Metrika
https://doi.org/10.1007/s00184-021-00807-4

Imprecise inference based on the log-rank test for accelerated life testing

Frank P. A. Coolen[1] · Abdullah A. H. Ahmadini[2] · Tahani Coolen-Maturi[1]

Received: 9 April 2019 / Accepted: 8 January 2021

Abstract
This paper presents an imprecise predictive inference method for accelerated life testing. The method is largely nonparametric, with a basic parametric function to link different stress levels. The log-rank test is used to provide imprecision for the link function parameter, which in turn provides robustness in the resulting lower and upper survival functions for a future observation at the normal stress level. An application using data from the literature is presented, and simulations show the performance and robustness of the method. In case of model misspecification, robustness may be achieved at the price of large imprecision, which would emphasize the need for more data or further model assumptions.

Keywords Accelerated life testing · Arrhenius link function · Failure data · Imprecise probability · Log-rank test · Lower and upper survival functions · Nonparametric predictive inference · Right-censored data

1 Introduction

To determine the reliability of a new product in a relatively short period of time, accelerated life testing (ALT) can be used. In ALT, units are exposed to higher than normal stress levels (e.g. lightbulbs to a higher than normal voltage) to induce failures more rapidly. There are several typical designs for ALT, including constant-, step- and progressive-stress testing, for a detailed introduction to statistical methods for ALT see Nelson (1990). Due to the complex nature of ALT scenarios, the modelling for statistical inference based on ALT data provides many challenges. The aim of this paper is to provide a robust method of statistical inference which can be widely used in practical ALT applications. The method generates imprecise survival functions based

✉ Frank P. A. Coolen
frank.coolen@durham.ac.uk

[1] Department of Mathematical Sciences, Durham University, Durham, UK

[2] Department of Mathematics, Faculty of Science, Jazan University, Jazan, Saudi Arabia

Published online: 09 February 2021

on relatively few model assumptions, if these fail to provide clear insight into practical issues then one may need to make stronger model assumptions or collect more data.

Yin et al. (2017) introduced an imprecise statistical method for ALT data using the power-Weibull model. They first fitted a fully parametric model for all data, assuming Weibull failure time distributions at the different stress levels with the scale parameters linked through the power-law link function. Then they introduced imprecision in the power-law link function by considering an interval around the parameter estimate, leading to observations at stress levels other than the normal level to be transformed into intervals at the normal level. They applied nonparametric predictive inference (NPI) at the normal stress level, using the original data at that level combined with transformed data intervals from other levels. Yin et al. (2017) did not give an argument, other than simulation studies, for the amount of imprecision in the parameter of the link function. Building on the work by Yin et al. (2017), Ahmadini and Coolen (2020) used a parametric statistical test between pairwise stress levels, to determine the level of imprecision. They assumed Weibull distributions for all stress levels, with the scale parameters linked by the Arrhenius link function. They derived an interval for the parameter of the Arrhenius link function by pairwise likelihood ratio tests.

This paper continues the research started by Yin et al. (2017) and Ahmadini and Coolen (2020). Different from that work, we do not assume a failure time distribution at each stress level, only a parametric link function between the levels. We use the log-rank test to provide imprecision for the link function parameter. We obtain an interval for the parameter of the link function by pairwise hypothesis tests between the stress levels to determine the level of imprecision, based on the idea that, if data from a higher stress level are transformed to the normal stress level, then the transformed data and the original data from the normal stress level should be indistinguishable if the model fits well. The main novelty of the approach is that, by using a nonparametric test, we do not need to assume a parametric failure time distribution at each stress level. This makes the method more widely applicable than the method presented by Ahmadini and Coolen (2020).

This paper is organized as follows. Section 2 gives an overview of the main aspects of NPI. Section 3 presents the new imprecise predictive inference method for ALT data based on the log-rank test. Section 4 presents an example to illustrate our method using data from the literature. Section 5 presents the results of a simulation study to investigate the performance of the proposed method, both for the case where the assumed model fits the data well and where the model is not correct, the latter provides insight into robustness of our method. Section 6 provides some concluding remarks.

