Journal of Econometrics and Statistics
Vol. 5, Issue 2, 2025, pp. 269-282

Health Technology Assessment of Long-Term Benefits of Interventions Using Flexible Parametric Models: Model Selection

Szilárd Nemes[a],

[a] *Institute of Clinical Sciences, Sahlgrenska Academy, University of Gothenburg, Gothenburg, Sweden*

ABSTRACT
Accurate survival predictions are crucial for cost-effectiveness evaluations in health technology assessment (HTA). However, model selection remains a challenge. Flexible survival models eliminate the need to identify the exact data-generating or best-fitting parametric model, though informed choices are still required. Using the Kaplan–Meier curve as a reference and selecting parametric survival models that minimize the squared distance to non-parametric survival estimates provides an objective approach to model selection for survival data extrapolation. This paper extends this methodology by incorporating flexible parametric models and employs resampling methods, which are described, evaluated, and illustrated using both simulated and real-world data. Our findings demonstrate that resampling-based model selection enhances predictive accuracy, making it a valuable tool for survival extrapolation in HTA.

KEYWORDS
Flexible parametric survival models, Model selection,Mean squared error, Resampling, Health technology assessment

1. Introduction

Accurate extrapolation of survival data is essential for Health Technology Assessments (HTAs), as clinical trials frequently have limited follow-up periods [1, 2]. The selection of an appropriate survival model plays a pivotal role in determining long-term cost-effectiveness; nevertheless, identifying the optimal model remains a challenge. Traditional model selection methods, such as the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC), emphasize goodness-of-fit rather than predictive accuracy, which may result in suboptimal extrapolations when applied to survival data [3]. Flexible parametric models, particularly spline-based approaches, provide greater adaptability than conventional parametric models by allowing the hazard function to vary dynamically over time [4, 5]. These models are able to capture complex survival patterns without imposing a fixed parametric form, making them especially valuable in the context of HTAs. Determining the appropriate number of

CONTACT Author[a]. Email: szilard.nemes@outlook.com

Article History
Received : 06 May 2025; Revised : 28 May 2025; Accepted : 10 June 2025; Published : 28 June 2025

To cite this paper

Szilàrd Nemes (2025). Health Technology Assessment of Long-Term Benefits of Interventions Using Flexible Parametric Models: Model Selection. *Journal of Econometrics and Statistics.* 5(2), 269-282.

spline knots is critical, as too few knots oversimplify the hazard function, whereas an excessive number may cause overfitting and unreliable extrapolations [6, 7].

A commonly adopted method for evaluating model fit involves visually comparing the estimated survival function to the non-parametric Kaplan–Meier curve. However, visual inspection lacks formal statistical justification and may introduce subjectivity. As an alternative, the Focused Information Criterion (FIC) has been proposed, selecting models that minimize the mean squared distance between survival estimates and the Kaplan–Meier estimator, thereby prioritizing accuracy in specific quantities of interest [8, 9].

In this work, the FIC framework is extended to flexible parametric survival models through the introduction of a resampling-based approach. This method estimates the mean squared error (MSE) to facilitate selection of the optimal number of knots. Performance is evaluated through simulation studies and a real-world application to Diffuse Large B-Cell Lymphoma survival data. The findings indicate that resampling-based model selection enhances predictive accuracy and improves the reliability of long-term extrapolations in HTAs.

2. Survival data survival models

For a sample size n, let the survival times $X_1, \dots, X_n$ be independently and identically distributed according to the distribution function F, which is assumed to be differentiable for $x > 0$. Similarly, let $C_1, \dots, C_n$ be independently and identically distributed according to the distribution function G, which is also assumed to be differentiable for $x > 0$. The survival times and the censoring times are assumed to be independent. The observable random variables are defined as $T_i = \min(X_i, C_i)$ and $\delta_i = I(X_i \leq C_i)$, where $I()$ is an indicator function, taking the value 1 if the condition is fulfilled and 0 otherwise. The value t_i is the observed follow-up time, and δ_i indicates whether the survival time is uncensored or censored. The observed and ordered follow-up times are denoted as $t_1 < \dots < t_n$, with $\delta_1, \dots, \delta_n$ being their corresponding unordered indicators. Additionally, we define an at-risk process $R(t) = \sum_{i=1}^{n} I(X_i \geq t)$.

The non-parametric Kaplan–Meier or product-limit estimator of survival is given by:

$$S_{km}(t) = \prod_{t_i \leq t} \left\{ 1 - \frac{\delta_i}{R(t_i)} \right\} \tag{1}$$

with the associated variance estimated using the Greenwood formula:

$$v_{km}(t) = S_{km}^2(t) \sum_{t_i \leq t} \frac{\delta_i}{R(t_i) \left[R\left(t_i\right) - 1 \right]} \tag{2}$$

where the subscript $_{km}$ indicates that the survival or associated variance estimate refers to the non-parametric Kaplan–Meier estimator.

