Impact of data maturity on the estimation of the within-trial hazard function: an example from metastatic castration resistant prostate cancer

Jack Williams[1], Rob Hettle[1], Thomas Haugli-Stephens[2]

Simulation and Analytics, AstraZeneca, Cambridge, UK[1] Health Economics, AstraZeneca, Oslo, Norway[2]

PT16

Introduction

- In oncology, cost-effectiveness analyses often use parametric survival models fitted to clinical trial data to predict the lifetime costs and outcomes of treatment.
- The choice of parametric survival model is often an area of uncertainty that is subject to critique by decision makers (e.g., Health Technology Assessment (HTA) organisations) because of its impact on results. This choice, therefore, requires a rigorous and comprehensive assessment of multiple criteria including model fit to the trial data and the plausibility of extrapolations.
- To support decision-making, HTA guidelines recommend that the empirical hazard, a non-parametric or semi-parametric estimate of the observed instantaneous hazard rate, is presented alongside the modelled hazard from the parametric survival models to assess goodness of fit to the observed clinical trial data.[1-3]
- Alongside statistical fit, recent guidance recommends that the long-term hazard is plotted to assess the plausibility of model extrapolation.[4,5] In the absence of data, the plausibility of the extrapolated hazard for new treatments is often informed by visual inspection of trends in the empirical hazard function.
- The accurate estimation of the empirical hazard is therefore critical to making an informed choice of extrapolation.
- Previous research has discussed that different algorithms and smoothing assumptions can impact on the shape of the empirical hazard function and advocate caution in their interpretation, particularly in the tail of curves where the number at risk is low.[4,6] However, the extent to which data maturity may influence the empirical hazard, and therefore the choice of survival model extrapolation, has not been well described.

Aim

- We compare different methods of estimating the empirical hazard across three reported data cut offs (DCO) of a large prostate cancer trial and assess whether existing methods produce consistent estimates of the empirical hazard function.

Methods

- Empirical hazards for overall survival (OS) were estimated across three DCOs (**Table 1**) from the PREVAIL trial (NCT01212991); a phase 3 randomised clinical trial of enzalutamide versus placebo in metastatic castration resistant prostate cancer (mCRPC).
- For each DCO, pseudo-patient level data for enzalutamide were recovered from digitised Kaplan-Meier (KM) plots (**Figure 1**) using the Guyot algorithm.[7,8]
- Following guidelines, unsmoothed hazards were estimated using the *muhaz* R package (*pehaz* function) and smoothed hazards were estimated using the *muhaz* and *bshazard* packages. The default settings were used for all functions. The analysis was performed in R version 4.0.2.
- The shapes of different hazard functions were compared to understand how these differed by estimation method and data availability.

Table 1 Details and sources of the three data cut-offs from the PREVAIL study

Data cut-off (DCO)	Median follow-up	Maximum follow-up	Overall survival maturity	Source
DCO1	22 months	34 months	27.6% (241 events)	[9]
DCO2	31 months	44 months	Not reported	[10]
DCO3	69 months	81 months	79.0% (689 events)	[11]

Results

- At the longest follow-up (DCO3), the empirical hazards initially increased (0 to ~24 months), before remaining broadly constant thereafter, with *muhaz* and *bshazard* estimating similar smoothed functions (**Figure 2**).
- The empirical smoothed hazards at DCOs 1-2 were generally consistent with DCO3 up to approximately 30 months of follow-up (~4% at risk, DCO1).
- After month 30, *muhaz* estimated increasing (DCO1) and then decreasing (DCO2) hazards whilst *bshazard* estimated increasing (DCO1) and constant (DCO2) hazards, respectively.
- In this example, *bshazard* provided smoother hazard functions than *muhaz* and was less sensitive to changes in the hazard at the tail.
- However, the smoother hazard function from *bshazard* did not lead to a more accurate estimate of the post-30-month hazard rate at DCO3. Additionally, at DCO1 *bshazard* predicted a shallower increase in the hazard than *muhaz*, but with the increase occurring at an earlier time point (~24 months versus ~30 months for *muhaz*).
- On visual inspection, *bshazard* suggests a clear change point in the hazard function which was not observed with longer follow-up. Therefore, the smoother fit may still result in incorrect selections of survival models based on the fit to the underlying hazards.
- Overall, both *muhaz* and *bshazard* hazard functions can be unstable at the tail, when the numbers of patients remaining at risk are low.