2 Nonparametric predictive inference

Nonparametric predictive inference (NPI) is a statistical method which provides lower and upper survival functions for a future observation based on past data using imprecise probabilities (Augustin et al. 2014; Coolen 2011a). Hill (1968) proposed the assumption $A_{(n)}$ which gives direct conditional probabilities for a future random quantity based on the observed values of related random qualities (Augustin and Coolen 2004; Coolen 2006, 2011a). It proposes that the rank of a future observation among the values already observed is equally likely to have each possible value $1, \dots, n+1$. Suppose that $X_1, X_2, \dots, X_n, X_{n+1}$ represent exchangeable and continuous real-valued possible random quantities. Let the ranked observed values of $X_1, X_2, \dots, X_n$ be denoted by $x_{(1)} < x_{(2)} < \cdots < x_{(n)}$, and let $x_{(0)} = 0$ and $x_{(n+1)} = \infty$. The assumption $A_{(n)}$ is

$$P(X_{n+1} \in (x_{(j-1)}, x_{(j)})) = 1/(n+1)$$

for all $j = 1, 2, \dots, n+1$. We assume that there are no tied observations for ease of presentation, any tied values can be dealt with by assuming that they differ by infinitesimally small amounts.

Statistical inference based on $A_{(n)}$ is nonparametric and predictive, and can be considered suitable if there is hardly any knowledge about the random quantity of interest, except for the n observations, or if one does not want to use any further information. $A_{(n)}$ is not sufficient to derive precise probabilities for many events of interest, but optimal bounds for probabilities can be derived through the ‘fundamental theorem of probability’ (De Finetti 1974), and these are lower and upper probabilities in imprecise probability theory (Augustin and Coolen 2004; Augustin et al. 2014).

The lower and upper probabilities for event A are denoted by $\underline{P}(A)$ and $\overline{P}(A)$, respectivily. These are open to interpretation in various ways (Augustin et al. 2014). $\underline{P}(A)$ can be regarded as the supremum buying price for a gamble on event A which pays 1 if A occurs and 0 if not. It can also be regarded as the maximum lower bound for the probability for A based on the assumptions made. $\overline{P}(A)$ can be regarded as the minimum selling price for the same gamble on A, or as the minimum upper bound based on the assumptions made. For lower and upper probabilities the logical relation $0 \leq \underline{P}(A) \leq \overline{P}(A) \leq 1$ holds, and under standard coherence conditions (Augustin et al. 2014) the conjugacy property $\underline{P}(A) = 1 - \overline{P}(A^c)$ holds, where A^c is the complementary event of A. These properties also hold for NPI for real-valued quantities (Augustin and Coolen 2004), hence also for the predictive inferences presented in this paper.

The NPI lower survival function for a future observation X_{n+1} is

$$\underline{S}_{X_{n+1}}(t) = \underline{P}(X_{n+1} > t) = \frac{n-j}{n+1}, \text{ for } t \in (x_{(j)}, x_{(j+1)}), \;\; j = 0, \dots, n \tag{1}$$

and the corresponding NPI upper survival function for X_{n+1} is

$$\overline{S}_{X_{n+1}}(t) = \overline{P}(X_{n+1} > t) = \frac{n+1-j}{n+1}, \text{ for } t \in (x_{(j)}, x_{(j+1)}), \;\; j = 0, \dots, n \tag{2}$$

In reliability and survival analysis interest is usually in failure events and data often include right-censored observations as some units are only known not to have failed before a specific time. Coolen and Yan (2004) presented a generalization of $A_{(n)}$, called rc-$A_{(n)}$, which enables NPI with right-censored data. Suppose that data from n units consists of u failure times $x_{(1)} < x_{(2)} < \cdots < x_{(u)}$ and $n-u$ observed right-censoring times $c_{(1)} < c_{(2)} < \cdots < c_{(n-u)}$, and set $x_{(0)} = 0$ and $x_{(u+1)} = \infty$. For ease of presentation we assume that there are no tied observations, any ties can be broken by adding infinitesimal amounts without noticeable effect on the resulting lower and upper survival functions. Let s_i denote the number of right-censored observations in the interval $(x_{(i)}, x_{(i+1)})$, denoted by $c^i_{(1)} < c^i_{(2)} < \cdots < c^i_{(s_i)}$, so $\sum_{i=0}^{u} s_i = n - u$. Let d^i_j denote the event time such that $d^i_0 = x_{(i)}$ and $d^i_j = c^i_{(j)}$ for $i = 1, 2, \ldots, u$ and $j = 1, 2, \ldots, s_i$, and let $d^i_{s_i+1} = d^{i+1}_0 = x_{i+1}$ for $i = 1, 2, \ldots, u-1$. Let $\tilde{n}_{c_u}$ and $\tilde{n}_{d^i_j}$ denote number of subjects in the risk set just before to time c_u and d^i_j, respectively, and let $\tilde{n}_0 = n + 1$. The risk set at a specific time consists of all units which have not failed or been censored prior to that time, hence for which the corresponding observed failure time or right-censoring time is greater than or equal to the specific time.

