Assessing survival models by interval testing

Ben Lee

June 1, 2024

Abstract

When considering many survival models, decisions become more challenging in health economic evaluation. In this paper, we present a set of methods to assist with selecting the most appropriate survival models. The methods highlight areas of particularly poor fit. Furthermore, plots and overall p-values provide guidance on whether a survival model should be rejected or not.

Keywords: Survival analysis, calibration, health economics, decision making

An R-package for these methods, including examples is available here.

1 Introduction

The primary results of a clinical trial are often right-censored survival data. To understand the efficacy of a treatment, survival models must be fitted to this data. A choice between fitted survival models must then be made, and this is often the primary driving factor in the economic evaluation of a new medical product. During this process, a large number of models are usually considered. However, with a growing emphasis on flexible survival mdoels[1] and a requirement for standard parametric models[2], the number of models considered can easily exceed 20, making decisions challenging. Our proposed methods can assist this process by indicating which survival models provide implausible estimates.

When we plot these survival models, we will refer to them as survival curves. Recommendations on selecting the most appropriate survival models are provided in NICE TSD 14[2]. These include visual curve inspection, AIC/BIC tests, clinical validity, and the use of external data. Except for AIC/BIC, these approaches are subjective and often time-consuming to implement, especially when considering many curves. It would be useful if curves were quickly shown to be implausible. This would reduce the number of feasible curves, making the decision process easier. One approach that may facilitate this is calibration, but this is rarely used in health technology assessment.

In survival analysis, calibration refers to the agreement between predicted probabilities and observed events within a specific time interval. The most common calibration methods are extensions of the Hosmer-Lemeshow goodness-of-fit test[3],[4]. These include the Grønnesby and Borgan test (specific to the Cox proportional hazard model)[5], the Nam and D'Agostino test (applicable to more general models)[6], as well as smoothed calibration curves and their associated metrics[7]. In all of these approaches, patients are assigned a risk score based on the fitted model and are grouped (often into deciles) based on these scores. The groups are then evaluated to determine how well-calibrated a survival model is. Our approach is different from this because patients do not need to be assigned risk scores. This is especially useful for models where the risk scores are equal among all patients, which is the case for our worked examples.

The paper is structured as follows: methods, application, and discussion. The methods section describes: a new calibration approach with intervals based on censoring, an extension to this where intervals can be specified, and finally methods for generating an overall p-value. The application section considers two examples. The first example evaluates a simple exponential model, providing an introduction to the developed methods. The second example is more practical and considers survival data which is challenging to model. In this example, seven parametric survival models are fitted. Subsequently, the full range of developed methods evaluate which curves are most appropriate. Finally, before the discussion is given, the results of a simulation experiment are summarized.

2 Methods

2.1 Interval testing

We consider right-censored survival data, where for each subject we have time and event given by $(t, E) \in (\mathbb{R}_{>0}, \{1, 0\})$. We split this data into two types:

- The observed data (event=1) with times $t_{i,1}$ such that $0 < t_{i,1} \leq t_{i+1,1}$ and $i = 1, \ldots, E$.
- The censored data (event=0) with times $t_{i,0}$ such that $0 < t_{i,0} \leq t_{i+1,0}$ and $i = 1, \ldots, C$.

Subsequently, we define the set of unique censor times $t'_{j,0}$ for $j = 1, \ldots, J$. That is, the values $t'_{j,0}$ are the unique values from the set of $t_{i,0}$ values, and thereby the $t'_{j,0}$ values satisfy $0 < t'_{j,0} < t'_{j+1,0}$. We now define the intervals of sequential censor times such that for $i = 1, ..., J-1$

$$
\begin{aligned}
I_0 &= (0,\ t'_{1,0}] \\
I_i &= (t'_{i,0},\ t'_{i+1,0}] \\
I_J &= (t'_{J,0},\ \infty).
\end{aligned}
$$

We are interested in the hypothesis that the event generating process is described by the model $M(\boldsymbol{\beta})$, with probability density function f_M and fixed parameters $\boldsymbol{\beta}$.

Thereby, imagine we have not seen the data. However, we still have the intervals I_j and the model $M(\boldsymbol{\beta})$. We now make the assumption that the intervals themselves, I_j, are uninformative of the event generating process. It follows that each patient that enters the interval I_j:

- either experiences an event or does not experience an event in interval I_j.
- has the same probability of experiencing an event as any other patient entering I_j.
- has a probability of experiencing an event defined by the event generating process $M(\boldsymbol{\beta})$ (when assuming the null hypothesis holds).

It follows that under the null, the following process occurs within each interval:

$$
events(I_j) \sim binomial(N_{I_j}, p_{I_j}), \tag{1}
$$

where $events(I_j)$ are the number of events observed within the interval I_j.

p_{I_j} ="the probability of a subject experiencing an event in interval I_j given the subject makes it to interval I_j", which is given by

$$
\begin{aligned}
p_{I_j} &= \mathbb{P}(T \in I_j \mid T > min(I_j)\ ) \\
&= \mathbb{P}(T \in I_j)\ /\ \mathbb{P}(T > min(I_j)) \\
&= \int_{I_j} f_M(t)dt \Big/ \int_{min(I_j)}^{\infty} f_M(t)dt.
\end{aligned} \tag{2}
$$

Finally, N_{I_j} ="the number of subjects at the start of interval I_j." That is, the number of starting patients $(C + E)$, minus the patients lost to events in the previous $j-1$ intervals, minus the patients lost to censors in the previous $j-1$ intervals. That is,

$$
N_{I_j} = \ C + E\ - \sum_{k=0}^{j-1} events(I_k)\ - \sum_{k=0}^{j-1} censors(I_k), \tag{3}
$$

where we define the number of censors observed within the interval I_j as $censors(I_j)$. Note that this quantity usually takes a value of 1. (However, there are exceptions to this. For example, the final interval J has $censors(I_J) = 0$ and intervals with multiple censors of equal time at the upper boundary of the interval have $censors(I_j) > 1$.)

