Performance of AIC and BIC for the extrapolation of survival data with different levels of censoring

Bütepage G, Vitor C, Carlqvist P[1]
Correspondence: greta.butepage@nordicmarketaccess.com
[1]Nordic Market Access, 113 59 Stockholm, Sweden

MSR71

Background

Survival is a key parameter in health economic assessments as it drives costs and benefits in the analyses. Due to limited follow-up of clinical trials, survival extrapolation is commonly needed [1]. The choice of survival model for extrapolation determines the outcome of the cost-effectiveness analyses [1].

The selection of the survival model is partly guided by goodness-of-fit statistics, namely the Akaike information criteria (AIC) and the Bayesian information criteria (BIC) [1]. Individual AIC and BIC values are not interpretable as they contain arbitrary constants and sample size. Rather, AIC/BIC are calculated for every candidate model and the "best" fitting model is the model with the smallest value. To guide the model selection, delta AIC (Δi = AICi – AICmin) and BIC values (Δi = BICi – BICmin) may be used, where AICmin and BICmin are the smallest values. Survival models with $\Delta i < 2$ are considered to be supported by the data and models with $\Delta i > 2$ may be considered not to be supported [2].

The use of goodness-of-fit statistics has some limitations [1]. For example, Beca et al. demonstrated, using the exponential distribution, short follow-ups of large samples produced a large error in AIC and BIC estimates [3].

The objective of this simulation study was to assess the performance of AIC and BIC when selecting between six standard parametric survival models to extrapolate survival data with varying levels of right censoring.

Methods

Survival data was simulated from six survival models (exponential, generalized gamma, Weibull, Gompertz, lognormal, and loglogistic), using the R programming language. Censoring of patients was simulated using a uniform distribution. To achieve different levels of censoring, the upper bound of the uniform distribution was increased gradually until 70% of events were censored. The level of censoring was defined as low (≈30%) to high (≈70%). The six models were fit to each of the six simulated data sets and the performance of AIC and BIC to identify the true survival model was assessed. The analysis was repeated 1000 times for each censoring level. The probability of the true survival model being the best fitting model was calculated using goodness-of-fit statistics. To assess the selection criteria of ΔAIC/BIC < 2, the number of times the correct model would have been rejected was estimated.

Results

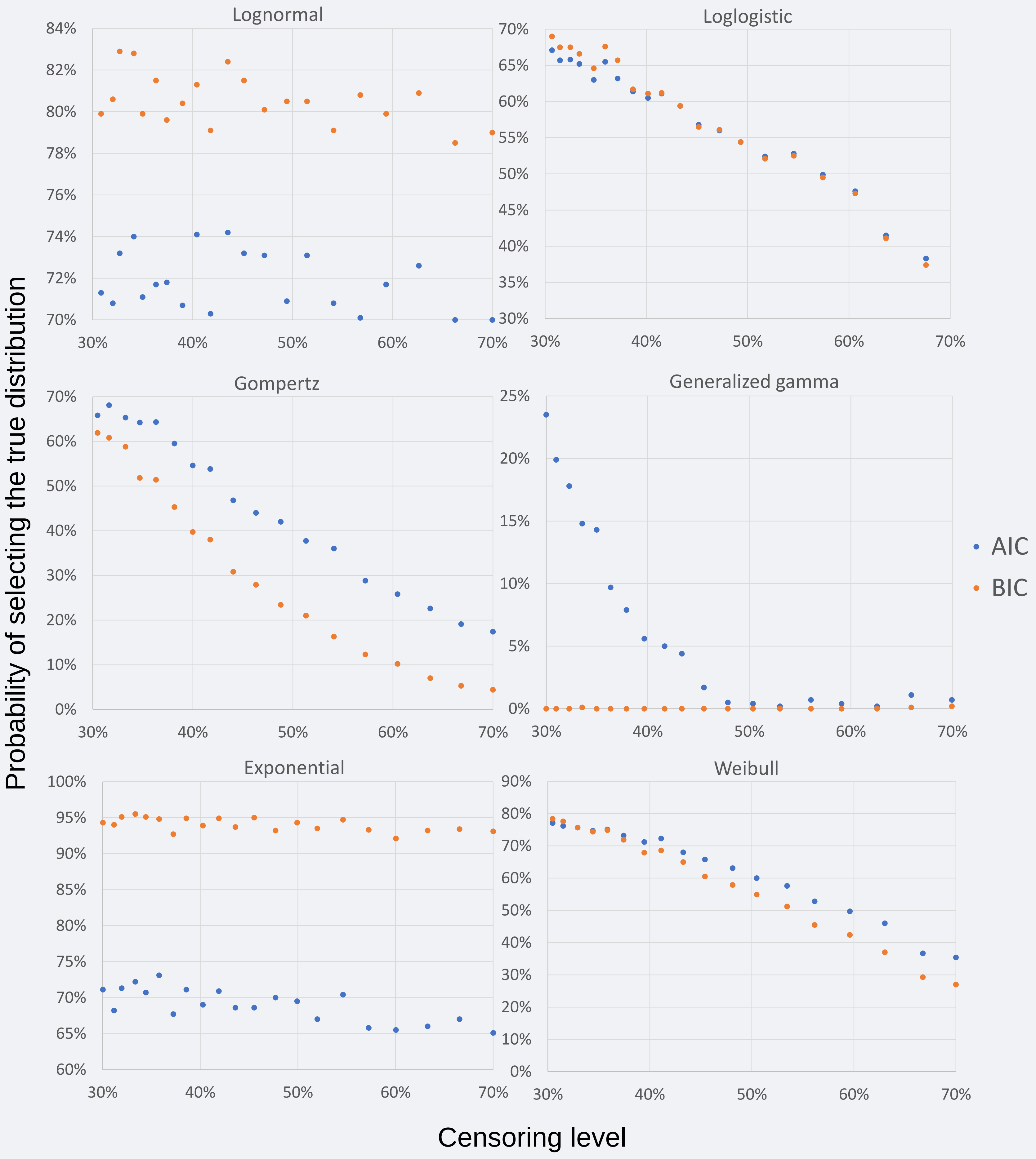
With increased censoring the probability of selecting the true survival model based on goodness-of-fit statistics decreased for the loglogistic, Gompertz, generalized gamma, and Weibull models. This was true for both AIC and BIC except for the generalized gamma model where BIC performed poorly at all censoring levels (Figure 1).