Coolen and Yan (2004) presented the NPI lower and upper survival functions for a future failure observation, denoted by $\underline{S}_{X_{n+1}}(t)$ and $\overline{S}_{X_{n+1}}(t)$, respectively, and they are available in the following product forms (Maturi 2010b; Maturi et al. 2010). For $t \in [d^i_j, d^i_{j+1})$, with $i = 1, 2, \ldots, u$ and $j = 1, 2, \ldots, s_i$, the NPI lower survival function is

$$\underline{S}_{X_{n+1}}(t) = \frac{1}{n+1}\tilde{n}_{d^i_j} \prod_{r: c_r \le d^i_j} \left(\frac{\tilde{n}_{c_r} + 1}{\tilde{n}_{c_r}} \right) \tag{3}$$

and for $t \in [x_{(i)}, x_{(i+1)})$, with $i = 1, 2, \ldots, u$, the NPI upper survival function is

$$\overline{S}_{X_{n+1}}(t) = \frac{1}{n+1}\tilde{n}_{x_i} \prod_{r: c_r \le x_i} \left(\frac{\tilde{n}_{c_r} + 1}{\tilde{n}_{c_r}} \right). \tag{4}$$

Note that $\underline{S}_{X_{n+1}}(t)$ decreases at each observation but $\overline{S}_{X_{n+1}}(t)$ decreases only at observed failure times. This illustrates the attractive informal interpretation that a lower probability for event A reflects the information in support of event A, while the upper probability (actually, $1 - \overline{P}(A)$) reflects the information against event A, so in support of the complementary event. A failure observation provides information against survival so leads to both $\underline{S}_{X_{n+1}}(t)$ and $\overline{S}_{X_{n+1}}(t)$ decreasing. A right-censored observation reduces information in support of survival past that point but it does not provides information against survival as the unit did not fail, hence $\underline{S}_{X_{n+1}}(t)$ decreases at such an observation but $\overline{S}_{X_{n+1}}(t)$ does not.

3 Imprecise predictive inference based on the log-rank test

In this section, we present a new semi-parametric statistical method for predictive inference based on constant stress ALT data. The method is similar to the method introduced in Ahmadini and Coolen (2020), but in this paper we do not assume a parametric failure time distribution at each stress level. As in Ahmadini and Coolen (2020), we assume a parametric model as a link between the different stress levels. While the new method can be used with a range of such parametric link functions, we only consider the Arrhenius model for ALT to link the different stress levels in this paper. This model is mainly used in situations where the failure mechanism is driven by temperature, it has been applied to various problems in engineering (Meeker and Escobar 1998; Nelson 1990). According to this model, an observation t^i at the stress level i, subject to stress K_i, can be transformed to an observation at the normal stress level K_0, by the equation

$$t^{i \to 0} = t^i \exp\left(\frac{\gamma}{K_0} - \frac{\gamma}{K_i}\right), \tag{5}$$

where K_i is a positive value representing stress level i (in scenarios where this is temperature, it is measured in Kelvin) and γ is the parameter of the Arrhenius model. This transformation is applied to all data, so failure times and right-censored observations.

The central idea behind the new method is as follows. If the Arrhenius model is suitable for linking the different stress levels, then the observations transformed from the increased stress levels to the normal stress level should not be distinguishable, so they should be reasonably mixed. We consider this by performing pairwise tests between the original data at level 0 and transformed data per level i, testing the hypothesis that both data sets come from a common underlying distribution. As we do not assume a parametric failure time distribution at each stress level, these pairwise tests have to be nonparametric. Because ALT data can include right-censored observations, we apply the well-known log-rank test (Mantel 1966; Peto and Peto 1972), which is also known as the Mantel-Cox test (Gehan 1965; Mantel 1967a).

