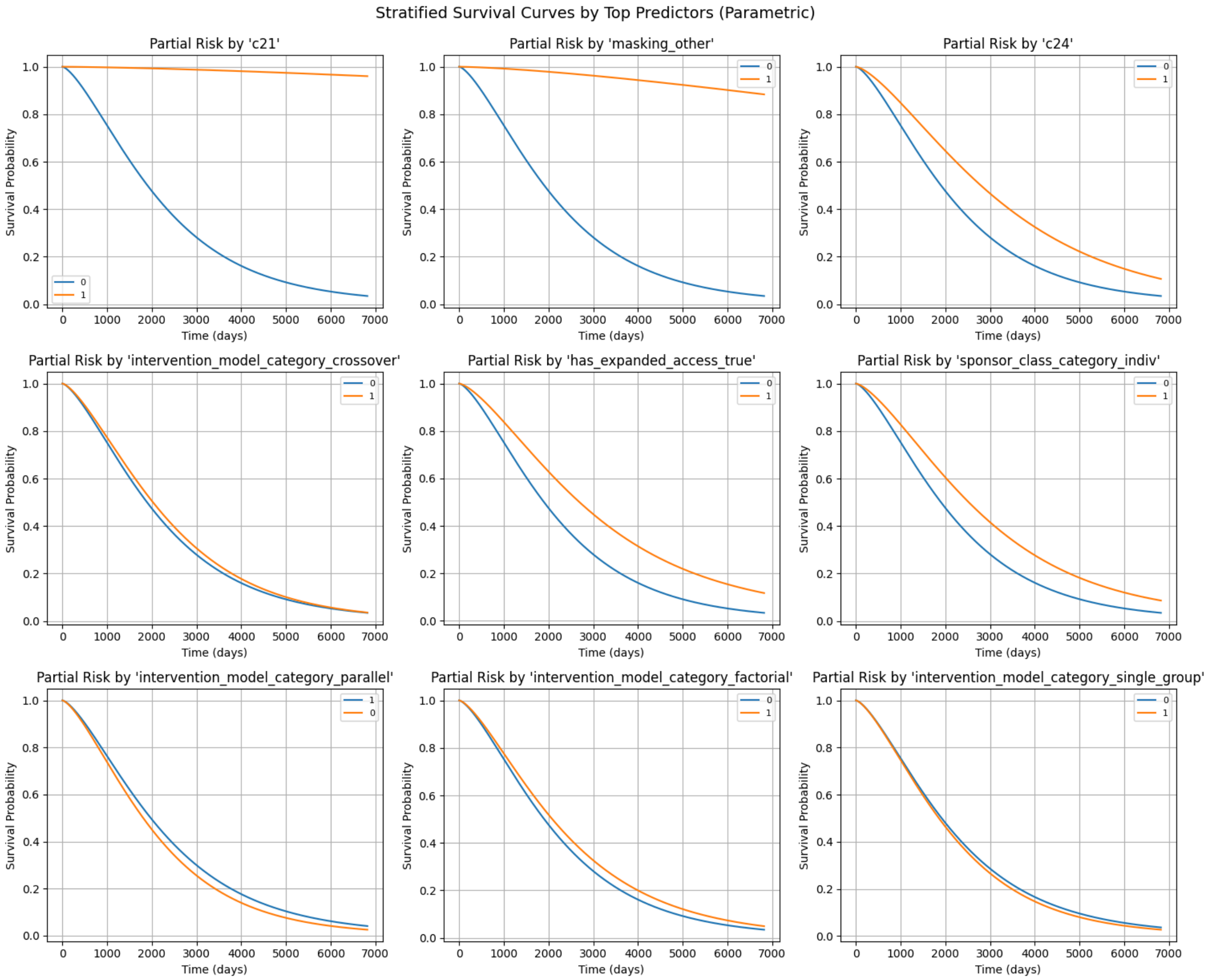
In short, the parametric models were well-behaved and computationally efficient, but they did not outperform their semi-parametric counterparts. This finding underscores the importance of aligning model complexity with data richness and cautions against assuming that stronger distributional assumptions will necessarily lead to better outcomes.

