MSD
INVENTING FOR LIFE

On selecting a parametric model to predict long-term survival

to support health technology assessment (HTA)

X. Gregory Chen, Sajjad Rafiq

Biostatistics and Research Decision Sciences (BARDS)

MSD

at 6th Conference of the Central European Network (CEN)
Sept 6, 2023

Declaration

This presentation only reflects our personal views and/or understandings, and does not represent the opinion of our employer.

Outline

1. Brief Introduction
   about Survival Extrapolation in the HTA space

2. Prediction-focused Performance Metric
   with and without external evidence

3. Model Selection Procedures
   based on (a sequence of) statistical tests

4. Ongoing work

Brief Introduction

about Survival Extrapolation in HTA space

Long-term Parametric Survival Extrapolation in an HTA Submission

Where do the results go?

Long-term extrapolations from parametric survival models for time-to-event (TTE) outcomes, fitted based on clinical trial data, are routinely used to **provide inputs for cost-effectiveness analysis** in support of health technology assessment (HTA).

The Incremental Cost Effectiveness Ratio (ICER) is a key metric in the economic evaluation of treatment benefit. A prevalent measure of ‘effectiveness’ in ICER is the **quality-adjusted life year** (**QALY**), where the results of survival extrapolations are required.

Two key challenges

Both challenges are in the task title.

One challenge pertains to the **long-term extrapolation**, or **external validity**. Usually, the time horizon for prediction is 30-40 years, while an oncology trial is typically no more than 5 years.

The other challenge relates to **parametric assumptions** to fit the observed data, or **internal validity**. Typically, several parametric models for TTE outcomes are required to be fitted for a dataset (e.g. NICE TSD 14). How to define and select the best model is an important question.

NICE (UK) TSD-14

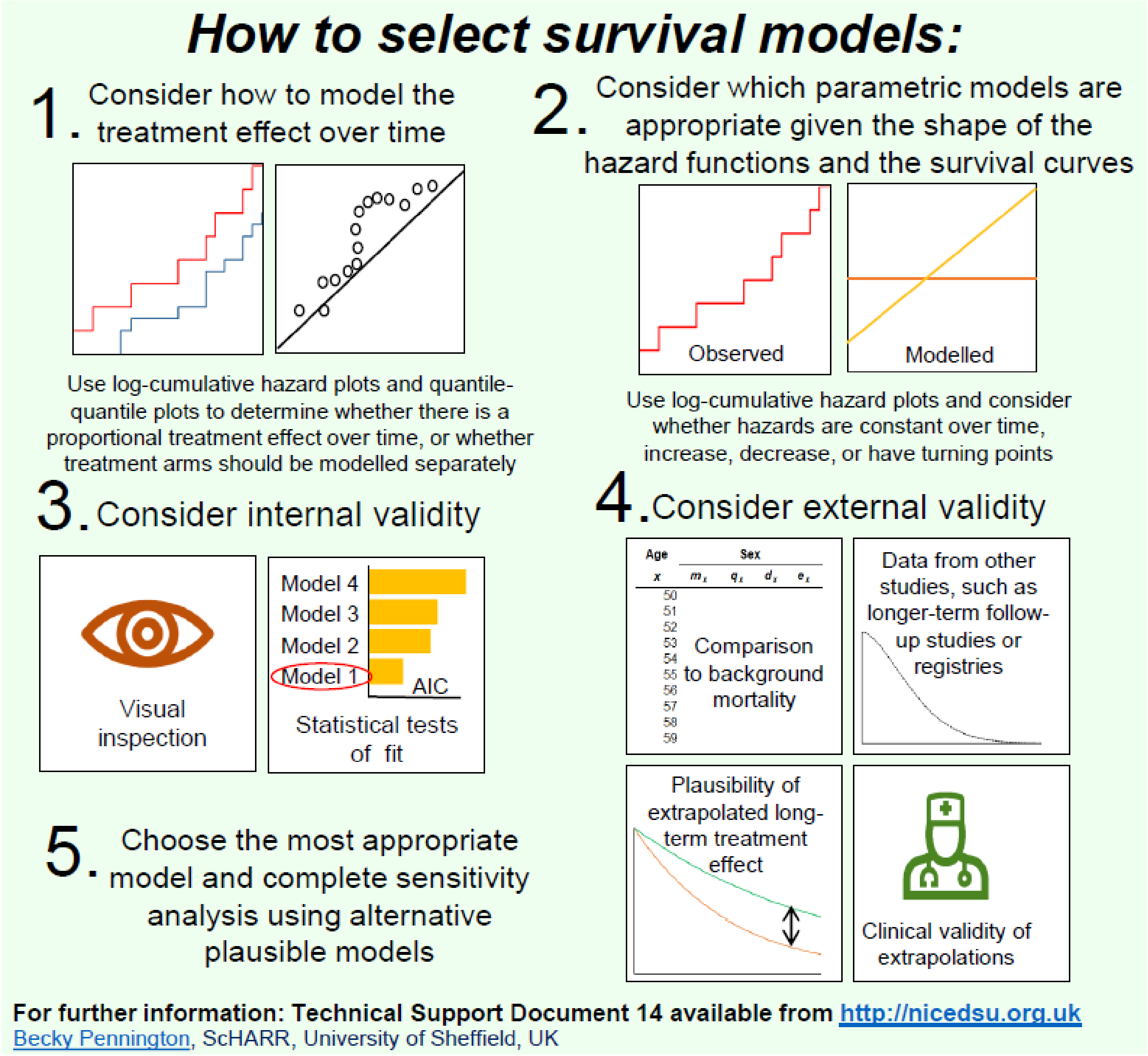
Latimer, N.R., 2013. Survival analysis for economic evaluations alongside clinical trials—extrapolation with patient-level data: inconsistencies, limitations, and a practical guide. Medical Decision Making, 33(6), pp.743-754.

TSD 14: Survival analysis for economic evaluations alongside clinical trials – extrapolation with patient-level data (last updated March 2013) (PDF, 395KB)

TSD 14 one page summary (PDF, 102KB)

Key words:

- ✓ "plausibility of assumed model"
- ✓ "internal validity" *(how well model fits the observed data)*
- ✓ "external validity" *(how well the model can predict the unobserved, particularly, far down the future)*
- ✓ "choose the most appropriate … sensitivity analysis"



How to select survival models:

1. Consider how to model the treatment effect over time

Use log-cumulative hazard plots and quantile-quantile plots to determine whether there is a proportional treatment effect over time, or whether treatment arms should be modelled separately

2. Consider which parametric models are appropriate given the shape of the hazard functions and the survival curves

Use log-cumulative hazard plots and consider whether hazards are constant over time, increase, decrease, or have turning points

3. Consider internal validity

4. Consider external validity

Age
Sex
x
m_x
q_x
d_x
e_x
50
51
52
53
54
55
56
57
58
59
Comparison to background mortality
Data from other studies, such as longer-term follow-up studies or registries
Plausibility of extrapolated long-term treatment effect
Clinical validity of extrapolations

5. Choose the most appropriate model and complete sensitivity analysis using alternative plausible models

For further information: Technical Support Document 14 available from http://nicedsu.org.uk
Becky Pennington, ScHARR, University of Sheffield, UK

Cited from: TSD 14 one page summary

What are we sharing today?