For each interval I_j, we have specified a binomial model with known N_{I_j}, p_{I_j}, and observed $events(I_j)$. Each interval I_j thereby has a midpoint p-value for observing $events(I_j)$ in that interval.

Extension to specified intervals

The methodology above describes how events can be modelled within each interval. However, the intervals themselves are entirely determined by the censor times. It is often of interest to understand events in specified intervals (for example 10 evenly-spaced intervals up to the last censor.) Below, we extend the methodology to allow for such an analysis.

Define the intervals of interest for $k = 1, \ldots, K$ as

$$\begin{aligned} V_1 &= (S_0,\ S_1] \\ &\ldots \\ V_k &= (S_{k-1},\ S_k]. \end{aligned}$$

We have unique censor times $t'_{j,0}$ for $j = 1, \ldots, J$. Define the set of unique times τ_i as the unique set of the union of the $t'_{j,0}$ values and the S_k values, with $\tau_i < \tau_{i+1}$ for $i = 1, ..., L_\tau - 1$, where L_τ denotes the number of unique τ values. We now construct the intervals

$$\begin{aligned} I'_0 &= (0, \tau_1] \\ &\ldots \\ I'_i &= (\tau_i, \tau_{i+1}] \\ &\ldots \\ I'_{L_\tau} &= (\tau_{L_\tau}, \infty]. \end{aligned}$$

As in equation 1, we now have a known distribution for each interval I'_i. The observations in said intervals are described by

$$events(I'_i) \sim binomial(N_{I'_i}, p_{I'_i}),$$

with $p_{I'_i}$ and $N_{I'_i}$ defined analogously to equations 2 and 3, with the sum in equation 3 now being over the I' intervals.

We now consider the I'_i intervals as the subintervals that make up each V_k interval. Define the collection of subintervals $l = 1, ..., L_k$ as

$$\{I'_{l,V_k}\} = \{I'_i \quad such\ that \quad I'_i \subseteq V_k\}.$$

It follows that for each interval V_k its subintervals follow

$$\begin{aligned} events(I'_{1,V_k}) &\sim binomial(N_{I'_{1,V_k}}, p_{I'_{1,V_k}}) \\ &\ldots \\ events(I'_{L_k,V_k}) &\sim binomial(N_{I'_{L_k,V_k}}, p_{I'_{L_k,V_k}}). \end{aligned}$$

Thereby $\sum_{l=1}^{L_k} events(I'_{l,V_k})$ is distributed as the realization of a sum of binomials with known parameters. That is, the number of observed events in each interval V_k follows a known distribution. We can thus obtain a midpoint p-value for each specified interval V_k.

2.2 Test statistics

The primary recommendation of this paper is the use of interval plots, such as Figure 1b. To support these plots, a single overall p-value can be a useful tool. Such a p-value can be obtained by condensing the plot's constituent p-values into a single overall p-value.

When deriving these test statistics, we assume that the midpoint p-value for each interval follows a continuous uniform[0,1] distribution under H0. When it comes to the examples, this assumption does not necessarily hold. This is justified in detail in Appendix A and C.

We derive two test statistics for obtaining an overall p-value. From Birnbaum (1954)[8], we know that selecting a best test statistic for H0 is not trivial, and a more modern discussion is given in [9]. Our first test statistic is continuous and allows for extreme p-values to have a high influence on the overall p-value. The second test statistic is discrete and aims to mitigate the influence of extreme p-values.

2.2.1 Transformed Fisher test (TFT) statistic

To motivate the test, we consider a p-value for a given interval. A high p-value indicates more events have occurred in the interval than would generally be expected under H0. A low p-value indicates fewer events than expected. It follows that both high and low p-values indicate poor model fit within a given interval. As such, we will construct a test statistic that can detect an uncommon amount of high and low p-values.

For each interval, we have a midpoint p-value, p_j. Under H0, we make the assumption these p-values are $uniform[0,1]$. The p-values are transformed to the form

$$U_j = \begin{cases} 2p_j & for \;\; p_j \leq 0.5 \\ 2(1-p_j) & for \;\; p_j > 0.5 \end{cases}$$

It follows that $U_j \sim uniform[0,1]$. By the classical result from Fisher, we obtain the test statistic

$$T_{cont} = -2\sum_{j=1}^{I} log(U_j) \sim \chi^2_{2I},$$

where we reject for large T_{cont} and I is the total number of intervals. The intuition for this test is that unfavourable p-values (close to 0 or 1) are transformed such that they correspond to U_j values close to 0. This makes the overall test able to detect unfavourable p-values.

2.2.2 Protection Against Very Small Intervals (PAVSI) test statistic

The transformed Fisher test is useful as it enables us to leverage the full range of reported p-values. However, in some situations, this may not be ideal. For example, a single p-value that is very close to 0 can heavily influence the test. We argue that a test that is not dominated by extreme p-values is a useful tool.