If we assume that the survival distribution F belongs to a parametric family of distributions with a parameter vector θ, then we can define the likelihood function as

$$\log L(\theta) = \sum_{i=1}^{n} \{\delta_i \log \lambda(t_i|\theta) + \log S(t_i|\theta)\} \tag{3}$$

where

$$\lambda(t|\theta) = \frac{\partial}{\partial t} H(t) \tag{4}$$

is the derivative of the cumulative hazard function $H(t)$.

We now introduce two competing approaches: a fully parametric model and a flexible parametric model, which we denote with the subscripts $_{pm}$ and $_{fs}$, respectively. In a fully parametric approach, it is necessary to identify $F(t)$ and subsequently $H(t)$. The parametric survival estimator is then estimated as $S_{pm}(t) = 1 - F(\theta; t)$. The flexible parametric survival model does not require the specification of $F(t)$ or $H(t)$. Instead, it utilizes splines to provide flexibility in modeling the baseline hazard (or log cumulative hazard). Specifically, it employs restricted cubic splines to allow the hazard function to vary smoothly over time without imposing strict parametric assumptions, thus yielding

$$\log H(t) = \mathcal{B}(t)\beta, \tag{5}$$

where $\mathcal{B}(t)$ represents a vector of spline basis functions evaluated at time t, and β represents a vector of spline coefficients estimated from the data.

Flexible survival models also incorporate an additional parameter known as knots. Knots represent points on the time axis where the spline transitions between different shapes, allowing for flexible and smooth modeling of the baseline hazard (or log cumulative hazard). The hazard functions estimated using restricted cubic splines closely align with the true function across a variety of complex hazard patterns, provided that a sufficient number of knots are used [6]. Typically, knots are positioned at quantiles of the log-observed event times to ensure that the spline accurately captures the overall shape of the hazard function [10]. The number and placement of knots significantly impact model fit. Too few knots may result in the spline failing to capture the complexity of the underlying hazard function [7], while an excessive number of knots can lead to overfitting, especially in areas with limited events, causing the hazard function to fluctuate excessively. Consequently, this overfitting could result in poor extrapolation beyond the observed trial period. The optimal number of knots cannot be directly estimated, so it is advisable to treat the number of knots as a tuning parameter, which becomes an integral part of the model selection process. Once the number of knots is determined and spline coefficients are estimated from the data, it becomes straightforward to transform them into survival functions, $S_{fs}(t) = \exp\left[-H(t)\right]$.

For both fully parametric and flexible survival models, the Delta method is used to approximate the variance. This method involves propagating the uncertainty in the parameter estimates (from the variance-covariance matrix) through to the predicted quantities using a Taylor series expansion, as follows:

$$\mathbf{v_{pm}} = \partial\mathbf{Spm}^{\mathbf{T}}\mathbf{I}^{-1}\partial\mathbf{Spm} \quad \text{and} \quad \mathbf{v_{fs}} = \partial\mathbf{Sfs}^{\mathbf{T}}\mathbf{I}^{-1}\partial\mathbf{Sfs} \tag{6}$$

where $\partial \mathbf{S}$ is the vector of partial derivatives of the survival function with respect to the parameter vector θ (for the fully parametric model) or β (for the flexible survival model), and $\mathbf{I}$ is the information matrix. For more details on the Fisher matrix under censoring, see [11]; for the model-agnostic information matrix used here, see [12].

Irrespective of the estimator, we assume that they are of the following form:

$$\hat{S}(t) = S(t) + \frac{1}{n}\sum_{i=1}^{n} \upsilon(T_i) + \epsilon_n \tag{7}$$

where $\upsilon(T_i)$ are the influence components with expectation 0 if the correct distribution is identified. In this setting, survival probabilities estimated from parametric and flexible parametric models represent marginal probabilities, i.e., intercept-only models.

3. Mean Squared Error and the Focused Information Criterion

MSE calculations usually require the true survival function, which is unknown in practical applications. Substituting the unknown true survival function with a consistent, unbiased estimate is fundamental to the Focused Information Criterion [8, 13]. In this framework, the Mean Squared Error (MSE) of a parametric survival function measures the average squared difference between the estimated survival probabilities and the observed survival, which is estimated by the Kaplan-Meier estimator.

The Kaplan-Meier estimator is uniformly consistent on intervals $[0, t]$ for which $S(t) > 0$ [14], and for any given t, the Kaplan-Meier estimator is almost unbiased [15, 16], satisfying $0 \leq S(t) - E[\hat{S}_{km}(t)] \leq e^{-R(t)}$. Asymptotically, the bias approaches zero as $n \to \infty$, making it effectively unbiased for large datasets. However, it is advisable to limit estimation to intervals where $R(t) \geq 5$, such that the small-sample bias of the Kaplan-Meier estimator is at most 0.0067. Thus, for $R(t) \geq 5$,

$$\mathrm{MSE}_{km}(t) \approx \frac{\upsilon_{km}(t)}{n}. \tag{8}$$

For the flexible parametric model,

$$\mathrm{MSE}_{fs}(t) = \frac{\upsilon_{fs}(t)}{n} + b^2_{fs}(t) \tag{9}$$

with the bias estimated as $\hat{b}_{fs}(t) = \hat{S}_{fs}(t) - \hat{S}_{km}(t)$. Regardless of the estimation method used, the survival estimates have a limiting normal distribution, and $E[\hat{b}_{fs}(t)] = b_{fs}(t)$ [9].