A more general framework is proposed by 7 experts: **4** Questions, **8** Steps, see details in Figure 1 of the reference below

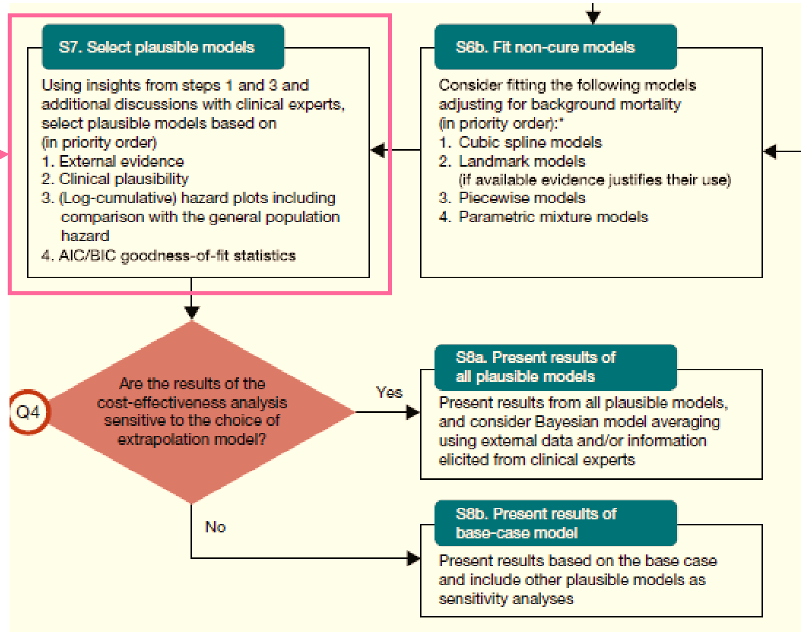
Palmer, S., Borget, I., Friede, T., Husereau, D., Karnon, J., Kearns, B., Medin, E., Peterse, E.F., Klijn, S.L., Verburg-Baltussen, E.J. and Fenwick, E., 2023. A guide to selecting flexible survival models to inform economic evaluations of cancer immunotherapies. *Value in Health*, *26*(2), pp.185-192.

Conducting a scientifically sound survival extrapolation is a complex process that contains a mixture of **qualitative** and **quantitative** steps with case-specific inputs from experts and data of different sources.

Fitting several models and selecting plausible models are an indispensable part. This presentation is to share parts of our ongoing research work to answer questions below, aiming to provide more relevant and reliable quantitative evidence for decision making:

- When and how to incorporate external data conceptually? How to derive a more prediction-focused performance metric, to support evaluation in direction of external validity when no suitable external evidence is available? Is it worthwhile to routinely report it?
- How to formulate proper test-based model selection procedures to supplement or replace absolute ranking of goodness-of-fit statistics (AIC/BIC), for better balancing fit and parsimony while accounting for sample uncertainty?

Figure 1. Flexible survival model selection algorithm. Where there is a specific preference ordering, this is shown in the algorithm as numbered bullets with the accompanying text "in priority order." Note that all of the models described here could be implemented in a relative survival framework to take account of background mortality (ie, other-cause mortality).[9,11] *In addition, standard parametric models should also be considered for comparative purposes as HTA agencies are likely to expect to see them. Among the standard parametric models, the generalized gamma, log-logistic, and log-normal would be most suitable where flexible modeling is suggested given that the other models are not able to capture turning points in the hazard.



S7. Select plausible models

Using insights from steps 1 and 3 and additional discussions with clinical experts, select plausible models based on (in priority order)
1. External evidence
2. Clinical plausibility
3. (Log-cumulative) hazard plots including comparison with the general population hazard
4. AIC/BIC goodness-of-fit statistics

S6b. Fit non-cure models

Consider fitting the following models adjusting for background mortality (in priority order):*
1. Cubic spline models
2. Landmark models (if available evidence justifies their use)
3. Piecewise models
4. Parametric mixture models

Are the results of the cost-effectiveness analysis sensitive to the choice of extrapolation model?

S8a. Present results of all plausible models

Present results from all plausible models, and consider Bayesian model averaging using external data and/or information elicited from clinical experts

S8b. Present results of base-case model

Present results based on the base case and include other plausible models as sensitivity analyses

Cited from: https://doi.org/10.1016/j.jval.2022.07.009

Prediction-focused Performance Metric

with and without external evidence

Suitable External Data to Assess External Validity

ppm-xx refers to the reference in this section (see slide 13)

> *“ In health technology assessment it is common for survival models fitted to clinical trial data to be used to extrapolate far into the future, with no external information taken into account. Sometimes extrapolations are compared to external evidence to provide a form of validation. It is relatively rare for external information to actually be incorporated in the model fitting process, …”*

- NICE TSD 21 [ppm-1]

Active research fields that are still lacking clear guidance:

? Are suitable external data available to facilitate long-term extrapolation beyond the observed trial period?

Ex: is cause-specific mortality rate available than the one in general population?

? Incorporate these external data directly in the model or use them to guide model validation?

When Suitable External Data is available for Validation

Example Sources of External Data:

See more discussion in [ppm-1,3,4]

- Trial data or real-world studies of the same product for different indications that may have a longer follow up time
- National administrative data (e.g. standardized mortality rates and lifetime tables)
- National or regional registries (e.g. SEER, Flatiron)
- Cohort studies
- Published literature related to the control arm, or have similar drug mechanisms as the experimental arm (rarely available due to the novelty of the experimental arm)
- Clinical opinions and/or expert opinions

How to Incorporate to Guide Model Selection

See more discussion in [ppm-5,6]

Note that

- Filtering and/or curating external data to represent the population of interests is essential (e.g. via probability weighting) [ppm-4]
- Additional assumptions may need to be made in conjunction with the available and good-quality external data

In case that the (adjusted) external data can be used to assess the prediction accuracy of the interested response (e.g. survival probability, hazard rate), one can quantify it by mean square error, mean absolute difference [ppm-7] or simply log-likelihood (see example in next slide).

Otherwise, usually the assessment would be conducted visually.

When Suitable External Data is NOT available for Validation

One may consider a more prediction-focused performance metric than AIC/BIC, for example, using K-fold Cross-Validation (NICE TSD-14 (2013), Section 3.4, it was also mentioned that this technique has not yet been used in any appraisals).

Stone in 1977 [ppm-8] has shown that AIC is asymptotically equivalent to leave-one-out validation procedure (i.e. N-fold Cross Validation procedure when sample size is N). However, in finite sample, it is quite common to see (in machine learning literature) that model selection based on AIC/BIC and based on Cross-Validation would lead to different "best" model.

The detailed procedure of Cross-Validation for a survival model is not enlisted in NICE TSD-14. We outline an example procedure on the right:

Given survival data $D = \{(t_i, e_i)\}_{i=1}^{N}$ for a sample of N subjects, where t_i is the time of death if $e_i = 1$, or the time of censored (drop-out or the end of follow up) if $e_i = 0$. Assume no prognostic factor for simplicity of illustration.