For each interval, we have a midpoint p-value, p_j. Under H0, we make the assumption these p-values are $uniform[0,1]$. The p-values are transformed to the form

$$V_j = \begin{cases} 1 & for \;\; p_j \geq 0.975 \;\; or \;\; p_j \leq 0.025 \\ 0 & else \end{cases}$$

It follows that under H0 we have $V_j \sim bernoulli(0.05)$, and thereby

$$T_{pavsi} = \sum_{j=1}^{I} V_j \sim binom(I, 0.05),$$

where we reject for large T_{pavsi} values. It is noted that T_{pavsi} is a discrete test statistic, and we recommend reporting the midpoint p-value for this, see Appendix A.

Compared to the transformed Fisher test, the PAVSI is more pragmatic in terms of extreme p-values. However, for PAVSI, a p-value of say 0.026 would contribute equivalently to a p-value of 0.5, which is arguably a limitation of this approach.

3 Application

The methods above described two general approaches to interval testing:

- Use the censor times to specify the intervals
- Specify the intervals yourself ("extension to specified intervals")

In example 1, we focus on the first approach only. In the second example, both approaches are described. Note that the two test statistics (2.2) can be used in either approach.

For our examples, we use real world overall survival data for melanoma as is described in [10]. Example 1, uses the 'Dabrafenib' treatment arm of the BREAK-3 trial[11]. Example 2, uses the 'Dabrafenib+Trametinib' treatment arm of the COMBI-d trial[12]. All data and code is available on Github here.

3.1 Example 1: A fitted exponential model

In this example, we fit the simplest survival model (exponential) to our data. Following this, we want to determine how appropriate the exponential model is. First, we do a visual check of the curve fit, Figure 1a. Visually, the Kaplan-Meier curve and the exponential curve seem to be misaligned, suggesting poor fit. However, an interval test provides a more rigorous assessment of this, Figure 1b. For this example, we are only interested in an interval test where the intervals are selected using the censor times.

To perform the test, we obtain the values I_j, N_{I_j}, and $events(I_j)$ directly from the data. From Equation 1, it remains to obtain p_{I_j} for the fitted exponential model. As the probability density function for the exponential distribution is given by $f_M(t) = \lambda e^{-\lambda t}$, using Equation 2, it follows that

$$p_{I_j} = 1 - exp(\lambda[min(I_j) - max(I_j)]\ ).$$

As the value λ is known (from fitting the exponential model to the data), it follows that we have fully specified binomial realisations for each interval I_j. This information for the first five intervals is presented in Table 1.

I_j	N.risk	p_{I_j}	$events(I_j)$	$\mathbb{E}$(events)	p-mid	p-rand
(0, 1]	187	0.026	1	4.8	0.027	0.034
(1, 1.6]	185	0.016	1	3	0.127	0.111
(1.6, 2.5]	183	0.024	1	4.4	0.037	0.059
(2.5, 3.4]	181	0.023	4	4.2	0.491	0.538
(3.4, 4.2]	176	0.021	5	3.6	0.771	0.762
...	...	...	...	...	...	...

Table 1: Summary of interval data for Example 1

The p-value for each interval is discrete as it comes from a binomial model. It follows that each discrete p-value does not follow a continuous uniform[0,1] distribution under H0. This can be rectified by using randomized p-values[13]. However, in the interest of reproducibility and clarity, we will focus our discussion on the results of the midpoint p-values. (This is discussed in more detail in Appendix A.)