Figure 1. Overall Survival Kaplan-Meier plots across three DCOs from the PREVAIL study (plots re-produced from digitized data)

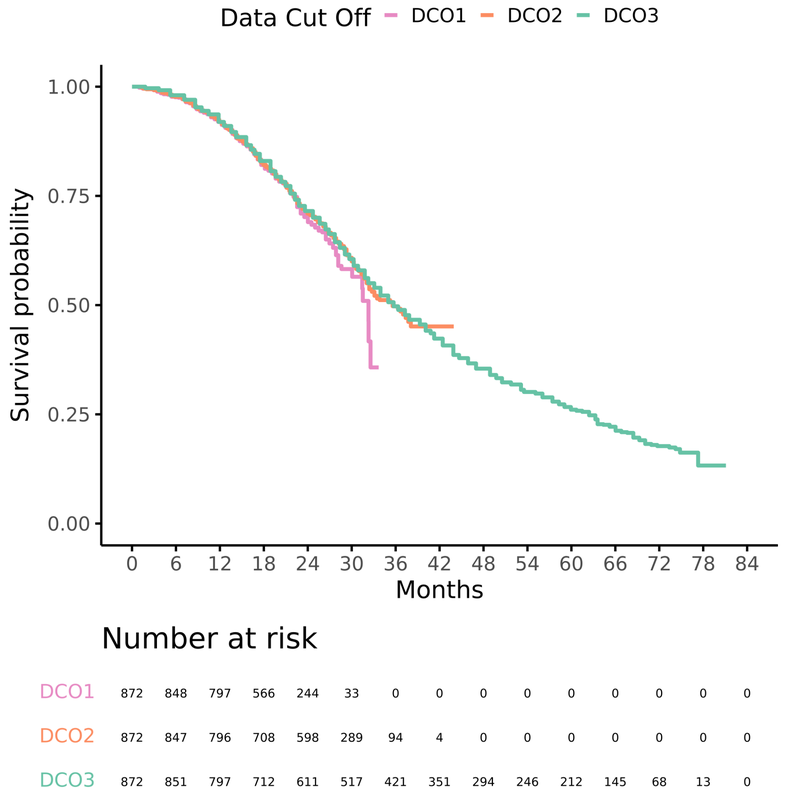
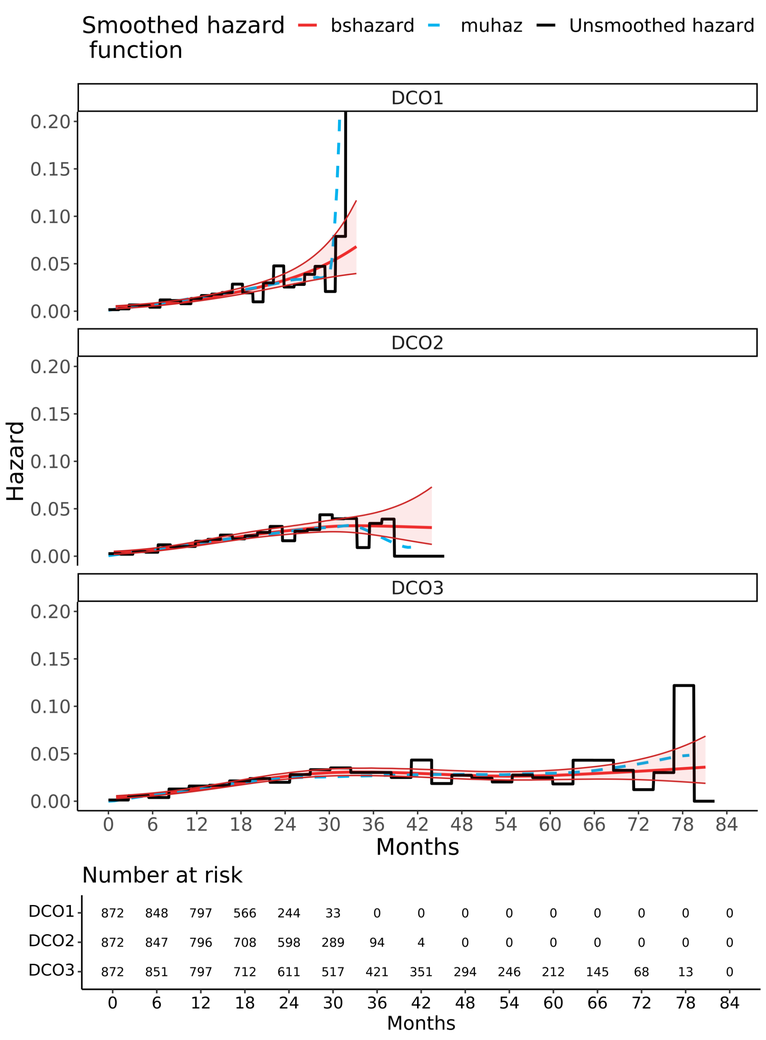