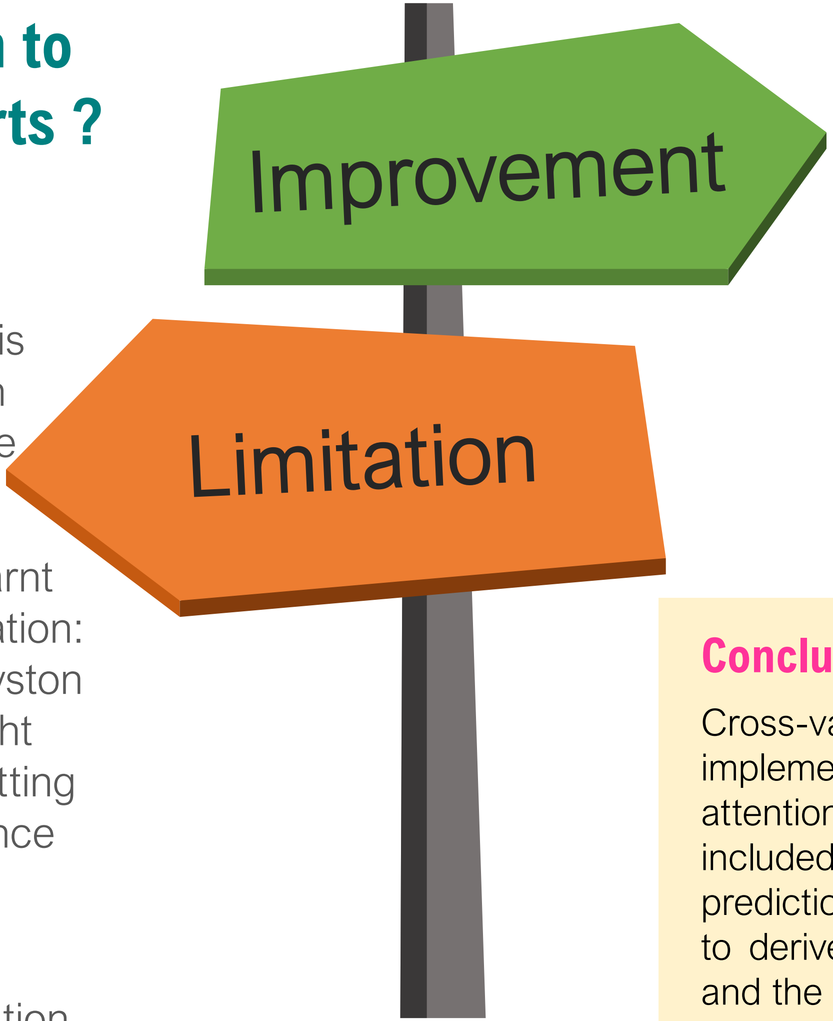
- For each b in $\{1, \dots, B\}$, where B is a large integer such as 10000
    - Randomly divide sample index $I = \{1,2, \dots, N\}$ into K sets $I_1^{(b)}, I_2^{(b)}, \dots, I_K^{(b)}$ that are mutually exclusive, $\mathrm{U}_{k=1}^{K} I_k^{(b)} = I$, and $\left|I_1^{(b)}\right| = \left|I_2^{(b)}\right| = \cdots = \left|I_{K-1}^{(b)}\right| = \lfloor N/K \rfloor$ (largest integer smaller than N/K)
    - Set $D_k = \{(t_j, e_j): j \in I_k^{(b)}\}$ for $k = 1, \dots, K$.
    - For each k in $\{1, \dots, K\}$,
        - Train a parametric survival model with density function $f\left(T; \hat{\theta}^{(k)}\right)$ where T is survival time and $\hat{\theta}^{(k)}$is the MLE of the governing parameters based on $\{D_h: h \neq k, h \in \{1, \dots, K\}\}$. Corresponding survival function is $S(\cdot)$
        - Calculate $l_k^{(b)} = \log\left(\prod_{i=1}^{N} f\left(t_i; \hat{\theta}^{(k)}\right)^{e_i} S\left(t_i; \hat{\theta}^{(k)}\right)^{e_i - 1}\right)$
    - Calculate $l^{(b)} = \sum_k l_k^{(b)}$
- Final performance metric is $l^* = \sum_b l^{(b)} / B$. Standard error can also be derived based on $\{l^{(b)}\}_{b=1}^{B}$

Is it worthwhile to add Prediction-focused Performance Metrics via Cross-Validation to Routine Extrapolation Reports ?

At the end of the day, no external data beyond the trial period is employed in this approach. At best, cross-validation is an internal validation procedure that has the potential to evaluate external validity.

Another important note that we have learnt through past experiments of implementation: incorporating flexible spline models (Royston & Parmar, 2010) into the procedure might need extra care to ensure consistent setting of knots and for potential non-convergence issue in some iterations.

When sample size is very small, this procedure should be used with high caution.



Unlike AIC/BIC, this performance metric is prediction-focused, hence one may argue it provides a more relevant depiction of the models for the task they are commissioned.

It may also reflect sample uncertainty by design. The standard error from the iterative procedure could be derived with small additional effort.

Conclusion:

Cross-validation (CV) is less straightforward to be implemented for routine reporting, particularly, extra attention is needed when flexible spline models are included in the candidate pool. However, it is a more prediction-focused performance metric with possibility to derive standard error to reflect sample uncertainty, and the ranking based on CV is not necessary the same as AIC/BIC in finite sample

References in this section

Indexed as [ppm-(number)]

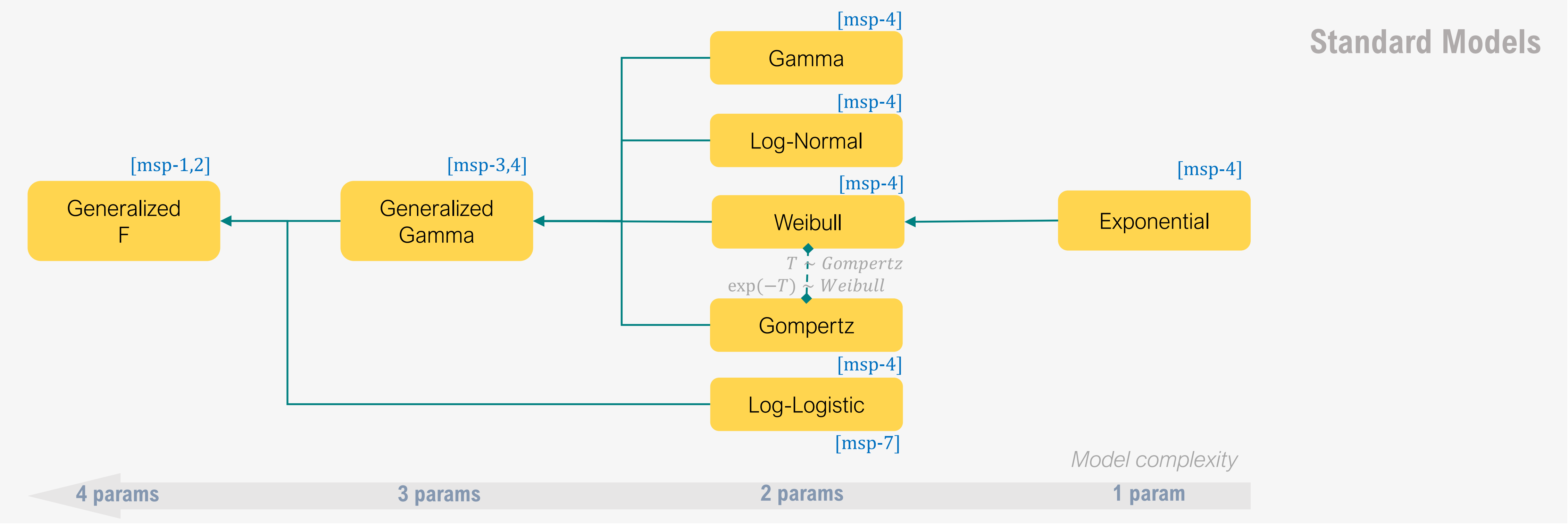
1. Rutherford, M., Lambert, P., Sweeting, M., Pennington, R., Crowther, M.J. and Abrams, K.R., 2020. Nice dsu technical support document 21: Flexible methods for survival analysis. Leicester, UK: Department of Health Sciences, University of Leicester, pp.1-97.
2. Latimer, N., 2011. NICE DSU technical support document 14: survival analysis for economic evaluations alongside clinical trials-extrapolation with patient-level data. Report by the Decision Support Unit.
3. Palmer, S., Borget, I., Friede, T., Husereau, D., Karnon, J., Kearns, B., Medin, E., Peterse, E.F., Klijn, S.L., Verburg-Baltussen, E.J. and Fenwick, E., 2023. A guide to selecting flexible survival models to inform economic evaluations of cancer immunotherapies. *Value in Health*, *26*(2), pp.185-192.
4. Jackson, C., Stevens, J., Ren, S., Latimer, N., Bojke, L., Manca, A. and Sharples, L., 2017. Extrapolating survival from randomized trials using external data: a review of methods. Medical decision making, 37(4), pp.377-390.
5. Royston, P. and Altman, D.G., 2013. External validation of a Cox prognostic model: principles and methods. BMC medical research methodology, 13, pp.1-15.
6. Ramspek, C.L., Jager, K.J., Dekker, F.W., Zoccali, C. and van Diepen, M., 2021. External validation of prognostic models: what, why, how, when and where?. Clinical Kidney Journal, 14(1), pp.49-58.
7. Pennington, M., Grieve, R., Van der Meulen, J. and Hawkins, N., 2018. Value of external data in the extrapolation of survival data: a study using the NJR data set. Value in Health, 21(7), pp.822-829.
8. Stone, M., 1977. An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. Journal of the Royal Statistical Society: Series B (Methodological), 39(1), pp.44-47.