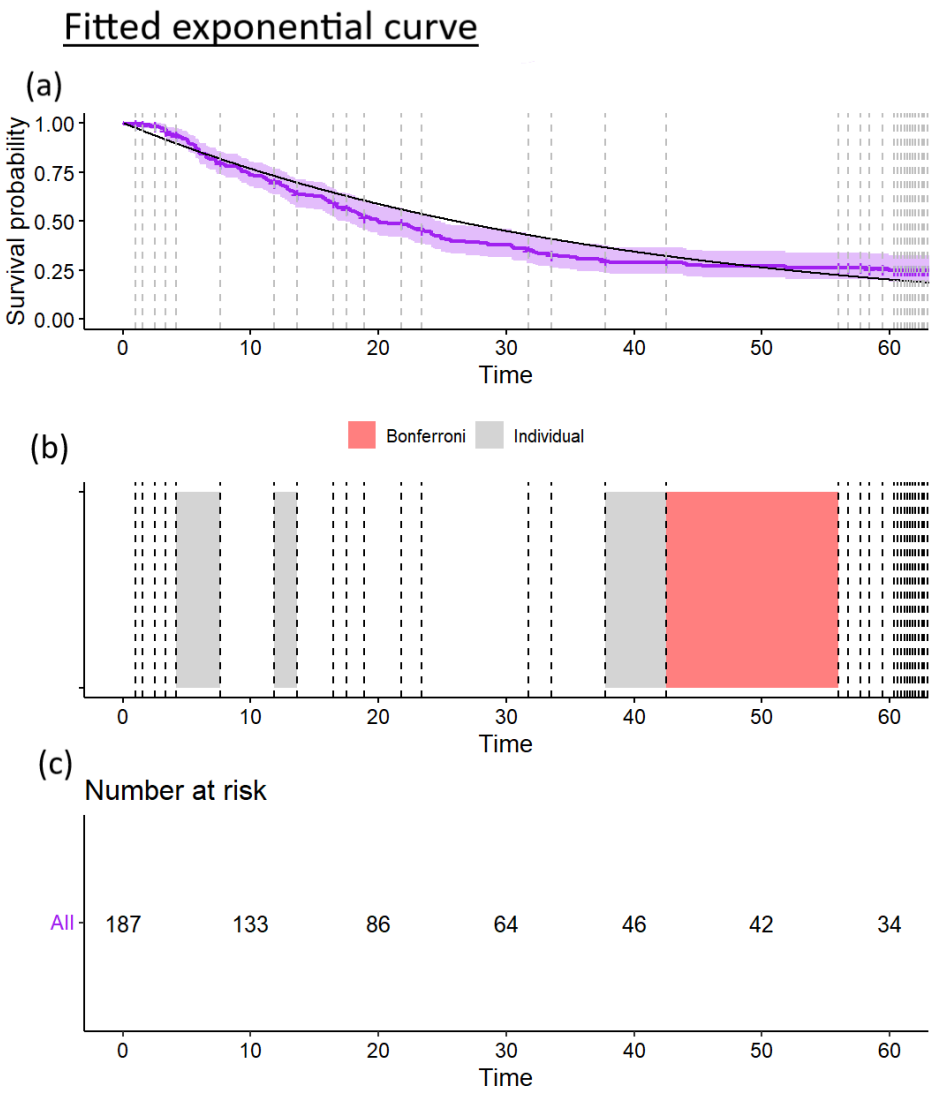

Figure 1: **Part a:** The KM curve for the data in purple, censors are given by dashed vertical black lines, and the fitted exponential curve is the overlaid curve in black. **Part b:** the binomial interval test for this data. Intervals are defined by the censors, where for each interval, red indicates a Bonferroni rejection, and grey indicates an individual flag. **Part c:** Numbers at risk for the data

For each interval, we consider the midpoint p-value with respect to a 2-sided test. That is, if $p_{I_j} \leq 0.025$ or $p_{I_j} \geq 0.975$, we have reason to suspect the interval generating said value. We call this the individual test.

Of greater concern, are the intervals that fail the Bonferroni test[14]. The Bonferroni test is chosen to strongly control the familywise error rate at a level of 0.05. In particular, we reject if $p_{I_j} \leq 0.025/I$ or $p_{I_j} \geq 0.975/I$, where I ="the total number of intervals". This means that under H0 (that the model holds for all intervals) a single Bonferroni rejection of an interval is only expected to occur for 1 in 20 times the data is generated under H0.

In Figure 1, we present the results for the Bonferroni and individual tests alongside the Kaplan-Meier (KM) curve, the fitted exponential model, and the numbers at risk. A single Bonferroni rejection of an interval is good reasoning for rejecting a model. In Figure 1, the extremely poor fit (Bonferroni rejection) is at the end of the curve. This is very unfavourable in health economic evaluation, where plausible extrapolations beyond the observed data are paramount to model selection. That is, the Bonferroni rejection at the end of the curve indicates that the extrapolation will be poor, which is more detrimental compared with a Bonferroni rejection at the start of the curve.

An overall p-value for the fitted exponential model is also of interest. For the transformed Fisher test statistic, there are $I = 43$ intervals, and our test statistic was computed as $-2\sum_{j=1}^{I} log(U_j) = 81.84$, giving a p-value of 0.607. For the PAVSI test statistic, there are $I = 43$ intervals. From Figure 1, we can read off that $T_{pavsi} = 4$ (as there are 3 individual rejections and 1 Bonferroni rejection). This gives us an overall midpoint p-value of 0.114. (These test statistics should be used with caution. In Appendix C, the test statistics are shown to be highly conservative when using the censor interval approach.)

As we can see, the two test statistics yield very different results. It is our recommendation that a model should not be accepted or rejected based entirely on the outcome of such tests. Instead, they are tools that provide an initial overview of model fit. The greatest detail and insight can be obtained by assessing the binomial interval plot alongside specific interval p-values and an overall p-value. For this particular example, we would recommend that the model is rejected. This is primarily due to the Bonferroni rejection at the end of the curve.

3.2 Example 2: Selecting a parametric survival model for challenging data

In this example, we fit seven standard parametric survival models. The survival data has a challenging shape, making it difficult for the models to provide good visual fit. Our aim is to use the interval test methods to determine whether any of the survival models are appropriate or whether flexible survival models are required.

Intervals selected based on censors

First, we perform the interval test with intervals chosen using censors as in Example 1. The 7 parametric models were fitted using flexsurv [15]. The final plots and tables were generated by the BITsurv package which is available here.

The test statistics for each parametric model are presented in Table 2. The PAVSI and TFT overall p-values are well above the 0.05 threshold for all of the survival models. This might lead us to believe that the models are providing good fit. However, as discussed in Appendix C, these test statistics are highly conservative when using the censor interval approach, meaning the overall p-values are not informative unless they indicate a rejection (less than 0.05.)