The proposed method in this paper consists of two steps. First, the log-rank test is used to test if the data transformed from level i to level 0, and the original data at level 0, may come from the same underlying distribution. This hypothesis test is performed with significance level α, in line with statistical tradition typically set at 0.01, 0.05 or 0.1. Instead of performing this test with a fixed value for the parameter γ of the Arrhenius link function, we derive the interval $[\underline{\gamma}_i, \overline{\gamma}_i]$ of values for γ for which we do not reject the null hypothesis. Note that the fact that the set of such values for γ is indeed an interval follows from a monotonicity property of the log-rank test proven by Coolen-Maturi and Coolen (2020b). Note further that $\underline{\gamma}_i$ and $\overline{\gamma}_i$ are increasing and decreasing functions of α, respectively, so the log-rank test with $\alpha = 0.05$ leads to a larger interval $[\underline{\gamma}_i, \overline{\gamma}_i]$, and hence to more imprecision in our method, than testing with $\alpha = 0.10$. This procedure is applied for each stress level $i = 1, \ldots, m$, and with the m pairs $(\underline{\gamma}_i, \overline{\gamma}_i)$ we define $\underline{\gamma}$ = max {min $\underline{\gamma}_i$, 0} and $\overline{\gamma}$ = max $\overline{\gamma}_i$. Because of the physical interpretation of failures generally occurring faster at increased stress levels, we exclude negative values which leads to some $\underline{\gamma}$ values being set at 0. We compute the $\underline{\gamma}_i$ and $\overline{\gamma}_i$ numerically using the statistical software R, in particular the function *survdiff* from the *survival* package to apply the log-rank test, and a basic search to find $\underline{\gamma}_i$ and $\overline{\gamma}_i$ with 3 decimals accuracy. For our entire procedure, very small changes in these $\underline{\gamma}_i$ and $\overline{\gamma}_i$ beyond the third decimal have extremely small effects on the resulting lower and upper predictive survival functions.

The second step of our method is as follows. Each observation at an increased stress level is transformed to an interval at level 0, by applying the transformation (5) with $\underline{\gamma}$ and with $\overline{\gamma}$, leading to the interval $[\underline{t}^{i\to 0}, \overline{t}^{i\to 0}]$ with

$$\underline{t}^{i\to 0} = t^i \exp\left(\frac{\underline{\gamma}}{K_0} - \frac{\underline{\gamma}}{K_i}\right)$$
$$\overline{t}^{i\to 0} = t^i \exp\left(\frac{\overline{\gamma}}{K_0} - \frac{\overline{\gamma}}{K_i}\right)$$

Then the NPI lower and upper survival functions, $\underline{S}$ and $\overline{S}$, are derived as described in Sect. 2, for a future observation at stress level 0. Crucially, $\underline{S}$ is based on the original data at level 0 combined with the values $\underline{t}^{i\to 0}$ for the transformed observations from higher stress levels, and $\overline{S}$ is based on the original data at level 0 combined with the values $\overline{t}^{i\to 0}$ for the transformed observations from higher stress levels. This leads to more imprecision than if only precisely observed data were used, which reflects uncertainty about the parameter γ in the Arrhenius link function. However, the imprecision also reflects possible problems with model fit, as we explain next.

If the model fits the data well, and there is a reasonable amount of data, we expect that most cases lead to $\underline{\gamma} = \underline{\gamma}_1$ and $\overline{\gamma} = \overline{\gamma}_1$, because a level 1 observation transferred using an interval of values for γ tends to lead to a smaller interval at level 0 than observations transferred from levels $i \geq 2$ using the same interval of values for γ, if the transferred intervals end up close to each other at level 0. Therefore, the pairwise log-rank tests will lead to more values for γ not being rejected for the pairwise test between levels 1 and 0, than for the tests between levels $i \geq 2$ and 0.

Crucially for our method, it is different if the model does not fit the data well. In this case, the intervals $[\underline{\gamma}_i, \overline{\gamma}_i]$ for different i will vary more, with less or even no overlap in extreme cases. This leads to the interval $[\underline{\gamma}, \overline{\gamma}]$ being much wider than is the case for good model fit. So, if a model is chosen for the link function which does not fit the data well, the lower and upper survival functions resulting from our method will have more imprecision than if the model fits well, which is an attractive feature of the method. It is important to emphasize that this effect is achieved because of the use of the pairwise hypothesis tests, and the conservative combination of the resulting intervals for γ. If, alternatively, we had performed a similar test for the data transformed from all stress levels simultaneously, which can be done by a generalization of the log-rank test (Collett 2015), then a poor model fit would lead to a small interval of values for γ for which the null-hypothesis would not be rejected, or even an empty interval. This would lead to less imprecision in the resulting inferences and would therefore not correctly reflect the poor model fit.

In Sect. 4 we present an example of application of our method using data from the literature, and we consider variations to the data due to right censoring. Section 5 presents a performance evaluation for our method based on simulations. For further discussion and more details of our method, including more examples which highlight more features of our method, comparison with the alternative approach where a parametric distribution is assumed at each stress level (Ahmadini and Coolen 2020), and discussion of computational aspects, we refer to the PhD thesis by Ahmadini (2019, Ch.4).