Model Selection Procedure

based on (a sequence of) statistical tests

Typical Parametric Models

msp-xx refers to the reference in this section (see slide 25)



Standard Models

Generalized F [msp-1,2]

Generalized Gamma [msp-3,4]

Gamma [msp-4]

Log-Normal [msp-4]

Weibull [msp-4]

Gompertz [msp-4]

Log-Logistic [msp-7]

Exponential [msp-4]

$T \sim Gompertz$

$\exp(-T) \sim Weibull$

Model complexity

4 params

3 params

2 params

1 param

Flexible Models

Flexible restricted cubic spline (Royston & Parmar, 2010) [msp-5]

Two-piece model (Nelson et al, 2008) [msp-6]

... ...

* Inference of these models can be derived using flexsurv (Jackson, [msp-7])

Selecting Plausible Parametric Models

The status-quo approach in practice

- Visual investigation of fitted curve versus KM plot
- Ranking based on goodness-of-fit statistics (AIC/BIC)

Intuitive, and it is a reasonable first step

Pain points:

- Subjectivity in assessment outcome
- Usually inflated uncertainty (due to small set at risk) in the tail of KM curve is not accounted for
- Unreliable when available survival data is immature (not a uncommon situation at the initial of HTA application)

Easy to implement, prevalent model selection procedure

Pain points:

- External validity is not reflected
- Containing sample uncertainty, e.g. small difference among top performing models might not be statistically meaningful.
- Seeking for the best fit, not necessarily needed in all cases

[msp-8] confirmed the observation of a similar reality:

> *"… All experts rated the importance of the AIC/BIC statistic below 5, on a scale of 1 (not important) to 10 (extremely important), with an average score, across all experts, of 3.7. Their opinion may be at odds with current practice. "*

Selecting Plausible Parametric Models

The status-quo approach in practice

- Visual investigation of fitted curve versus KM plot
- Ranking based on goodness-of-fit statistics (AIC/BIC)

Intuitive, and it is a reasonable first step

Pain points:

- Subjectivity in assessment outcome
- Usually inflated uncertainty (due to small set at risk) in the tail of KM curve is not accounted for
- Unreliable when available survival data is immature (not a uncommon situation at the initial of HTA application)

Easy to implement, prevalent model selection procedure

Pain points:

- External validity is not reflected
- Containing sample uncertainty, e.g. small difference among top performing models might not be statistically meaningful.
- Seeking for the best fit, not necessarily needed in all cases

Motivates our test-based procedure

[msp-8] confirmed the observation of a similar reality:

> *"… All experts rated the importance of the AIC/BIC statistic below 5, on a scale of 1 (not important) to 10 (extremely important), with an average score, across all experts, of 3.7. Their opinion may be at odds with current practice. "*

Statistical Test-based Procedure for Model Selection

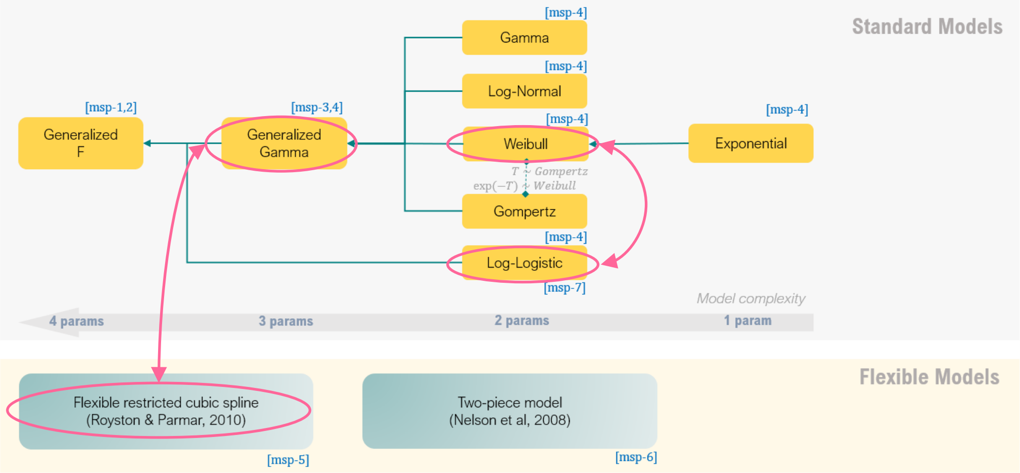
Ranking-based procedure begin with a given set of models, aim to select *the best one* with respect to a performance metric. Optimal use would be as a part of a quantitative decision process. [msp-9,11]

- All models under consideration are treated symmetrically (no pathway)
- Ends with a definite outcome (not ideal in all cases)
- Parsimony of model may be accounted for in the performance metric in a specific parametric form (via penalty on model complexity), user has little control
- Sample uncertainty of the performance metric is not accounted for

Test-based procedure begin with a set of models and null hypotheses / a pathway. It aims to answer if there is any *statistically significant evidence* of deviation from the null hypotheses towards one or more alternative hypotheses (i.e. alternative parametric forms). Optimal use is to assess empirical validity of a theoretical prediction [msp-11] to support decision making.

- In every test (pairwise or joint), models are treated asymmetrically by design (based on null and alternative status)
- With proper interpretation of test outcome, does not always leads 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 main analysis), definition of pathway (a sequence of tests) can be made, which would incorporate the model preference (e.g. standard over flexible) and model complexity specified by the user

Two Challenges to Formulate a (Sequential) Test-based Procedure



Standard Models

[msp-4] Gamma

[msp-4] Log-Normal

[msp-1,2] Generalized F

[msp-3,4] Generalized Gamma

[msp-4] Weibull

[msp-4] Exponential

$T \sim$ Gompertz

$\exp(-T) \sim$ Weibull

Gompertz

[msp-4] Log-Logistic [msp-7]

Model complexity

4 params · 3 params · 2 params · 1 param

Flexible Models

Flexible restricted cubic spline (Royston & Parmar, 2010) [msp-5]

Two-piece model (Nelson et al, 2008) [msp-6]

Non-Nested Testing

Main difference between nested and non-nested setting:

The usual log-likelihood or Wald statistics used in the nested setting are centered at zero under the null hypothesis, while this is not true in a non-nested setting [msp-11]. Majority of pairwise comparisons of the models in the above map is in non-nested setting (e.g. see next slide). In section 3.3 of NICE TSD-14 (2013) highlighted that this mistake was made in past NICE TAs.