For each model, there are between 1 and 4 individual flags. However, these flags are primarily descriptive. Of greater interest are the Bonferroni rejections, of which there are none. This seems good, and from Table 2 alone we might believe that our curves provide good fit. However, this is an unfortunate result of using intervals based on censors, see poor fit in Figure 2. Due to the large number of censors, the majority of intervals are very small, which makes it difficult to detect poor fit within these intervals. For example, there are many small intervals for the last 5 months of the curve. It would be more informative to consider a single interval for those 5 months. Because of these limitations, a specified interval approach is required.

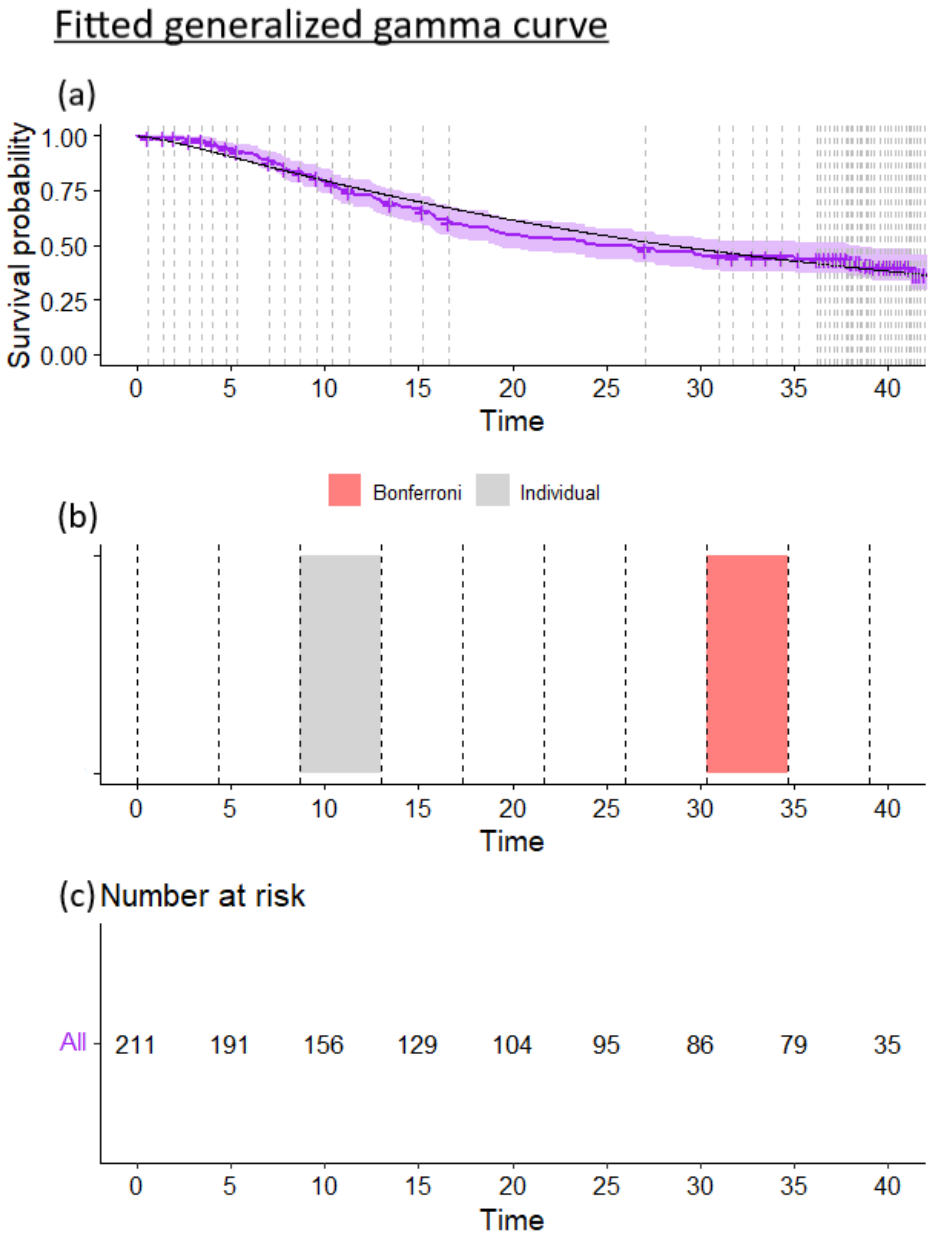

Figure 2: **Part a:** The KM curve (purple), censors (vertical dashed lines), and the fitted generalized gamma curve (black). **Part b:** The binomial interval test results for 10 evenly-spaced intervals. **Part c:** Numbers at risk.

Parametric survival model	Bonferroni rejections	Individual flags	PAVSI	TFT
Exponential	0	3	0.468	0.854
Gamma	0	2	0.686	0.887
Generalised gamma	0	1	0.881	0.955
Gompertz	0	4	0.267	0.884
Log-logistic	0	1	0.881	0.977
Log-normal	0	2	0.696	0.968
Weibull	0	3	0.468	0.871

Table 2: Example 2: Interval test results using censor defined intervals. (Note that Bonferroni rejected intervals are also counted as individual flags.)

Specified intervals (10 evenly-spaced)

There are many ways these intervals could be chosen, and several approaches are briefly mentioned in the discussion. For the sake of brevity, we will outline one of the simpler approaches. The data is split into 10 evenly-spaced intervals (owing to the classical convention of the Hosmer-Lemeshow test.) These intervals start at time 0 and end at the largest censor time. That is, for $k = 1, ..., 10$

$$V_k = \frac{1}{10} \cdot t.max \cdot (k-1,\ k].$$

To simplify code implementation, p-values were calculated using the sinib R package [16], which uses a saddle point approximation for the distribution of the sum of binomials. The results are presented in Table 3, and the interval test plot for generalised gamma is presented in Figure 2.