The quantity of interest is not the bias but its square. While the estimate of the bias is consistent, the squared bias may overestimate the quantity of interest by $n^{-1}\upsilon_b(t) + O(n^{-2})$ [17], where $\upsilon_b(t)$ is the variance of the bias at time t. Thus, estimation of MSE requires a corrected squared bias estimate for both the fully parametric and flexible parametric survival models by subtracting its variance. A caveat of this correction is that occasionally we may obtain a negative squared bias estimate. Thus, following [8], the correction is executed as

$$\max\left\{0, \hat{b}_{fs}^2(t) - \frac{\hat{v}_{b_{fs}}(t)}{n}\right\} \tag{10}$$

leading to

$$\mathrm{MSE}_{fs}(t) = \frac{v_{fs}(t)}{n} + \max\left\{0, \hat{b}_{fs}^2(t) - \frac{\hat{v}_{b_{fs}}(t)}{n}\right\}. \tag{11}$$

This correction requires an estimate for the variance of the bias, which is given by

$$v_{b_{fs}}(t) = v_{km}(t) + v_{fs}(t) - 2v_{km,fs}(t), \tag{12}$$

where $v_{km,fs}(t)$ is the covariance between the non-parametric Kaplan-Meier survival estimates and the flexible survival model estimates. The covariance between the Kaplan-Meier survival estimate and fully parametric models can be estimated using the influence function [9, 18]. However, extending this method to $v_{km,fs}(t)$ is more complex. Instead of relying on analytical estimates, we can turn to resampling methods, which will be described in the next section. There are two possible approaches: either use resampling to obtain an estimate of $v_b(t)$, or estimate $v_{km,fs}(t)$ directly and use it as a plug-in to estimate $v_b(t)$.

If one or more of the quantities involved in estimating MSE are asymptotically biased, the approximation has an error of order $\mathcal{O}(1/n)$, which can result in the variance contributing relatively little to the expected squared error [19]. Daehlen et al. [19] introduced a more precise bias estimate and a corrected MSE that is of order $o(1/n)$ by considering the error term of Equation 7 and estimating

$$c(t) = \lim_{n\to\infty} nE\left[\hat{S}(t) - S(t)\right] = \lim_{n\to\infty} nE\left[\epsilon_n(t)\right] \tag{13}$$

and correcting the MSE estimator as

$$\mathrm{MSE}_{fs}(t) = \frac{v_{fs}(t)}{n} + \max\left\{0, \hat{b}_{fs}^2(t) - \frac{\hat{v}_{b_{fs}}(t)}{n} + \frac{2\hat{b}_{fs}(t)\hat{c}(t)}{n}\right\}. \tag{14}$$

For correctly specified parametric models, the two formulations of MSE have the same limit since $c = 0$ with increasing sample size. For flexible parametric models, we cannot assume $c = 0$, as these models are, by definition, approximations of the true model.

4. Resampling survival models

4.1. *Bootstrap*

This section aims to provide the necessary background information for replicating the numerical evaluations. It is not intended to be a comprehensive review, but instead to offer the essential context for reproducing the specific calculations. For further details, readers are referred to [20], specifically Sections 3.5 and 7.3. We did not evaluate resampling with replacement of the pairs (t_i, δ_i) where $i = 1, \ldots, n$. Although this type of bootstrapping is widely used, it does not guarantee inclusion of the maximum observed follow-up time in the bootstrapped samples. This limitation renders the estimation of bootstrapped versions of key quantities impractical, as survival probability projections from the flexible parametric model would be required for $v_b(t)$.

The conditional bootstrap method involves simulating failure times from the estimated survival distribution. Subsequently, for each observation, its simulated censoring time is set equal to the observed censoring time if the observation was censored. If the observation was uncensored, the censoring time is generated from the estimated censoring distribution, conditioned on it being greater than the observed failure time.

First, we consider the bootstrapped version of $v_b^*(t)$, where the superscript $*$ indicates that this is a bootstrap variance estimate of the bias, given by

$$\hat{v}_{b^*}(t) = \frac{1}{B}\sum_{j=1}^{B}\left\{\hat{b}_{fs}^*(t) - \hat{b}_{fs}(t)\right\}^2 \tag{15}$$

and the covariance is calculated as

$$\hat{v}_{km,fs}^*(t) = E\left[\hat{S}_{km}^*(t)\hat{S}_{fs}^*(t)\right] - E\left[\hat{S}_{km}^*(t)\right]E\left[\hat{S}_{fs}^*(t)\right], \tag{16}$$

where B is the number of bootstrap resamples.

The bootstrapped version of the correction factor c is given by

$$\hat{c}_{boot} = \frac{n}{B}\sum_{j=1}^{B}\left\{\hat{S}^*(t) - \hat{S}(t)\right\}. \tag{17}$$

4.2. *Jackknife*

The Jackknife is a resampling technique used to estimate the bias and variability of a statistic. It works by systematically removing one observation at a time from the dataset and recalculating the statistic for each reduced sample. The results are then used to assess how much each individual data point influences the overall estimate. This method is particularly useful for constructing confidence intervals and reducing bias, especially in small sample sizes [21]. It serves as a computationally simpler alternative to other resampling methods, such as the bootstrap.