Control overall error rate for a sequence of tests

The same dataset is used to fit several models. They are now tested pairwise (if the alternative model fits better than the null model) in a pre-specified sequence with a stopping rule, in order to find an appropriate model for the main analysis.

One might want to control the overall error rate (e.g. in the sense of FDR or FWER) of the whole procedure.

Illustration: Distribution of p-value comparing 2 and 3 knots spline model

In general, 2-knots and 3-knots of Restrict Cubic Spline (RCS) Model are not nested if the knots are determined based on default algorithm by flexsurv::flexsurvspline.

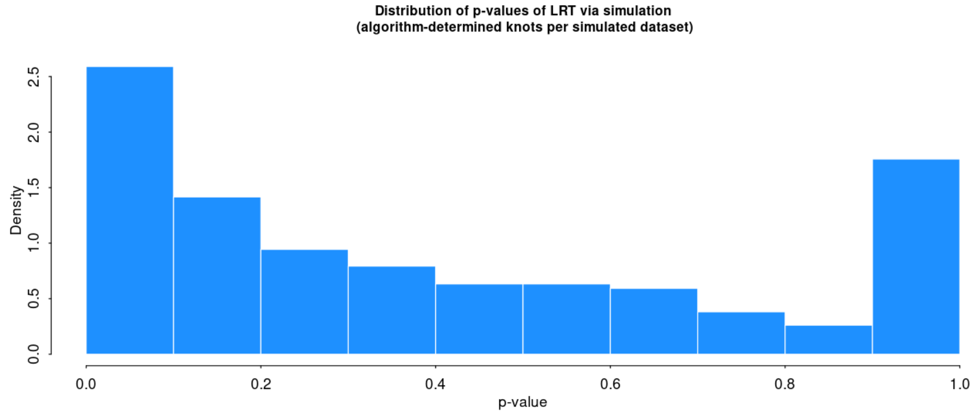
True model is based on open-source ovarian cancer survival data (included in {survival} R package). 1000 samples, each has 100 subjects, are then simulated.

```
library(flexsurv)
fit     <- flexsurvspline(formula=Surv(futime,fustat)~1, data=ovarian, k=2)
fit_a  <- flexsurvspline(formula=Surv(futime,fustat)~1, data=ovarian, k=3)
nd     <- data.frame(id=1:100)
simdat <- simulate(fit, nsim=1000, seed=12345, censtime = max(ovarian$futime), newdata = nd)
```

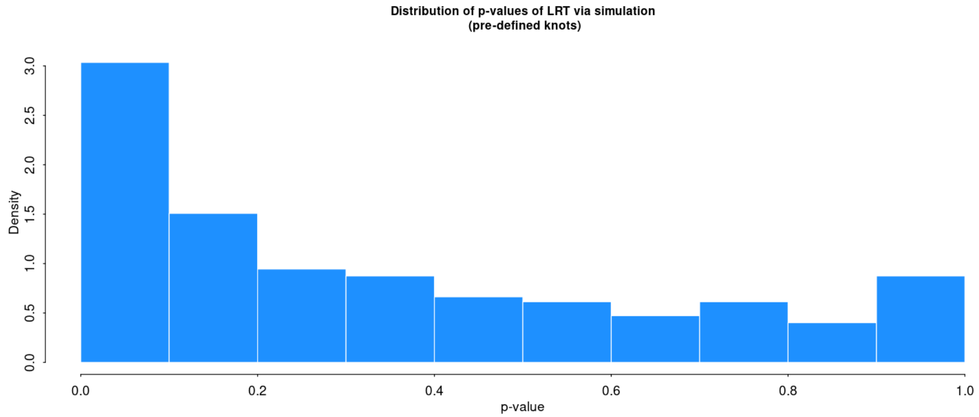
Distribution of p-value of likelihood-ratio (LR) between two RCS models is approximated based on this simulation. See plots on the right.

Let "usedat" be a simulated dataset of 100 subjects (a subset of "simdat")

Scenario 1: default location of knots determined by flexsurvspline given a "usedat"

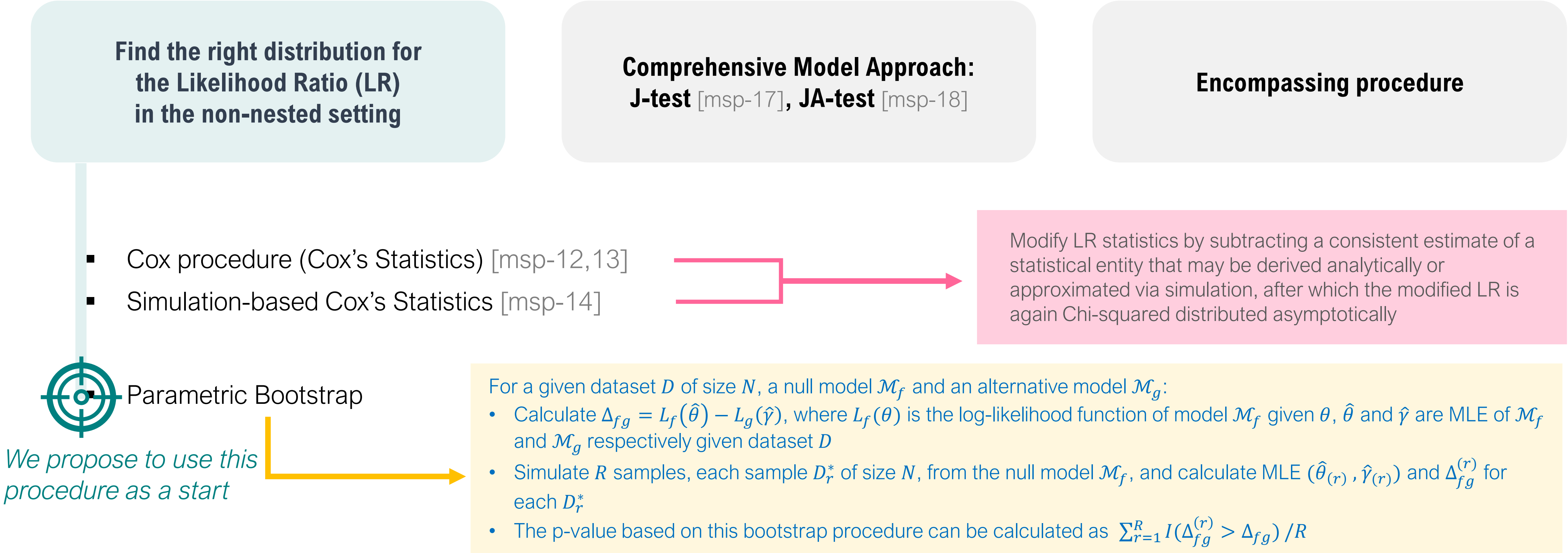

Distribution of p-values of LRT via simulation
(algorithm-determined knots per simulated dataset)

Scenario 2: 2-knots at (5.73, 6.09) vs. 3-knots at (5.73, 6.09, 6.15) across all "usedat"


Distribution of p-values of LRT via simulation
(pre-defined knots)

Non-Nested Hypothesis Testing in Model Selection

In general terms, two models ($\mathcal{M}_f$ and $\mathcal{M}_g$) are defined to be non-nested, if it is not possible to derive $\mathcal{M}_f$ (or $\mathcal{M}_g$) from the other model either by means of an exact set of parametric restrictions or as a result of a limiting process. More precise definition via Kullback-Leibler Information Criteria (KLIC) is given by [msp-8,9].