Table 1 Failure times at three temperature levels

$K_0 = 393$	$K_1 = 408$	$K_2 = 423$
3850	3300	2750
4340(+)	3720	3100
4760(+)	4080(+)	3400
5320(+)	4560	3800
5740	4920	4100
6160	5280	4400
6580	5640	4700
7140(*)	6120	5100
7980(*)	6840(*)	5700
8960(*)	7680(*)	6400

Table 2 $[\underline{\gamma}_i, \overline{\gamma}_i]$ and $[\underline{\gamma}, \overline{\gamma}]$, data Table 1

α	0.01	0.05	0.10
i	$[\underline{\gamma}_i, \overline{\gamma}_i]$	$[\underline{\gamma}_i, \overline{\gamma}_i]$	$[\underline{\gamma}_i, \overline{\gamma}_i]$
1	$[-1874.191, 5169.809]$	$[-1108.280, 4403.899]$	$[-624.387, 3920.005]$
2	$[38.751, 3690.236]$	$[435.786, 3293.202]$	$[686.627, 3042.360]$
	$[\underline{\gamma}, \overline{\gamma}]$	$[\underline{\gamma}, \overline{\gamma}]$	$[\underline{\gamma}, \overline{\gamma}]$
	$[0, 5169.809]$	$[0, 4403.899]$	$[0, 3920.005]$

4 Example

In this example, we apply the presented method to a data set from the literature (Vassilious and Mettas 2001), resulting from an accelerated life test with temperature as the accelerated factor. The failure data were collected at the normal temperature $K_0 = 393$ Kelvin and at $K_1 = 408$ and $K_2 = 423$. Ten units were tested at each temperature, the 30 failure times (in hours) are given in Table 1, for now neglect the $(+)$ and $(*)$. The intervals of the values for γ resulting from our method are given Table 2. The resulting lower and upper survival functions for a future observation at the normal temperature are presented in Fig. 1, indicated as $\underline{S}(\gamma)$ and $\overline{S}(\gamma)$, respectively, with the specific corresponding value for γ. At all three levels of significance α, $\underline{\gamma}_1 < 0$ hence $\underline{\gamma} = 0$ and the lower survival functions for the three α values are all equal. The upper survival functions differ, with imprecision increasing for decreasing α. These lower and upper survival functions can be used for further inference on the random failure time for a future item at the normal stress level.

We now illustrate the effect of censored observations in the data by changing some of the observations to right-censored observations. We consider three cases, represented in Table 1. In Case 1, we assume that the 4 observations indicated by $(+)$ in Table 1 are instead right-censored observations at 4000. In Case 2, we assume that the 5 observations indicated by $(*)$ in Table 1 are instead right-censored observations at 6600. In Case 3, all the right-censored observations of Cases 1 and 2 occur.

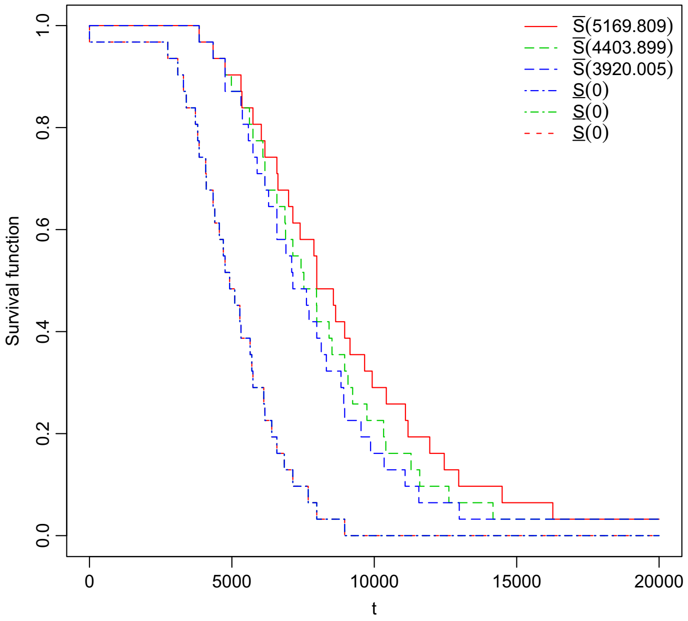

Fig. 1 Lower and upper survival functions

The intervals $[\underline{\gamma}_i, \overline{\gamma}_i]$ for Cases 1, 2, and 3 are given in Table 3. For all these cases we have $\overline{\gamma} = \overline{\gamma}_1$ and $\underline{\gamma} = 0$, except for Case 1 with $\alpha = 0.10$ where $\underline{\gamma} = 409.614$. The resulting lower and upper survival functions for $\alpha = 0.01$ are presented in Fig. 2. Clearly, more right-censored observations tend to lead to more imprecision, but there is an interesting detail in Fig. 2, namely it shows that the upper survival function for Case 1 is mostly being below the upper survival function for Case 2, but for some small intervals of values for t. This can happen if the order of an observed event and a right-censored observation, from two different stress levels, is different under two different transformations. It is important if a right-censored observation is before or after a fully observed event time, because the probability mass that is divided at a right-censoring time among the intervals to the right of it, depends on the number of observations to the right of the right-censoring time. If there are fewer observations to the right of the right-censoring time, the intervals between them all get a bit more probability mass according to the NPI approach.