Figure 2. Smoothed hazard estimations at each DCO of the PREVAIL study, using '*muhaz*' and '*bshazard*' R packages, alongside unsmoothed hazards



Limitations

- We have focused on two popular methods for estimating smoothed hazards, however there are numerous other methodologies.[12]
- The analysis compares the default hazard function settings and did not extensively explore different options or truncation points for each function.
- We do not evaluate which of the two smoothed hazard functions is the most appropriate.

Conclusions

- In this study, the empirical hazard functions appeared generally robust to the choice of smoothing function until the numbers at risk were small.
- The hazard trend at the tail of the curve can differ considerably depending on function used, and these trends at the tail may not be reliable to inform the long-term hazards (as demonstrated with additional follow-up in our example).
- As such, caution should be taken when interpreting the tail of the hazard function when the number of patients remaining at risk are low, as this could suggest a different long-term hazard and therefore choice of curve extrapolation.
- Therefore, we recommend that hazard functions are presented and interpreted alongside the number of patients at risk, and that the median follow-up is also taken into consideration.
- Our findings suggest more research is needed to determine both the minimum evidence required to reliably estimate hazards, and the preferred methodology for hazard estimation. Where uncertain, the clinical plausibility of the empirical hazard function should also be considered when informing extrapolation.

References

1. Statens legemiddelverk (Norwegian Medicines Agency). Guidelines for the submission of documentation for single technology assessment (STA) of pharmaceuticals 2021.
2. Medicinrådet (Danish Medical Council). Anvendelse af forløbsdata i sundhedsøkonomiske analyser. 2020.
3. Medicinrådet (Danish Medical Council). Application form for assessment of new pharmaceuticals. 2023.
4. Palmer S et al. Value Health. 2023;26(2):185-92.
5. Rutherford MJ et al. NICE DSU Technical Support Document 21. 2020.
6. Wang J-L. Smoothing Hazard Rates: Contribution to the Encyclopedia of Biostatistics. 2003.
7. Rohatgi A. WebPlotDigitizer 2022
8. Guyot P et al. BMC Medical Research Methodology. 2012;12(1):9.
9. Beer TM et al. New England Journal of Medicine. 2014;371(5):424-33.
10. Beer TM et al. European Urology. 2017;71(2):151-4.
11. Armstrong AJ et al. European Urology. 2020;78(3):347-57.
12. Hagar Y, Dukic V. Comparison of hazard rate estimation in R. arXiv preprint:150903253. 2015.

Funded by AstraZeneca

Presented at ISPOR Europe 2023, Copenhagen