In our case, we aim to obtain jackknife estimates of either the covariance between the Kaplan-Meier and flexible parametric estimates of survival or a jackknife estimate of the variance for $\hat{b}_{fs}(t)$.

For both estimators, the jackknife estimate of survival is given by

$$\hat{S}^{jk}(t) = \frac{1}{n}\sum_{i=1}^{n} \hat{S}^{(-i)}(t), \tag{18}$$

where the superscript jk indicates the jackknife estimate. Subsequently, standard estimators are applied to compute the covariance and variance, respectively. The jackknife estimate for the correction factor c is given by

$$\hat{c}_{jack} = (n-1)\left\{\hat{S}(t) - \frac{1}{n}\sum_{i=1}^{n} \hat{S}^{(-i)}(t)\right\}. \tag{19}$$

5. Model Selection and the Behavior of the Selection Criterion

Conventional practice suggests selecting the model with the lowest mean squared error (MSE), which, in this context, is expressed as $MSE_{fs}(t) < MSE_{np}(t)$ or, equivalently:

$$RE(t) = \frac{MSE_{fs}(t)}{MSE_{km}(t)} < 1, \tag{20}$$

where RE denotes relative efficiency. This notation is chosen for pedagogical convenience and reflects that, when comparing two unbiased estimators, RE represents the ratio of variances, aligning with the concept of asymptotic relative efficiency. The limiting probability that RE is less than 1 is $P(\chi_1^2 < 2) = 0.843$ [8, 22]. This criterion should be interpreted pointwise across different time points and can be evaluated graphically by plotting the event times against the ratio of mean squared errors [22, 23]. If the time-indexed ratios are consistently below 1, the alternative models should be favored over the Kaplan-Meier estimator.

It should be noted, however, that it is not expected for the entire curve to remain below 1. The mean squared error (MSE) is estimated at the times when an event occurs, meaning that a total of $\sum_i \delta_i$ assessments will be conducted. If we define an indicator function I_i that takes the value 1 if $RE(t_i) < 1$, each I_i represents a Bernoulli trial with parameter $p = 0.842$. By the linearity of expectation:

$$E[\mathbf{I}] = E\left[\sum_i I_i\right] = \sum_i E[I_i] = 0.842\sum_i \delta_i, \tag{21}$$

Thus, if the true model is included in the set of competing models, it is expected that approximately 84.2% of the $RE(t)$ values will fall below 1. It should also be noted that these represent dependent Bernoulli trials with a complex covariance structure; since survival estimates are positively correlated, the variance of $\mathbf{I}$ will exceed 0.132, which is the variance of an independent Bernoulli trial with probability 0.843. When comparing competing models, preference should be given to those whose proportion of the $RE(t)$ curve below 1 is as close as possible to 84.2% (or higher), as this would indicate a better alignment with the expected behavior for the true model.

An alternative approach is to calculate the area under the $RE(t)$ curve, normalized by the maximum event time $t_{\max}$, as follows:

$$AUC_{RE(t)} = \frac{1}{t_{\max}} \int_0^{t_{\max}} RE(u), du \tag{22}$$

Heuristically, this area should be below 1 for the true model; however, no asymptotic results are currently available to formally support this assumption.

6. Numerical evaluations

6.1. *Data generating process*

We conducted simulation studies to assess the feasibility of model selection using bootstrap or jackknife resampling. In the first set of simulations, we generated survival times from an exponential distribution with a rate of 1/365, mimicking a scenario with a mean survival of one year. To replicate conditions typical in clinical trials, we incorporated low levels of random censoring, modeled using an exponential distribution with a rate of 1/1460, along with administrative censoring at day 365. For each simulated dataset (with sample sizes $n = 100, 200, 300$), we fitted Kaplan-Meier survival curves and flexible parametric survival models. Additionally, we applied a fully parametric exponential model as a benchmark. This process was repeated 1,000 times.

To provide a more realistic framework for testing the selection procedure, we conducted a second set of simulations using a mixed distribution. In this scenario, we again assumed sample sizes of 100, 200, and 300. The survival times were generated from a mixture of three distributions: one-third followed a log-logistic distribution (shape parameter 1.5, scale parameter 150), another third followed a uniform distribution between 0 and 1000 days, and the final third followed an exponential distribution with a rate of 1/365. Censoring was modeled as in the first simulation set. This procedure was also repeated 1,000 times. We used this second dataset to compare and assess knot selection and its effect on extrapolation accuracy. For this purpose, we fitted flexible survival models with 1 to 4 knots, extrapolating beyond the administrative censoring at 365 days out to a maximum of 3,650 days. We then assessed these projections against the actual observed survival probabilities as estimated by the Kaplan-Meier estimator.

For each iteration in both simulation sets, we estimated the jackknifed and bootstrapped covariance matrices between the competing estimators, along with the variance of the bias, and ultimately the corresponding $MSE(t)$ and $RE(t)$. This approach allowed us to evaluate the performance of our model selection method under both simple and complex survival scenarios, providing insights into its effectiveness and potential limitations.