Table 1: Performance Comparison of Parametric Survival Models

Model	Penaliser	AIC	Log-likelihood	Concordance Index
Weibull	0.001	129241.53	–64524.77	0.6268
Log-logistic	0.001	129335.67	–64571.84	0.6277
Log-normal	0.001	130097.22	–64952.61	0.6268

The log-logistic model produced the highest concordance index, though the margin over the Weibull and log-normal models was slight. All three models yielded similar log-likelihood values, suggesting broadly equivalent fit. The differences in AIC were similarly modest, with none offering a decisive improvement over the others.

Stratified survival curves from the best-performing parametric model, showing survival probabilities over time by selected binary predictors.
Survival curves are grouped by key design variables, including masking type, intervention model, and sponsor class. While some features show mild separation in survival probability, the overall curves remain relatively close, reflecting limited risk stratification by the model.



3. Performance Evaluation and Observations

After selecting the best-performing version of each parametric model, we applied it to a held-out test set to see how well it could generalise. Each model produced median survival predictions, and these were evaluated using the concordance index, which measures how well the model can rank trials by their risk of early failure.

The results were rather modest. The log-logistic model, which had shown the best performance during training, achieved a concordance index just above 0.36 on the test set. This is not far above random chance and suggests that the model struggled to meaningfully distinguish higher-risk trials from more stable ones. Similar results were observed for the Weibull and log-normal models, both of which produced concordance scores in the same general range.

When the predictions were broken down by subgroups, there were few bright spots. For instance, trials sponsored by industry had slightly higher scores than those led by non-industry groups, but the difference was small. Trials with regulatory involvement, such as those flagged as FDA-regulated, performed a little better, yet still remained far below the level one would hope for in a practical decision-making tool. Certain subgroups, especially those with fewer observations, showed erratic behaviour and very low scores, further highlighting the limitations of the model.

Across continuous variables like site count, age ranges, or exclusion criteria, the results were equally unconvincing. Binned groups based on these features produced concordance values that hovered around 0.36 or below. This consistency at low levels suggests that the model was not able to pick up on useful patterns in these features, at least not in a way that transferred well to new data.

In truth, these models may have been too rigid for the task. While parametric survival models can perform well in clinical settings with structured, controlled data, the trial metadata we used here was highly variable, with many subtle interactions and missing context. It is quite possible that more flexible models, or richer feature engineering, would be needed to capture the complexities at play.

Nevertheless, the process was valuable. It helped establish a benchmark for what is currently achievable using parametric forms and reinforced the importance of careful validation when deploying any survival model in real-world settings.

4. Reflections and Practical Use

Although the parametric models explored in this study did not deliver strong predictive performance, their evaluation has still provided useful insight. It is often tempting to focus only on models that perform well, but understanding where and why models fall short is just as important, particularly in high-stakes domains such as clinical trial planning.

What became clear is that while parametric models are elegant and interpretable, their success depends heavily on the structure of the data and the accuracy of their underlying assumptions. In this case, the survival distributions they relied on may not have captured the complex risk landscape of clinical trials. The features available, while structured and clearly defined, likely missed many of the subtle contextual and operational factors that influence trial success. Timing of approvals, sponsor priorities, global logistics, and evolving eligibility interpretations are just a few examples of dynamics that fall outside what these models can see.

Despite this, there is still value in using parametric approaches as part of a broader toolkit. They offer closed-form expressions for survival probabilities, which can be helpful in specific scenarios, such as simulation studies or trial duration forecasting under controlled settings. Moreover, their transparency makes them useful for educational purposes and for grounding expectations before more flexible or opaque models are introduced.

Within the wider TrialIQ framework, these models help set the lower boundary of what predictive accuracy looks like when relying on protocol metadata alone. They reinforce the need for richer data inputs, careful feature design, and perhaps a shift towards models that can better capture non-linearities and interactions. Just as importantly, they serve as a reminder that no modelling effort should begin or end with performance metrics alone. Interpretation, transparency, and real-world utility remain just as vital.

In the end, while the parametric models did not meet the bar for deployment, their inclusion in this project has strengthened the overall understanding of what is required to predict trial viability with confidence. They will continue to play a role as reference models and as stepping stones towards more effective approaches.