Find the right distribution for the Likelihood Ratio (LR) in the non-nested setting

Comprehensive Model Approach: J-test [msp-17], **JA-test** [msp-18]

Encompassing procedure

- Cox procedure (Cox's Statistics) [msp-12,13]
- Simulation-based Cox's Statistics [msp-14]

Modify LR statistics by subtracting a consistent estimate of a statistical entity that may be derived analytically or approximated via simulation, after which the modified LR is again Chi-squared distributed asymptotically

Parametric Bootstrap

We propose to use this procedure as a start

For a given dataset D of size N, a null model $\mathcal{M}_f$ and an alternative model $\mathcal{M}_g$:

- Calculate $\Delta_{fg} = L_f(\hat{\theta}) - L_g(\hat{\gamma})$, where $L_f(\theta)$ is the log-likelihood function of model $\mathcal{M}_f$ given θ, $\hat{\theta}$ and $\hat{\gamma}$ are MLE of $\mathcal{M}_f$ and $\mathcal{M}_g$ respectively given dataset D
- Simulate R samples, each sample D_r^* of size N, from the null model $\mathcal{M}_f$, and calculate MLE $(\hat{\theta}_{(r)}, \hat{\gamma}_{(r)})$ and $\Delta_{fg}^{(r)}$ for each D_r^*
- The p-value based on this bootstrap procedure can be calculated as $\sum_{r=1}^{R} I(\Delta_{fg}^{(r)} > \Delta_{fg})\,/R$

Control Overall Error Rate for a Sequence of Tests

G'sell, Wager, Chouldechova and Tibshirani [msp-15] offers an easy-to-implement rejection rule, which can reject a sequence of null hypotheses, $H_1, H_2, \dots, H_m$ in an ordered fashion while controlling the false discovery rate (FDR).

* A rejection rule is a function of the ordered p-values $p_1, \dots, p_m$ that returns a cut-off $\hat{k}$ such that $H_1, \dots, H_{\hat{k}}$ are rejected. FDR is defined as $\mathbb{E}\left(V(\hat{k})/\max(1, \hat{k})\right)$, where $V(k) = |\{i \in Q : i \le k\}|$, set $Q \subset \{1, \dots, m\}$ of those p-values that are null with the property that $\{p_i : i \in Q\} \sim_{iid} U[0,1]$

Procedure	**Steps**	**Notes**
ForwardStop	1. Transform p-values $p_1, p_2, \dots, p_m$ into a monotone increasing sequence $0 \le q_1 \le q_2 \le \cdots \le q_m$. 2. Reject hypotheses $1, \dots, \hat{k}_F$, where $\hat{k}_F = \max\left\{k \in \{1, \dots, m\} : -\frac{1}{k}\sum_{i=1}^{k} \log(1 - p_i) \le \alpha\right\}$	Moderately robust to potential misspecification of the null distribution of the p-values at high indices, since it will always reject the first hypotheses regardless of the last hypotheses
StrongStop	1. Transform p-values $p_1, p_2, \dots, p_m$ into a monotone increasing sequence $0 \le q_1 \le q_2 \le \cdots \le q_m$. 2. Reject hypotheses $1, \dots, \hat{k}_F$, where $\hat{k}_S = \max\left\{k \in \{1, \dots, m\} : \exp\left(\sum_{j=k}^{m} \frac{\log(p_j)}{j}\right) \le \frac{\alpha k}{m}\right\}$	Stronger guarantee than ForwardStop, as it controls both FDR and FWER given the non-null p-values precede the null p-values. If false discovery has a high cost, this rule might be attractive. A weakness is that the decision to reject at $\boldsymbol{k}$ depends on all the p-values after $\boldsymbol{k}$. If the very last p-values are larger than they should be under the uniform hypothesis, the rule suffers a considerable loss of power.

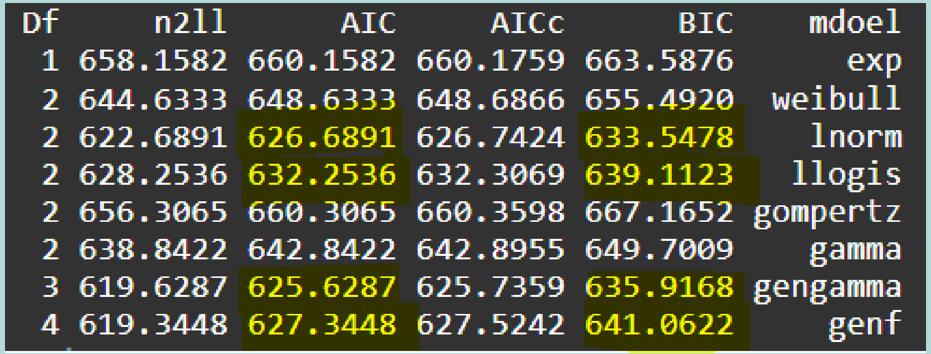
Open-source Data used in the next 2 examples

A subset (prognostic group = “Poor”) of Breast cancer data from {flexsurv} R package that consists of 226 subjects.

The 8 parametric standard survival models are fitted to the above data with no predictor

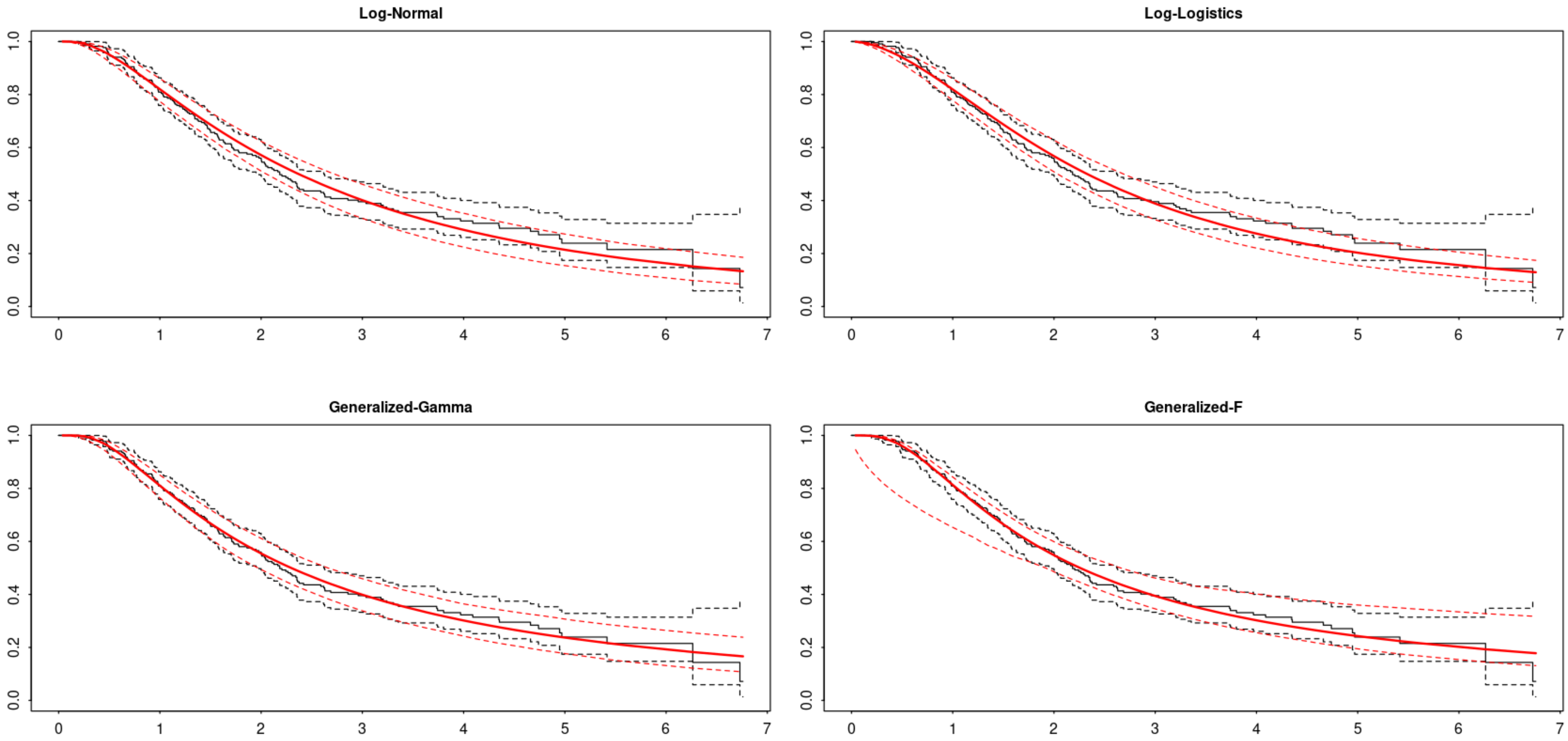
```
fit0 <- flexsurvreg(formula=Surv(recyrs,censrec)~1,data=bc[bc$group=="Poor",],dist="exp")
fit1 <- flexsurvreg(formula=Surv(recyrs,censrec)~1,data=bc[bc$group=="Poor",],dist="weibull")
fit2 <- flexsurvreg(formula=Surv(recyrs,censrec)~1,data=bc[bc$group=="Poor",],dist="lnorm")
fit3 <- flexsurvreg(formula=Surv(recyrs,censrec)~1,data=bc[bc$group=="Poor",],dist="llogis")
fit4 <- flexsurvreg(formula=Surv(recyrs,censrec)~1,data=bc[bc$group=="Poor",],dist="gompertz")
fit5 <- flexsurvreg(formula=Surv(recyrs,censrec)~1,data=bc[bc$group=="Poor",],dist="gamma")
fit6 <- flexsurvreg(formula=Surv(recyrs,censrec)~1,data=bc[bc$group=="Poor",],dist="gengamma")
fit7 <- flexsurvreg(formula=Surv(recyrs,censrec)~1,data=bc[bc$group=="Poor",],dist="genf")
```