6.2. *Criteria for evaluating the performance*

To assess the performance of our model selection approach, we employed both pointwise and global measures. For each iteration, we conducted pointwise assessments on a predefined time grid ranging from 0 to 365 days, with increments of 5 days. At each of these time points, we evaluated whether the relative efficiency $RE(t)$ was less than

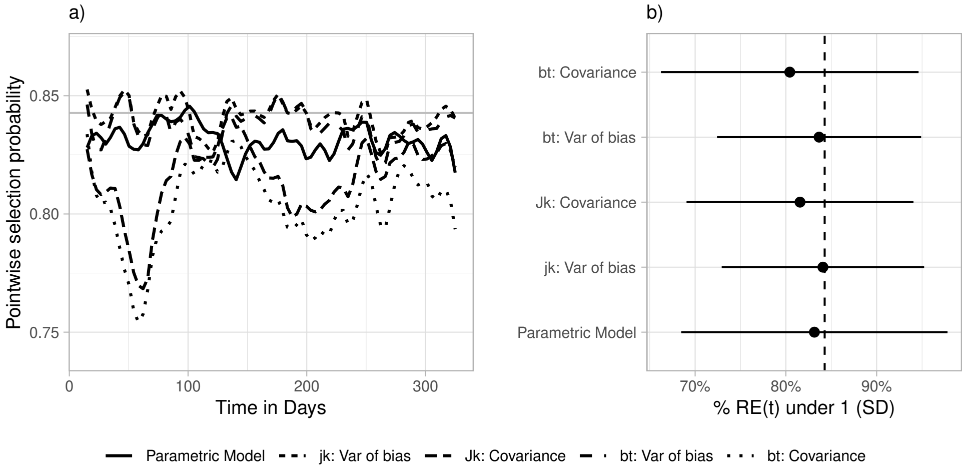

Figure 1. Panel a) presents the pointwise selection probabilities for the four competing selection estimators, along with the reference provided by the fully parametric model. Panel b) shows the percentage of the $RE(t)$ curve that falls below one, with a vertical reference line at 82.4%

1, indicating superior performance of the flexible parametric model compared to the Kaplan-Meier estimator.

As a global measure, we calculated the percentage of the time-indexed $RE(t)$ curve that falls below 1. This metric provides an overall indication of the flexible model's performance across the entire follow-up period. Additionally, we computed the area under the $RE(t)$ curve, $AUC_{RE(t)}$, normalized by the maximum follow-up time, as another summary measure of model performance.

For the second set of simulations, which involved extrapolation beyond the administrative censoring at 365 days, we extended our evaluation to assess the accuracy of long-term projections. We fitted Kaplan-Meier and flexible survival models to the time interval $(0, 365]$ and then extrapolated these models to $(365, \min(t_{max}, 3650)]$. To evaluate the extrapolation performance, we calculated both bias and percent bias at each observed event time beyond 365 days, averaging these measures across all such time points.

To compare different knot selection approaches, we estimated $RE(t)$, $AUC_{RE(t)}$, AIC, and BIC for flexible survival models with 1 to 4 knots. This allowed us to assess how different model selection criteria perform in identifying the optimal number of knots for extrapolation.

These performance measures were aggregated across the 1,000 iterations for each simulation scenario, providing a comprehensive summary of model performance and selection accuracy.

6.3. *Numerical Results*

Figure 1 summarizes the results from the simulation with exponential survival times, providing a detailed comparison of the different estimators across the follow-up period. The patterns of model selection were largely consistent across the different types of estimators. However, selection procedures based on jackknifed or bootstrapped covariance estimates exhibited lower selection probabilities compared to those based on the jackknifed or bootstrapped estimates of the variance of the bias.

The proportion of the relative efficiency, $RE(t)$, curve that fell below 1 consistently

Table 1. Agreement in knot selection between the four competing selection criteria, average number of knots and observed average bias and percent bias of the extrapolated survival curves.

	$AUC_{RE(t)}$	AIC	BIC	$Knots$	$Bias$	$\%Bias$
$P(RE(t)) < 1$	84%	24%	25%	2.31	-0.00095	12.2%
$AUC_{RE(t)}$	-	25%	27%	2.09	0.00568	19.7%
AIC	-	-	97%	2.58	0.0310	44.8%
BIC	-	-	-	2.49	0.0297	43.5%

favored estimation of the variance of the bias over the covariance between the two survival estimators (Figure 1b). This finding was further corroborated by the area under the $RE(t)$ curve. The variance of the bias estimate, whether using the jackknife (0.923, SD 0.061) or bootstrap (0.925, SD 0.612), was closest to the area under the curve for the fully parametric exponential estimator (0.812, SD 0.368).

For the second set of simulations using the mixed distribution, the results obtained from jackknifing and bootstrapping were largely similar. Thus, we present only the results for the jackknifed estimators. These results further confirmed the superiority of the selection procedure based on the resampled variance of the bias (Figure 2). Both the pointwise selection probabilities and the percentage of the $RE(t)$ curve falling below 1 improved as the sample size increased, indicating better performance of our method with larger datasets.