Parametric survival model	Bonferroni rejections	Individual flags	PAVSI	TFT
Exponential	1	4	0.0005	0.0040
Gamma	1	2	0.0488	0.0052
Generalised gamma	1	2	0.0488	0.0345
Gompertz	1	4	0.0005	0.0056
Log-logistic	1	1	0.2437	0.0631
Log-normal	0	1	0.2437	0.0699
Weibull	1	3	0.0063	0.0041

Table 3: Example 2: Interval test results using 10 evenly-spaced intervals.

As we can see, Table 3 tells a very different story to the previous interval selection approach. The overall PAVSI p-value is below 0.05 for all models except log-logistic and log-normal, indicating poor fit in 5 out of 7 models. Similarly, the overall TFT p-value rejects 5 out of 7 models and is below 0.07 in the final two models, which again raises concerns about poor fit. Finally, all of the parametric models have a Bonferroni rejection except log-normal, suggesting only the log-normal is a feasible model. We investigate the log-normal model further by presenting the results for each of its 10 intervals in Table 4.

From Table 4, we see that the interval resulting in an individual flag for the log-normal model was between times 30.35 and 34.69 with a p-value of 0.0087, which is above the 0.0025 threshold of a Bonferroni rejection. Furthermore, we note that the 4th interval had a p-value of 0.971. If this value was only 0.004 higher, then the log-normal would have received a second individual flag and thereby a PAVSI overall p-value rejection. Furthermore, the log-normal performed the worst of all 7 models for both AIC and BIC (not presented). For these reasons, as well as the overall TFT p-value of 0.0699

V_j	$\mathbb{E}$(Events)	Obs Events	p-mid	Individual test	Bonferroni test
(0, 4.34]	18.76	11	0.0289	Accept	Accept
(4.34, 8.67]	22.52	23	0.5534	Accept	Accept
(8.67, 13.01]	17.90	26	0.9677	Accept	Accept
(13.01, 17.34]	13.67	21	0.9710	Accept	Accept
(17.34, 21.68]	10.70	11	0.5549	Accept	Accept
(21.68, 26.01]	8.99	7	0.2550	Accept	Accept
(26.01, 30.35]	7.71	9	0.6947	Accept	Accept
(30.35, 34.69]	6.12	1	0.0087	Flag	Accept
(34.69, 39.02]	5.02	5	0.5247	Accept	Accept
(39.02, 43.36]	1.72	3	0.8279	Accept	Accept

Table 4: Example 2: Interval test for the log-normal model with 10 evenly-spaced intervals.

(close to 0.05) we can be justifiably suspicious of the log-normal model. However, we have no strong evidence to outright reject it.

In conclusion, due to the Bonferroni rejections, we would suggest that all presented models except for the log-normal are unfeasible. This leaves us with only the log-normal model remaining, which is somewhat restrictive and has also been flagged for concern as described above. For this reason, it is recommended that exploring flexible models such as splines and piecewise would be appropriate.

4 Simulation summary

A simulation experiment was performed to verify results under a true null hypothesis. Additional details are provided in Appendix C. The simulation experiment aimed to confirm our methodological assumptions with a simple but realistic example. For the approach where intervals were chosen using censoring, the test statistics had a very low type I error rate, suggesting that they are implausibly conservative in practice. For the 10 evenly-spaced interval approach, the test statistics performed as expected with a type I error rate close to 0.05. The Bonferroni test performed well in both approaches but was somewhat conservative for the censor approach. In conclusion, the two primary assumptions hold fairly well for the 10 evenly-spaced intervals approach. (The two primary assumptions are that censor intervals I_j are uninformative of the event generating process and that the mid-point p-values are approximately uniform[0,1].) We also note that varying the maturity of the data had only a minor impact on the evenly-spaced interval approach.

5 Discussion

In terms of applying the interval test, two approaches were presented. In the first, intervals of interest are defined entirely by the censors. In the second, intervals of interest were specified as 10 evenly-spaced intervals up to the final censor. There are strengths and weaknesses to both approaches. We address this in the discussion, where the following topics are covered: the censor interval approach, how to specify intervals, the benefits of the 10 evenly-spaced interval approach, testing with limited patient numbers, and a conclusion.

In our first approach, intervals are selected by the data using the censor times. Unfortunately, this approach is generally unhelpful in describing regions with heavy censoring. This is because very small intervals are unlikely to be rejected due to low power. Additionally, with this approach, our test statistics are very conservative in standard cases (see Appendix C). On the other hand, we argue that in some cases (not presented) this approach is indicative of informative or unusual censoring. That is, it is very apparent if a lot of censoring has occurred, and furthermore, we can see if an unusual number of events occurred within these intervals. As a final point, this approach underlies the specified interval

approach, so it will always provide helpful supporting evidence.

In the second approach, intervals can be specified. There are several ways this could be done. In our example, we provide one approach (10 evenly-spaced intervals). If submitting to a health technology assessment body, we recommend using this approach. This prevents intervals from being systematically chosen such that model selection favours a treatment's efficacy. However, in terms of further research, there are certainly additional options for how the intervals are chosen. One option is automatic interval selection based on equal statistical power. Another topic of consideration is restricting chosen intervals to have boundaries that are censor times.