5 Simulation study

In this section, we present results of a simulation study to investigate the performance of the imprecise predictive inference method for ALT data proposed in this paper. We consider two cases: Case A with perfect model fit, while in Case B the model does not fit perfectly. Case A serves to illustrate the performance of the method under ideal

Table 3 $[\underline{\gamma}_i, \overline{\gamma}_i]$ for 3 right-censoring cases

	α	0.01	0.05	0.10
	i	$[\underline{\gamma}_i, \overline{\gamma}_i]$	$[\underline{\gamma}_i, \overline{\gamma}_i]$	$[\underline{\gamma}_i, \overline{\gamma}_i]$
Case 1	1	[−1119.318, 5982.068]	[−353.408, 4948.068]	[409.614, 4636.459]
	2	[606.301, 4332.095]	[1200.640, 3805.070]	[1222.635, 3575.298]
Case 2	1	[−3673.884, 8158.461]	[−2357.513, 6408.005]	[−1652.449, 5653.132]
	2	[38.751, 5239.500]	[435.786, 4332.095]	[795.557, 3940.782]
Case 3	1	[−2357.513, 10677.282]	[−1119.319, 7377.025]	[−414.253, 7220.264]
	2	[686.627, 6545.213]	[1352.626, 5287.021]	[1864.493, 4834.417]

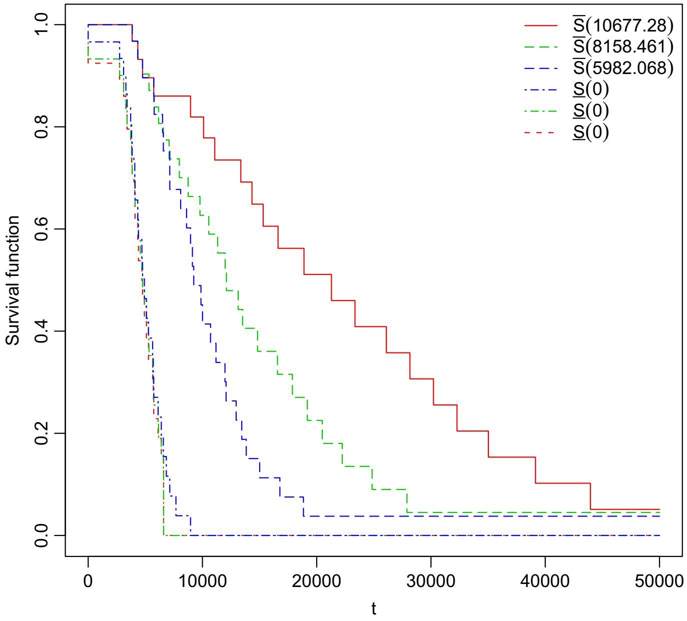

Fig. 2 Lower and upper survival functions for Cases: 1-blue, 2-green, 3-red

circumstances, with different numbers of data. Cases B illustrates the robustness of the method. Each case consists of 10,000 simulated data sets.

In *Case A*, we set three temperature stress levels $K_0 = 283$, $K_1 = 313$, and $K_2 = 353$. We generated random samples from the Arrhenius-Weibull model, which for stress level i with temperature K_i, has the Weibull survival function

$$S(t) = \exp[-(\frac{t}{\phi_i})^{\beta}]$$

Table 4 Proportion of runs with future observation greater than quartiles, Case A