The results of the extrapolation analysis are presented in Table 1. The two MSE-based measures ($AUC_{RE(t)}$ and $P(RE(t) < 1)$) agreed on the optimal number of knots, just as the two likelihood-based criteria (AIC and BIC) demonstrated similar agreement. However, these two types of measures generally disagreed with each other. The likelihood-based criteria tended to select models with a higher number of knots, which resulted in reduced accuracy of the projections. The MSE-based knot selection method that used the jackknifed variance of bias demonstrated better extrapolation performance.

Additionally, we applied a correction to the mean squared error (MSE) as proposed by [19]. The results were largely identical to our original formulation, with only a minor decrease in selection probabilities (< 0.01) when using bootstrapping. This difference is statistically insignificant and may be attributed to the use of bootstrapping for the additional bias correction [24].

Overall, these numerical results demonstrate the effectiveness of our proposed resampling-based model selection approach, particularly when using the variance of the bias estimate. The method exhibits improved performance with increasing sample size and appears to be more robust in selecting models that provide accurate extrapolations compared to traditional likelihood-based criteria.

7. Applied example

Accurately modeling survival in diseases with complex hazard functions, such as Diffuse Large B-Cell Lymphoma (DLBCL), presents a significant challenge. Survival patterns in DLBCL often exhibit periods of high risk (e.g., during treatment), followed by remission phases, and potentially increasing risk due to late-stage recurrence or metastasis. Such variability makes it difficult to identify a standard parametric model that effectively balances bias and variance.

The dataset from [25] provides an example in which traditional parametric survival models—such as exponential, Weibull, gamma, and log-logistic distributions—struggle to accurately capture survival dynamics. Previous analyses suggest that these models

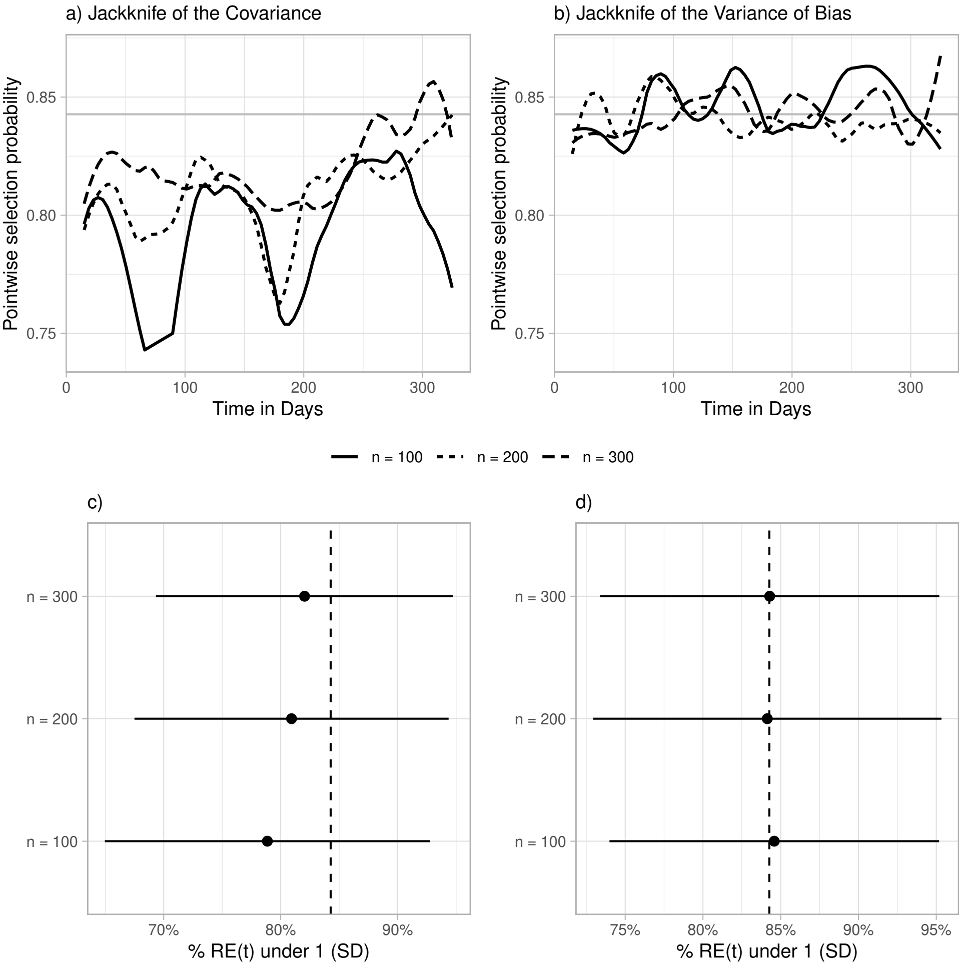

Figure 2. Panels (a) and (b) show the pointwise selection probabilities for the jackknifed covariance-based and variance of bias-based selection procedures as a function of sample size. These panels illustrate how the likelihood of selecting the correct model changes with increasing sample size. Panels (c) and (d) display the percentage of the $RE(t)$ curve that falls below 1, with a vertical reference line at 84.2%, representing the expected proportion.

tend to either over-smooth or misrepresent key features of the survival function [22]. We applied our flexible parametric model selection approach to evaluate whether these models could offer a better alternative to the Kaplan-Meier estimator.