There are several benefits to the 10 evenly-spaced intervals approach compared to the censor interval approach. The 10 intervals are generally easier to interpret with plots. Furthermore, the approach is informative for areas with heavy censoring (such as the tail of the KM curve), which is often the area of greatest importance for long-term extrapolation. Additionally, overall p-values perform much better for this method. It is for these reasons that in general we recommend the fixed interval approach. However, we acknowledge that the censor interval approach is still valid and has its own benefits.

In situations with limited data, say under 20 patients, the tests can still be used. However, they will likely be under-powered and simply accept the fitted model. This is not necessarily a bad thing. Visual inspection of curve fit can be challenging for such data. As a result of visual inspection alone, flexible survival models might be fitted to data with only 20 patients. That is, we have potentially over-fitted the data. The interval tests can prevent this by providing an objective evaluation. The tests will likely confirm that the observed data is feasible for simple models, which indicates that flexible models are not required.

In conclusion, the interval tests provide additional information that can support the model selection process. The tests are able to highlight areas of particularly poor fit. It is recommended that they are especially useful in aiding decisions when the curve fit is poor.

6 Appendix

A In defense of midpoint p-values

When deriving test statistics and results, we will often make the assumption that midpoint p-values are uniform[0,1] under H0. If the validity of this approach is in question, then randomized p-values can be used, which are known to be uniform[0,1] under H0[13]. For this reason, the randomized p-values are superior from this purely statistical standpoint. However, we argue below that the midpoint p-values are more practical and that for our applications the uniform[0,1] assumption is often reasonable.

In the case of reducing the number of intervals to say 10, as is described in example 2, the generating process is the sum of binomials. If the number of patients is reasonable, one would expect these intervals to contain a reasonable number of feasible discrete outcomes. This means the uniform[0,1] assumption is more valid, compared to when there are only a small number of possible outcomes.

We also acknowledge that in practice, randomized p-values are more difficult to deal with. When an analysis is rerun, the p-values are rerandomized. This would then involve rerunning the analysis to determine the level of variation in the randomised p-values and their resulting test statistic. This is helpful. However, this approach leads you down the path of using the expected value of these reruns as your final quoted value. That is, the midpoint p-value, which is the expected value of randomized p-value.

B Extension to individual patient models

In the introduction, we mentioned that the most commonly used calibration methods rely on risk scores [4]. However, in our approach, risk scores were not considered. Furthermore, in our examples, all of our models inherently assume equal risk scores for each patient. It is, however, possible to extend our methodology to models where the risk scores are not assumed constant. This would involve amending Equation 1 such that each individual is considered separately for each interval. That is, every patient is assigned a specific probability of experiencing an event (as defined by the model) for that interval. Subsequently, Equation 1 would be replaced by a collection of N_{I_j} Bernoulli random variables with different probabilities.

C Additional details on simulation

In practice, the parametric model never truly underlies the data, meaning type I errors don't make logical sense. However, it does provide an interesting thought-experiment; if the observed data **was** generated by its fitted model, is the observed data behaving as you would expect? If the answer is yes, then the model is reasonable. If the answer is no, then the model is likely unreasonable.

The primary purpose of this investigation by simulation was to confirm that the methodology behaves as expected with respect to type I errors. The realistic model was described by an event generating process $T \sim exp(\lambda)$ and a censor generating process for C given by

$$\begin{aligned} C1 &\sim uniform[0, 100], \\ C2 &\sim uniform[18, 22], \\ C &= min(C1, C2). \end{aligned}$$

Finally, $Time = min(C, T)$; and if $T \leq C$, $Event = 1$, else $Event = 0$. Taking the time to be in months, we are saying some patients are heavily censored in the last 4 months of the trial, patients have an expected event time of $1/\lambda$, and some patients are censored throughout the trial.

For the censored interval approach, test statistics were shown to be unrealistically conservative. This is due to an inflated number of interval p-values that are close to 0.5, which occurs due to using

the midpoint p-value approach. That is, most of the small intervals will yield a midpoint p-value of 0.5. These are then transformed for the TFT to give effective values of 1, which inflates the overall TFT p-value, pushing it closer to 1. In general, more censor intervals result in the tests being even more conservative. In our simulation, with 200 patients, $\lambda = 1/10$, and 10,000 iterations the PAVSI type I error rate was 0.0041 and for TFT this was 0.0001. This is not ideal as it indicates that in practice these test statistics would be overly conservative. Thereby, both test statistics are not recommended in practice when using the censor interval approach.

For the 10 evenly-spaced intervals approach, the test statistics behaved as expected. When the number of patients increased, the test statistics appeared to converge to the expected values. Even with small numbers of patients, the test statistics remained reasonable. For example, with 50 patients, $\lambda = 1/10$, and 10,000 iterations the PAVSI type I error rate was 0.0331 and for TFT this was 0.0146.

For the 10 evenly-spaced intervals approach with 200 patients, the PAVSI type I error rate was 0.071 and for TFT this was 0.04. Note that here the PAVSI type I error rate is above 0.05. This is as expected. In particular, when assuming the interval p-values are uniform[0,1], the type I error is 0.08614. (This result can be shown by considering our approach, which uses 10 intervals, a 0.05 rejection level, and a test statistic obtained by using the mid-point p-value.) As the number of patients increases, our PAVSI test statistic appears to converge towards this value, 0.08614. That is, the midpoint p-values are behaving as uniform[0,1] random variables. Similarly, the TFT appears to converge towards 0.05 as the number of patients increases.