AIC typically performed better than BIC at all censoring levels. The declining trend in probability of selecting the true model was not observed for the lognormal and exponential models, and notably BIC performed better for these models.

Except for the generalized gamma model, the probability of selecting the true model was approximately 70 – 90% at a censoring level of 30% and dropped to 40% or lower at a censoring level of 70%. Only for the exponential and lognormal survival models, the probability remained at the same level, irrespective of censoring level.

AIC and BIC performed poorly when fitting the generalized gamma model (the probability of selecting this model was 0% at a censoring level of 50% or higher).

Figure 1. Probability of AIC and BIC selecting the true distribution by censoring level for six survival models



Across all censoring levels, ΔAIC < 2 was observed for all the true models; whereas ΔBIC < 2 was only observed for the lognormal, loglogistic, exponential and Weibull models, irrespective of censoring level. For the Gompertz and the generalized gamma models, ΔBIC > 2 was observed in 53% and 100% of simulations, respectively.

Conclusions

Overall, goodness-of-fit statistics AIC and BIC are sensitive to censoring and should be interpreted with caution when selecting a survival model for extrapolation. In most simulations, ΔAIC/BIC supported the use of the true distribution. Using ΔAIC as guidance may be appropriate irrespective of censoring while ΔBIC should be used with more caution.

The exponential and the lognormal models were the least sensitive to high censoring. In contrast, the generalized gamma and Gompertz models showed the highest sensitivity. The good and the poor performance of BIC for the exponential and the generalized gamma models, respectively, was expected, as the use of additional parameters (one and three parameters, respectively) is penalised more highly by the BIC than by the AIC.

Our results have shown that, at high levels of censoring, neither AIC nor BIC can guide the choice of a survival model. At 70% or more censoring, weight should be given to other selection criteria such as the clinical plausibility or visual fit rather than AIC and BIC. Distributions within Δi of two should be considered, as well as the distribution with the minimum AIC or BIC.

References

1. Latimer N, *NICE DSU Technical Support Document 14: Survival analysis for economic evaluations alongside clinical trials*. 2011.
2. Burnham, K.P. and D.R. Anderson, *Multimodel Inference: Understanding AIC and BIC in Model Selection.* Sociological Methods & Research, 2004. **33**(2): p. 261-304.
3. Beca, J.M., et al., *Impact of limited sample size and follow-up on single event survival extrapolation for health technology assessment: a simulation study.* BMC Medical Research Methodology, 2021. **21**(1): p. 282.