Figure 3(a) presents survival estimates obtained using the Kaplan-Meier method (stepwise function) and flexible parametric models (continuous curves). The flexible parametric models closely follow the Kaplan-Meier estimates, suggesting a favorable bias-variance trade-off. Model selection using the mean squared error (MSE) criterion indicated that flexible parametric models outperformed the non-parametric Kaplan-Meier estimator, as shown in Figure 3(b), though not at the extremes of the follow-up time.

Table 2 compares different model selection approaches for determining the optimal number of spline knots. Both AIC and BIC agreed on a two-knot solution, aligning with the jackknife-based MSE selection. However, the bootstrap-based MSE selection process struggled to distinguish between different knot numbers.

Interpreting these results in conjunction with the relative efficiency (RE) curve suggests that extrapolation should be approached with caution. Approximately 65% of

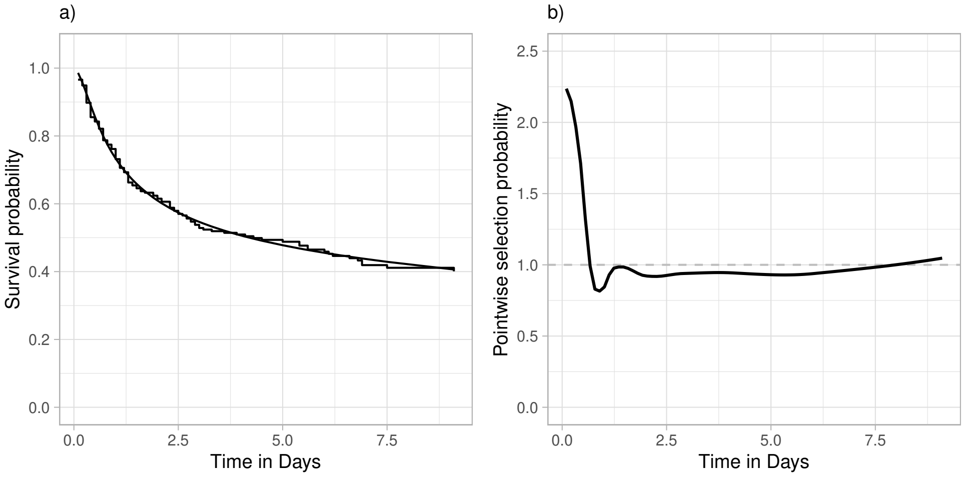

Figure 3. Panel (a) presents the survival probabilities estimated by the Kaplan-Meier estimator (stepwise function) and by flexible parametric models (continuous smooth curve) for mortality associated with Diffuse Large B-Cell Lymphoma. Panel (b) depicts the pointwise probability of selecting the flexible parametric models over the Kaplan-Meier estimator, accounting for the bias-variance trade-off.

Table 2. Knot number selection for flexible parametric survival model applied to Diffuse Large-B-Cell Lymphoma survival

Knots	AUC	$P(RE < 1)$	AIC	BIC
1	2.489	0.269	745.48	752.40
2	0.933	0.655	710.38	720.76
3	0.961	0.646	711.84	725.67
4	0.983	0.553	713.18	730.48
5	0.979	0.507	713.69	734.45

the RE curve falls below 1. However, deviations at the extremes of follow-up indicate potential discrepancies between the modeled and observed survival probabilities. Notably, early deviations may have minimal impact on extrapolation, but discrepancies at longer follow-up times suggest a risk of inaccurate survival projections.

This analysis underscores the importance of combining graphical assessments with scalar selection criteria when applying flexible parametric survival models. The results indicate that flexible models can provide robust survival estimates for diseases like DLBCL, but careful consideration of bias-variance trade-offs and the reliability of extrapolations is essential.

8. Discussion

Flexible parametric survival models, introduced by [26], have gained significant attention because these models are adept at handling complex hazard functions without the challenge of needing to specify the parametric family of distributions that generated the data. Choosing the right model is crucial, as extrapolated survival probabilities can vary significantly depending on the model selected.

The FIC framework [8, 9, 19], which uses the Kaplan-Meier curve as a reference, provides an objective mechanism for determining the appropriate number of knots and quantifies the trade-off between bias and variance. This ensures that the selected model not only outperforms alternatives but also aligns closely with the underlying data. Although this additional step introduces some complexity, the trade-off is justified by its potential to improve long-term extrapolations. Applying this framework to flexible parametric survival models requires several decisions, including the selection of the MSE variant and the methods used for its estimation. The original FIC/MSE framework introduced by [8, 9] does not account for the inherent (small-sample) bias of estimators. However, the updated framework proposed by [19] offers a more precise bias estimate and ensures that variability is given adequate weight in the calculated FIC/MSE. The updated procedure eliminates bias in maximum likelihood estimators, which helps reduce the MSE of parameter estimates [27]. Such bias estimation and correction are inherently "corrective" [28], yet these adjustments are typically not applied to extrapolated survival models. Moreover, in the context of flexible parametric survival models, the bias in survival estimates is often small enough that additional correction may have little practical impact.