We make a quick digression here to discuss the PAVSI type I error converging towards 0.08614. If we were to select a different number of intervals, say 7, then the PAVSI type I error would converge towards 0.0444 instead. However, this is not ideal. With lower patient numbers, this value decreases even further due to the midpoint p-values. Thereby the 7 interval approach would always be conservative, which we argue is a bad property for model selection. The 10 evenly-spaced interval method approaches a type 1 error of 0.08614 as the patients increase. However, a type I error slightly above 0.05 isn't necessarily a bad thing, especially in practice. Also, note that when using 10 intervals with a simulation of 100 patients, the PAVSI type I error rate was 0.0556, which is approximately what we want as it is close to 0.05.

The Bonferroni type I error rates were approximately 0.02 for the censored interval approach and 0.04 for the 10 evenly-spaced interval approach. These values appear to remain approximately stable as the number of patients changes. That is, approximately within $(0.03, 0.05)$ for the 10 evenly-spaced interval approach and within (0.017,0.022) for the censor approach, when considering 50 to 2000 patients with $\lambda = 1/10$.

Finally, we experimented with different levels of maturity in the data. In particular, we used $\lambda = 1/10$ (mature data), $\lambda = 1/30$ (somewhat mature), and $\lambda = 1/70$ (immature). For these simulated data, the KM curves finished at approximate survival levels of 0.11, 0.48, and 0.73 respectively. In the censored interval approach, when the data was less mature, the test statistics performed worse. (They became even more conservative.) This is due to an increased amount of censoring in the less mature data. For 10 evenly-spaced interval approach, the different levels of maturity in the data had a somewhat minimal impact on the type I error rates. In particular, when weakening the data from mature to immature ($\lambda = 1/10$ to $\lambda = 1/70$), this would approximately result in a decrease of around 0.01 in the type 1 error rate for the PAVSI, TFT and Bonferroni tests.

References

[1] Mark J Rutherford et al. (2020) *NICE DSU Technical Support Document 21: Flexible Methods for Survival Analysis.*

[2] Nicholas Latimer. (2011) *NICE DSU Technical Support Document 14: Survival Analysis for Economic Evaluations alongside Clinical Trials - Extrapolation with Patient-Level Data* .

[3] David W. Hosmer and Stanley Lemeshow. (2000) *Applied Logistic Regression 2nd edition*, Wiley series in probability and statistics: Section 5.2.2

[4] Olga V. Demler, Nina P. Paynter, and Nancy R. Cook. (2015) *Tests of Calibration and Goodness of Fit in the Survival Setting*, Statistics in Medicine, Vol.34: 1659–1680

[5] Grønnesby JK, Borgan Ø. (1996) *A method for checking regression models in survival analysis based on the risk score*, Lifetime Data Analysis. 2:315–320. 9

[6] D'Agostino RB and Byung-Ho N. (2004) *Evaluation of the performance of survival analysis models: discrimination and calibration measures*, Handbook of statistics. 2004; 23:1–25

[7] Peter C. Austin, Frank E. Harrell, and David van Klaveren. (2020) *Graphical calibration curves and the integrated calibration index (ICI) for survival models*, Statistics in Medicine, Vol.39: 2714–2742

[8] Allan Birnbaum (1954) *Combining Independent Tests of Significance,* Journal of the American Statistical Association, Vol 49: No 267 559-574

[9] Nicholas A. Heard and Patrick Rubin-Delanchy (2018) *Choosing Between Methods of Combining p-values,* Biometrika, Vol 105: 239-246

[10] Suzanne C Freeman et al. (2022) *Challenges of modelling approaches for network meta-analysis of time-to-event outcomes in the presence of non-proportional hazards to aid decision making: Application to a melanoma network*, Statistical Methods in Medical Research, Vol. 31(5) 839–861

[11] Axel Hauschild et al. (2020) *Long-term outcomes in patients with BRAF V600-mutant metastatic melanoma receiving dabrafenib monotherapy: Analysis from phase 2 and 3 clinical trials*, European Journal of Cancer, Vol. 125:114-120.

[12] Georgina Long et al. (...) *Dabrafenib plus trametinib versus dabrafenib monotherapy in patients with metastatic BRAF V600E/K-mutant melanoma: long-term survival and safety analysis of a phase 3 study*, Annal of Oncology, Vol. 28(7):1631-1639.

[13] Patrick Rubin-Delanchy, Nicholas A. Heard, and Daniel J. Lawson (2019) *Meta-Analysis of Mid-p-Values: Some New Results based on the Convex Order,* Journal of the American Statistical Association, Vol 114, No. 527: 1105–1112: Theory and Methods

[14] Richard A. Armstrong. (2014) *When to use the Bonferroni correction,* Ophthalmic and Physiological Optics, Vol 34: 502-508

[15] Christopher Jackson. (2016) *flexsurv: A Platform for Parametric Survival Modeling in R*, Journal of Statistical Software, Vol. 70 Issue 8

[16] Boxiang Liu and Thomas Quertermous. (2017) *Sinib: Approximating the Sum of Independent Non-Identical Binomial Random Variables,* arXiv