Goodness-of-Fit can be summarized as below



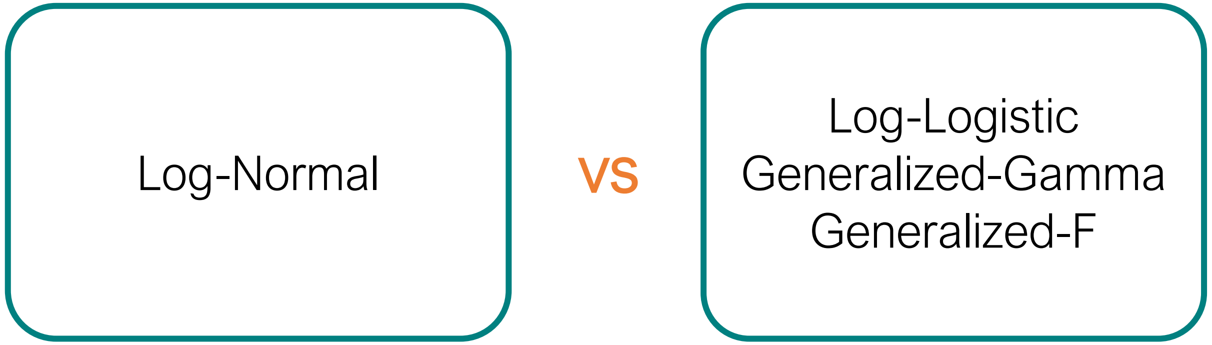
```
Df     n2ll      AIC     AICc      BIC    mdoel
 1 658.1582 660.1582 660.1759 663.5876      exp
 2 644.6333 648.6333 648.6866 655.4920  weibull
 2 622.6891 626.6891 626.7424 633.5478    lnorm
 2 628.2536 632.2536 632.3069 639.1123   llogis
 2 656.3065 660.3065 660.3598 667.1652 gompertz
 2 638.8422 642.8422 642.8955 649.7009    gamma
 3 619.6287 625.6287 625.7359 635.9168 gengamma
 4 619.3448 627.3448 627.5242 641.0622     genf
```

KM versus Fitted Curves in the previous example data



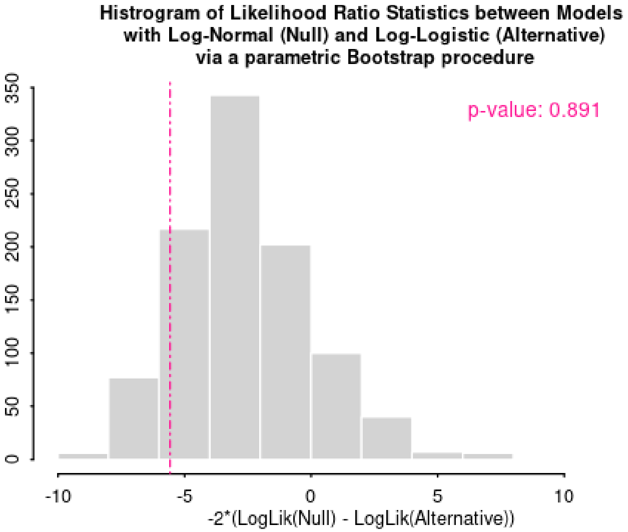
Example 1: Selection among Top Ranked Standard Models

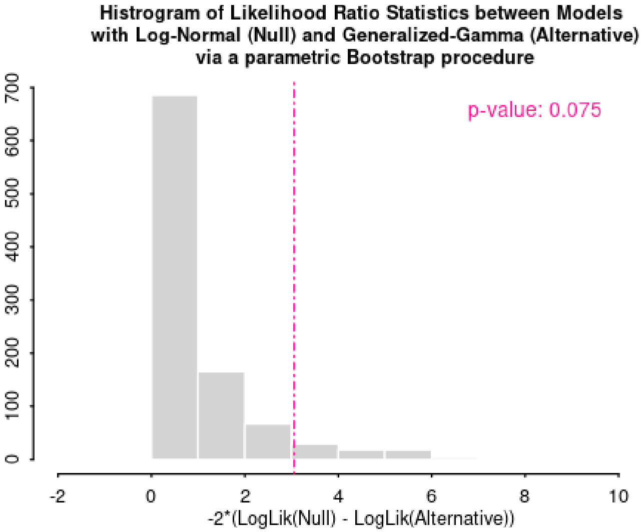
3 pairwise hypothesis tests, no model pathway (non-sequential)

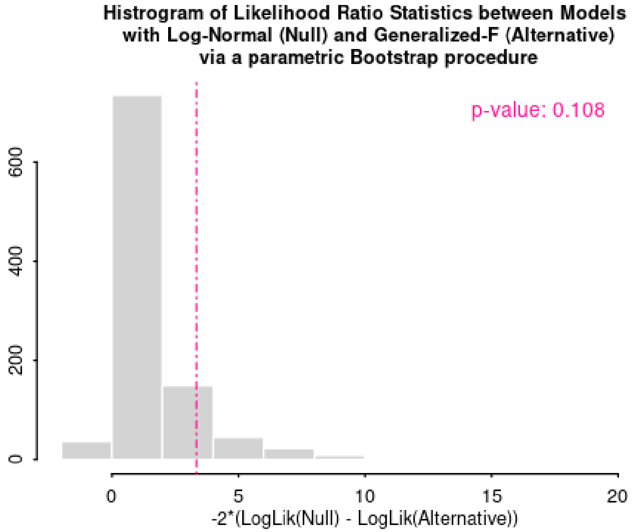


Conclusion:

There is no significant evidence in the data in support that Log-logistic, Generalized-Gamma, or Generalized-F fitted better than Log-Normal


Histrogram of Likelihood Ratio Statistics between Models with Log-Normal (Null) and Log-Logistic (Alternative) via a parametric Bootstrap procedure


Histrogram of Likelihood Ratio Statistics between Models with Log-Normal (Null) and Generalized-Gamma (Alternative) via a parametric Bootstrap procedure


Histrogram of Likelihood Ratio Statistics between Models with Log-Normal (Null) and Generalized-F (Alternative) via a parametric Bootstrap procedure

Example 2: Selection between Standard and Flexible Models

Consider a sequential test, with the following pathway of models that reflect the two principles in model choice prioritization: (1) less complex model is better, (2) standard model is preferred over the restricted cubic spline (RCS) model.