α	q	$n = 10$		$n = 50$		$n = 100$	
		qL	qU	qL	qU	qL	qU
0.01	0.25	0.9498	0.4857	0.8565	0.6251	0.8302	0.6696
	0.50	0.8415	0.1237	0.6756	0.3192	0.6358	0.3673
	0.75	0.5917	0.0113	0.4344	0.0869	0.3840	0.1244
0.05	0.25	0.9197	0.5360	0.8363	0.6539	0.8156	0.6919
	0.50	0.7777	0.1939	0.6374	0.3596	0.6058	0.3988
	0.75	0.5323	0.0306	0.4006	0.1189	0.3528	0.1511
0.10	0.25	0.9036	0.5650	0.8239	0.6692	0.8075	0.7007
	0.50	0.7437	0.2327	0.6205	0.3811	0.5916	0.4130
	0.75	0.4960	0.0483	0.3815	0.1356	0.3387	0.1639

and the scale parameters are linked by the Arrhenius link function

$$\phi_i = \phi_0 \exp\left(\frac{\gamma}{K_i} - \frac{\gamma}{K_0}\right)$$

We set $\phi_0 = 7000, \beta = 3, \gamma = 5200$. At each stress level we simulated $n = 10, 50, 100$ observations from this model. Then we applied our method to derive the intervals $[\underline{\gamma}, \overline{\gamma}]$ for each data set and for significance levels $\alpha = 0.01, 0.05, 0.10$. With these intervals we derived the lower and upper survival functions for a future observation at the normal stress level.

To evaluate the performance of our method, for each simulated data set we also simulated one future observation at the normal stress level (with K_0), and we compared this to the quartiles of the lower and upper survival functions. For our method to perform well, the future observations should exceed the first, second, and third quartiles of the lower survival functions in proportions of the runs exceeding 0.75, 0.50 and 0.25, respectively, and also in proportions less than these values for the upper survival functions. Of course, the closer these proportions are to the respective values 0.75, 0.50, 0.25, the better the method performs.

The results for Case A are presented in Table 4, where under qL and qU these proportions are given according to the lower and upper survival functions, respectively, and the first, second and third quartiles are indicated by $q = 0.25, 0.50, 0.75$, respectively. The proportions of runs in which the future observation exceeds quartiles are as they should be, as explained above, but clearly there is a lot of imprecision in the method which is shown by the difference between corresponding qL and qU entries. The imprecision clearly increases if the significance level decreases, and there is substantial decrease of imprecision for increasing sample size n. For more details on this simulation we refer to the PhD thesis by Ahmadini (2019), where also detailed comparison of the intervals $[\underline{\gamma}_i, \underline{\gamma}_i]$, for $i = 1, 2$, is presented.

An important aim of our new method, in line with our related papers (Ahmadini and Coolen 2020; Yin et al. 2017) is to develop a quite straightforward method of predictive inference for ALT data based on few assumptions, where imprecision in

Table 5 Proportion of runs with future observation greater than quartiles, Case B

α	q	$n = 10$		$n = 50$		$n = 100$	
		qL	qU	qL	qU	qL	qU
0.01	0.25	0.9742	0.4957	0.9197	0.6443	0.9000	0.6754
	0.50	0.8663	0.1446	0.7669	0.3438	0.7404	0.3690
	0.75	0.5844	0.0145	0.4711	0.1058	0.4473	0.1273
0.05	0.25	0.9563	0.5554	0.9018	0.6612	0.8864	0.6871
	0.50	0.8248	0.2219	0.7373	0.3687	0.7189	0.3871
	0.75	0.5390	0.0389	0.4474	0.1267	0.4275	0.1399
0.10	0.25	0.9446	0.5802	0.8938	0.6691	0.8800	0.6928
	0.50	0.7999	0.2599	0.7220	0.3802	0.7067	0.3965
	0.75	0.5129	0.0574	0.4367	0.1344	0.4176	0.1460

the link function between different stress levels provides robustness against the model assumptions. To investigate robustness, in *Case B* we simulated data from the same setting as in Case A, but all data from stress level 1, with $K_1 = 313$, are multiplied by 1.2. The results of this simulation are presented in Table 5. The proportions of the runs in which the future observation exceeds the respective quantiles are still all in line with the requirement for the method to work well, as explained above. Compared to Case A, the entries in the table for corresponding qL and qU now indicate increased imprecision, which results from the fact that the intervals $[\underline{\gamma}, \overline{\gamma}]$ tend to become wider due to the model not fitting the data well anymore. This reflects the necessary pay-off for a robust method in case of problems with model fit, namely the resulting inferences are still meaningful but they become more imprecise. In practical applications, if the imprecision is too large to draw the inference of interest, or to make the required decision, then this indicates the need for more data to be collected or stronger model assumptions to be made. More detailed investigation for this case is included in the PhD thesis of Ahmadini (2019), where also other scenarios of model misspecification are investigated, including data simulated from ALT models with different link functions. For all simulations the main conclusion is the same, namely that the robustness works well but the price of severe model misspecification is high imprecision.

