Proposal for a Summer Intern to Work on

Compiling and Evaluating More Efficient Statistical Testing Procedure for Non-Nested Model Selection for Survival Extrapolation and Event Projection

Supervisor: Gregory Chen, Ruifeng Xu, Geoffrey Johnson, Yujie Zhao, Larry Leon

In the cost-effectiveness analysis of health technology assessment (HTA), it is a complex process when conducting a scientifically sound survival extrapolation, since it contains a mixture of qualitative and quantitative steps with case-specific inputs from experts and data from different sources. Fitting several models and selecting plausible models are indispensable parts. In practice, one usually relies heavily on visual assessment between KM and the fitted curve and the ranking of AIC/BIC. Visual assessment is a reasonable first step but has subjectivity in assessment outcome, and usually the inflated uncertainty in the tail of the KM curve (due to a smaller risk set) is not quantitatively accounted for. The ranking of AIC/BIC seeks the best model which is not necessary in all cases, and overlooks estimation uncertainty (e.g. small differences among top performers might not be statistically nor practically meaningful).

A similar need for survival extrapolation and struggle in selecting plausible models also exist outside of the HTA space. For example, event projection during Phase III to make an educated guess on the timing of interim analyses.

Our aim is to formulate proper test-based model selection procedures to supplement or replace absolute ranking of goodness-of-fit statistics (AIC/BIC), for better balancing the goodness-of-fit and the parsimony of models, while accounting for sample uncertainty. With proper interpretation, test-based procedures do not always lead to a favorite model (e.g. all models could be observationally equivalent). When a definite outcome is needed (e.g. to choose the model for the main analysis), the definition of a pathway (a sequence of tests) can be made, where users' model preference (e.g. standard over flexible, simple over complex) can be incorporated.

The key challenge in the way (also recognized by NICE TSD 14) is that most pairs of interested parametric models are not nested one way or the other (e.g. Royston-Parmar spline model with 2 knots is not nested in its 3-knot setup even when the same 2 knots are included). In such a case, the usual log-likelihood or Wald statistic is no longer centered at zero, i.e. the conventional likelihood ratio test procedure is not applicable.

Initial research has been done and a parametric bootstrapping-based solution has been proposed in our research work in 2023 and it is under evaluation via simulation for their statistical properties. However, further work needs to be done to make it more well-understood and pragmatically useable in our daily business. **This intern will** conduct further literature research to find a more efficient bootstrapping procedure (e.g. via identifying appropriate pivot statistics to replace the vanilla likelihood ratio statistics for bootstrapping), evaluate the impact of different ways to formulate bootstrapped datasets to give the null model fitted based on the full data (e.g. setting of follow up time and necessity to mimic drop-out pattern when generating a bootstrapped dataset from the fitted parametric model), and lastly identify relevant performance metrics and compare test-based model selection procedure and the AIC/BIC ranking based on selected cases from literature or past NICE technical appraisals, or simulation.

The research will benefit HEDS modelers to develop a statistically sound evidence-based procedure to perform model selection for their base-case cost-effectiveness analysis, and interim analysis planning.