Step	P-Value	ForwardStop Statistics	Stop or not at $\alpha = 0.05$
1	0.000	0.000	Stop here
2	Not needed	--	--
3	Not needed	--	--

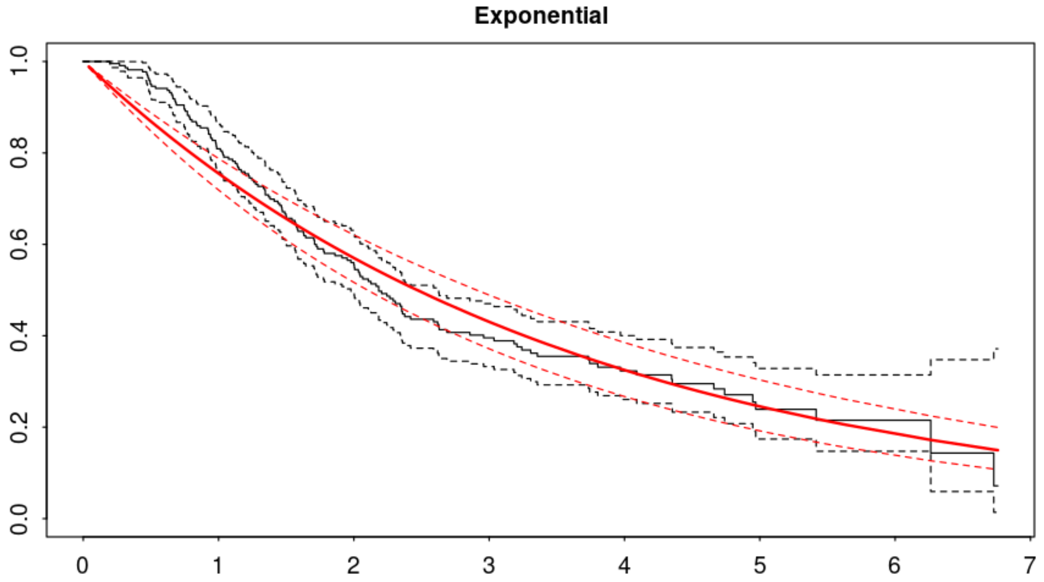
Pathway

(0) Exponential

(1) → Log-Normal

(2) → RCS with 2 nodes

(3) → RCS with 3 nodes



Reference in this section

Indexed as [msp-(number)]

1. Cox, C., 2008. The generalized F distribution: an umbrella for parametric survival analysis. Statistics in medicine, 27(21), pp.4301-4312.
2. Prentice, R.L., 1975. Discrimination among some parametric models. Biometrika, 62(3), pp.607-614.
3. Prentice, R.L., 1974. A log gamma model and its maximum likelihood estimation. Biometrika, 61(3), pp.539-544.
4. Cox, C., Chu, H., Schneider, M.F. and Munoz, A., 2007. Parametric survival analysis and taxonomy of hazard functions for the generalized gamma distribution. Statistics in medicine, 26(23), pp.4352-4374.
5. Royston, P. and Parmar, M.K., 2002. 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, 21(15), pp.2175-2197.
6. Nelson, C.L., Sun, J.L., Tsiatis, A.A. and Mark, D.B., 2008. Empirical estimation of life expectancy from large clinical trials: use of left-truncated, right-censored survival analysis methodology. Statistics in medicine, 27(26), pp.5525-5555.
7. Jackson, C.H., 2016. flexsurv: a platform for parametric survival modeling in R. Journal of statistical software, 70.
8. Palmer, S., Borget, I., Friede, T., Husereau, D., Karnon, J., Kearns, B., Medin, E., Peterse, E.F., Klijn, S.L., Verburg-Baltussen, E.J. and Fenwick, E., 2023. A guide to selecting flexible survival models to inform economic evaluations of cancer immunotherapies. *Value in Health*, *26*(2), pp.185-192.
9. Granger, C.W.J. and Pesaran, M.H., 2000. A decision theoretic approach to forecast evaluation. In Statistics and finance: An interface (pp. 261-278).
10. Vuong, Q.H., 1989. Likelihood ratio tests for model selection and non-nested hypotheses. Econometrica: journal of the Econometric Society, pp.307-333.
11. Pesaran, M.H. and Weeks, M., 2001. Non-nested hypothesis testing: an overview. A companion to theoretical econometrics, pp.279-309.
12. Cox, D.R., 1961, January. Tests of separate families of hypotheses. In Proceedings of the fourth Berkeley symposium on mathematical statistics and probability (Vol. 1, No. 0, p. 96).
13. Cox, D.R., 1962. Further results on tests of separate families of hypotheses. Journal of the Royal Statistical Society Series B: Statistical Methodology, 24(2), pp.406-424.
14. Pesaran, M.H. and Pesaran, B., 1993. A simulation approach to the problem of computing Cox's statistic for testing nonnested models. Journal of Econometrics, 57(1-3), pp.377-392.
15. G'Sell, M.G., Wager, S., Chouldechova, A. and Tibshirani, R., 2016. Sequential selection procedures and false discovery rate control. Journal of the Royal Statistical Society Series B: Statistical Methodology, 78(2), pp.423-444.
16. Jackson, C.H., Sharples, L.D. and Thompson, S.G., 2010. Survival models in health economic evaluations: balancing fit and parsimony to improve prediction. The international journal of biostatistics, 6(1).
17. Davidson, R. and MacKinnon, J.G., 1981. Several tests for model specification in the presence of alternative hypotheses. Econometrica: Journal of the Econometric Society, pp.781-793.
18. Fisher, G.R. and McAleer, M., 1981. Alternative procedures and associated tests of significance for non-nested hypotheses. Journal of Econometrics, 16(1), pp.103-119.

Ongoing work

Unfinished Work in Collaboration with Other Colleagues in MSD

Non-nested Test Procedure

- Simulation-based Evaluation of Type I error and Power of Proposed Parametric Bootstrap Procedure for Single and Joint test with various sample sizes
- Exploration of the change of evaluation result when setting a restricted cubic spline model as the null, due to its data-dependent base functions.
- More efficient Bootstrap Procedure via seeking an appropriate pivot statistics

ForwardStop Sequential Test

- Simulation-based Evaluation of false discovery rate (FDR) and familywise error rate for several pre-defined model pathways, assuming true model is a generalized gamma or a 3-knots restricted cubic spline model
- Exploration of remedy to control FDR for a linear pathway with composition nodes (i.e. a composition node contains multiple candidate models)

Thank You

Contact: xiangyi.gregory.chen@msd.com