6 Concluding remarks

In this paper we presented a novel statistical method of imprecise parametric inference for ALT data, which does not require a parametric distribution to be assumed for each stress level, and uses a basic parametric link function between the different stress levels. The method combines the log-rank test for pairwise comparison of the survival distributions at different stress levels with the Arrhenius model as link between the stress levels. Transformation of observations at increased stress levels result in interval observations at the normal stress level, reflecting doubt about the quality of the basic assumed link function. By using pairwise log-rank tests it is assured that imprecision increases for worse model fit. The pairwise comparisons are between each increased stress level and the normal stress level, which reflects that data from the normal stress level are considered to be the main basis for the statistical inference, as interest is explicitly in a future observation at the normal stress level. We have assumed that failure data are available at all stress levels including the normal stress level, which may not be realistic. If there are no, or very few, failure data at the normal stress level, or only right-censored observations, then the method can be applied similarly but using a higher stress level as the basis for the combinations, so data from higher stress levels can be transformed to that stress level. Then the combined data at that level could be transformed all together to the normal stress level.

The log-rank test in this approach could be replaced by other comparison tests, this is left as an interesting topic for future research. The method can also be developed for different ALT scenarios, e.g. stepwise increased stress, as long as transformations of observations from higher stress levels to the normal stress level are possible. For an example application of our method to investigate basic warranty contracts we refer to the PhD thesis of Ahmadini (2019).

Acknowledgements We thank two reviewers for very detailed comments on an earlier version of this paper, which led to substantial improvements in presentation. Abdullah A.H. Ahmadini gratefully acknowledges the financial support from Jazan University in Saudi Arabia and the Saudi Arabian Cultural Bureau (SACB) in London for pursuing his PhD at Durham University.

Compliance with ethical standards

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

Ahmadini AAH (2019) Imprecise statistical methods for accelerated life testing. Ph.D. Thesis, Durham University. http://www.npi-statistics.com

Ahmadini AAH, Coolen FPA (2020) Statistical inference for the Arrhenius-Weibull accelerated life testing model with imprecision based on the likelihood ratio test. J Risk Reliabil 234:275–289

Augustin T, Coolen FPA (2004) Nonparametric predictive inference and interval probability. J Stati Plan Inference 124:251–272

Augustin T, Coolen FPA, de Cooman G, Troffaes MCM (2014) Introduction to imprecise probabilities. Wiley, Chichester

Collett D (2015) Modelling survival data in medical research, 3rd edn. Chapman and Hall/CRC, Boca Raton

Coolen FPA (2006) On nonparametric predictive inference and objective Bayesianism. J Logic Lang Inform 15:21–47

Coolen FPA (2011) Nonparametric predictive inference. In: International encyclopedia of statistical science. Springer, Berlin, pp 968–970

Coolen FPA, Yan KJ (2004) Nonparametric predictive inference with right-censored data. J Stat Plan Inference 126:25–54

Coolen-Maturi T, Coolen FPA (2020) A monotonicity property of weighted log-rank tests. arXiv:2008.00848v1 [stat.ME]

De Finetti B (1974) Theory of probability. Wiley, Chichester

Gehan E (1965) A generalized Wilcoxon test for comparing arbitrarily singly-censored samples. Biometrika 52:203–224

Hill BM (1968) Posterior distribution of percentiles: Bayes' theorem for sampling from a population. J Am Stat Assoc 63:677–691

Mantel N (1966) Evaluation of survival data and two new rank order statistics arising in its consideration. Cancer Chemother Rep 50:163–170

Mantel N (1967) Ranking procedures for arbitrarily restricted observation. Biometrics 8:65–78

Maturi TA (2010) Nonparametric predictive inference for multiple comparisons. Ph.D. Thesis, Durham University. http://www.npi-statistics.com

Maturi TA, Coolen-Schrijner P, Coolen FPA (2010) Nonparametric predictive inference for competing risks. J Risk Reliabil 224:11–26

Meeker WQ, Escobar LA (1998) Statistical methods for reliability data. Wiley, New York

Nelson WB (1990) Accelerated testing: statistical models, test plans, and data analysis. Wiley, Hoboken, New Jersey

Peto R, Peto J (1972) Asymptotically efficient rank invariant test procedures. J R Stat Soc Ser A 135:185–207

Vassilious A, Mettas A (2001) Understanding accelerated life-testing analysis. In: Annual reliability and maintainability symposium, Tutorial Notes. Citeseer, pp 1–21

Yin Y, Coolen FPA, Coolen-Maturi TA (2017) An imprecise statistical method for accelerated life testing using the power-Weibull model. Reliabil Eng Syst Saf 167:158–167