Regardless of the chosen MSE/FIC approach, using jackknifing instead of bootstrapping, and estimating the variance of the bias directly rather than through the covariance of competing estimators, offers both numerical and practical advantages.

While the increased complexity of this method may pose challenges, integrating the FIC/MSE framework into extrapolation-based HTA assessments enables researchers to achieve a better balance between bias and variance, thereby improving model selection and subsequent extrapolation.

References

[1] Bell Gorrod H, Kearns B, Stevens J, Thokala P, Labeit A, Latimer N, et al. A review of survival analysis methods used in NICE technology appraisals of cancer treatments: consistency, limitations, and areas for improvement. Medical Decision Making. 2019;39(8):899-909.

[2] Gallacher D, Kimani P, Stallard N. Extrapolating parametric survival models in health technology assessment: a simulation study. Medical Decision Making. 2021;41(1):37-50.

[3] Latimer NR. Survival analysis for economic evaluations alongside clinical trials—extrapolation with patient-level data: inconsistencies, limitations, and a practical guide. Medical Decision Making. 2013;33(6):743-54.

[4] Gray J, Sullivan T, Latimer NR, Salter A, Sorich MJ, Ward RL, et al. Extrapolation of survival curves using standard parametric models and flexible parametric spline models: comparisons in large registry cohorts with advanced cancer. Medical Decision Making. 2021;41(2):179-93.

[5] Chen EYT, Leontyeva Y, Lin CN, Wang JD, Clements MS, Dickman PW. Comparing Survival Extrapolation within All-Cause and Relative Survival Frameworks by Standard Parametric Models and Flexible Parametric Spline Models Using the Swedish Cancer Registry. Medical Decision Making. 2024:0272989X241227230.

[6] Rutherford MJ, Crowther MJ, Lambert PC. The use of restricted cubic splines to approximate complex hazard functions in the analysis of time-to-event data: a simulation study. Journal of Statistical Computation and Simulation. 2015;85(4):777-93.

[7] Syriopoulou E, Mozumder SI, Rutherford MJ, Lambert PC. Robustness of individual and marginal model-based estimates: A sensitivity analysis of flexible parametric models. Cancer epidemiology. 2019;58:17-24.

[8] Jullum M, Hjort NL. Parametric or nonparametric: The FIC approach. Statistica Sinica. 2017:951-81.

[9] Jullum M, Hjort NL. What price semiparametric Cox regression? Lifetime Data Analysis. 2019;25(3):406-38.

[10] Lambert PC, Royston P. Further development of flexible parametric models for survival analysis. The Stata Journal. 2009;9(2):265-90.

[11] Miller RG. What price Kaplan-Meier? Biometrics. 1983:1077-81.

[12] Hjort NL. On inference in parametric survival data models. International Statistical Review/Revue Internationale de Statistique. 1992:355-87.

[13] Claeskens G, Hjort NL. The focused information criterion. Journal of the American Statistical Association. 2003;98(464):900-16.

[14] Andersen PK, Borgan O, Gill RD, Keiding N. Statistical models based on counting processes. Springer Science & Business Media; 2012.

[15] Meier P. Estimation of a distribution function from incomplete observations. Journal of Applied Probability. 1975;12(S1):67-87.

[16] Borgan Ø. Kaplan–Meier Estimator. Wiley StatsRef: Statistics Reference Online. 2014.

[17] Lehmann EL, Casella G. Theory of point estimation. Springer Science & Business Media; 2006.

[18] Reid N. Influence Functions for Censored Data. The Annals of Statistics. 1981;9(1):78 92.

[19] Dæhlen I, Hjort NL, Hobæk Haff I. Accurate bias estimation with applications to focused model selection. Scandinavian Journal of Statistics. 2024;51(2):724-59.

[20] Davison AC, Hinkley DV. Bootstrap methods and their application. 1. Cambridge university press; 1997.

[21] Miller RG. The jackknife-a review. Biometrika. 1974;61(1):1-15.

[22] Nemes S. Parametric analysis and model selection for economic evaluation of survival data. Model Assisted Statistics and Applications. 2024;19(2):123-31.

[23] Klein JP, Moeschberger ML. The robustness of several estimators of the survivorship function with randomly censored data. Communications in Statistics-Simulation and Computation. 1989;18(3):1087-112.

[24] Efron B, Tibshirani RJ. An introduction to the bootstrap. Chapman and Hall/CRC; 1994.

[25] Rosenwald A, Wright G, Chan WC, Connors JM, Campo E, Fisher RI, et al. The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. New England Journal of Medicine. 2002;346(25):1937-47.

[26] Royston P, Parmar MK. Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. Statistics in medicine. 2002;21(15):2175-97.

[27] Mardia K, Southworth H, Taylor C. On bias in maximum likelihood estimators. Journal of statistical planning and inference. 1999;76(1-2):31-9.

[28] Firth D. Bias reduction of maximum likelihood estimates. Biometrika. 1993;80(1):27-38.