A simulation-based comparison of statistical methods for time-to-event data analysis under non-proportional hazards

Tobias Fellinger[1], Florian Klinglmüller[1], co-authors from the CONFIRMS consortium[2,3,4,5]

1 Agentur für Gesundheit und Ernährungssicherheit (AGES), Vienna, Austria
2 University Medical Center Göttingen, Department of Medical Statistics, Göttingen, Germany
3 Federal Institute for Drugs and Medical Devices, Bonn, Germany
4 Uppsala University, Department of Pharmacy, Uppsala, Sweden
5 Medical University of Vienna, Center for Medical Data Science, Vienna, Austria

Introduction

Well-established methods for time-to-event data are available when the proportional hazards assumption holds. Under non-proportional hazards (NPH) there is no consensus on the best inferential approach. However, a wide range of parametric and non-parametric methods for testing and estimation in this scenario have been proposed.

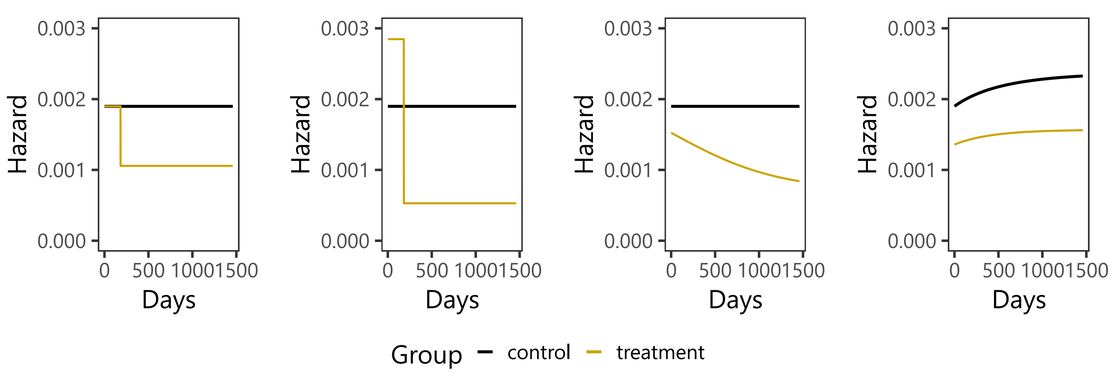
To provide recommendations on the statistical analysis of clinical trials where NPH are expected, we conducted a comprehensive simulation study. The simulated scenarios included delayed onset of treatment effect, crossing hazard curves, subgroups with different treatment effect and changing hazards after disease progression.

We examined a wide range of methods including weighted log-rank tests, the MaxCombo test, summary measures such as the restricted mean survival time (RMST), average hazard ratios, and milestone survival probabilities as well as accelerated failure time regression models.

Scenarios/Methods

The parameters of the simulated scenarios were derived from a review of European public assessment reports (EPARs) that identified 16 marketing authorization procedures where issues with non-proportional hazards were noted [1]. Summary measures like median survival in both treatment groups, time of separation of the survival curves were extracted from the EPARs and realistic parameter values for the simulation were derived.

The simulation study [2,3] investigated different departures from PH and different parameter values for the hazard in the control arm, sample size, recruitment time and amount of random censoring. In addition scenarios from parametric joint models were simulated and data was sampled from individual participant data reconstructed from published Kaplan-Meier curves. Different trial designs were simulated for each scenario, a simple event driven design in which the trial is stopped after a fixed number of observed events and group sequential designs with a possibility to stop for efficacy after half the number of observed events.



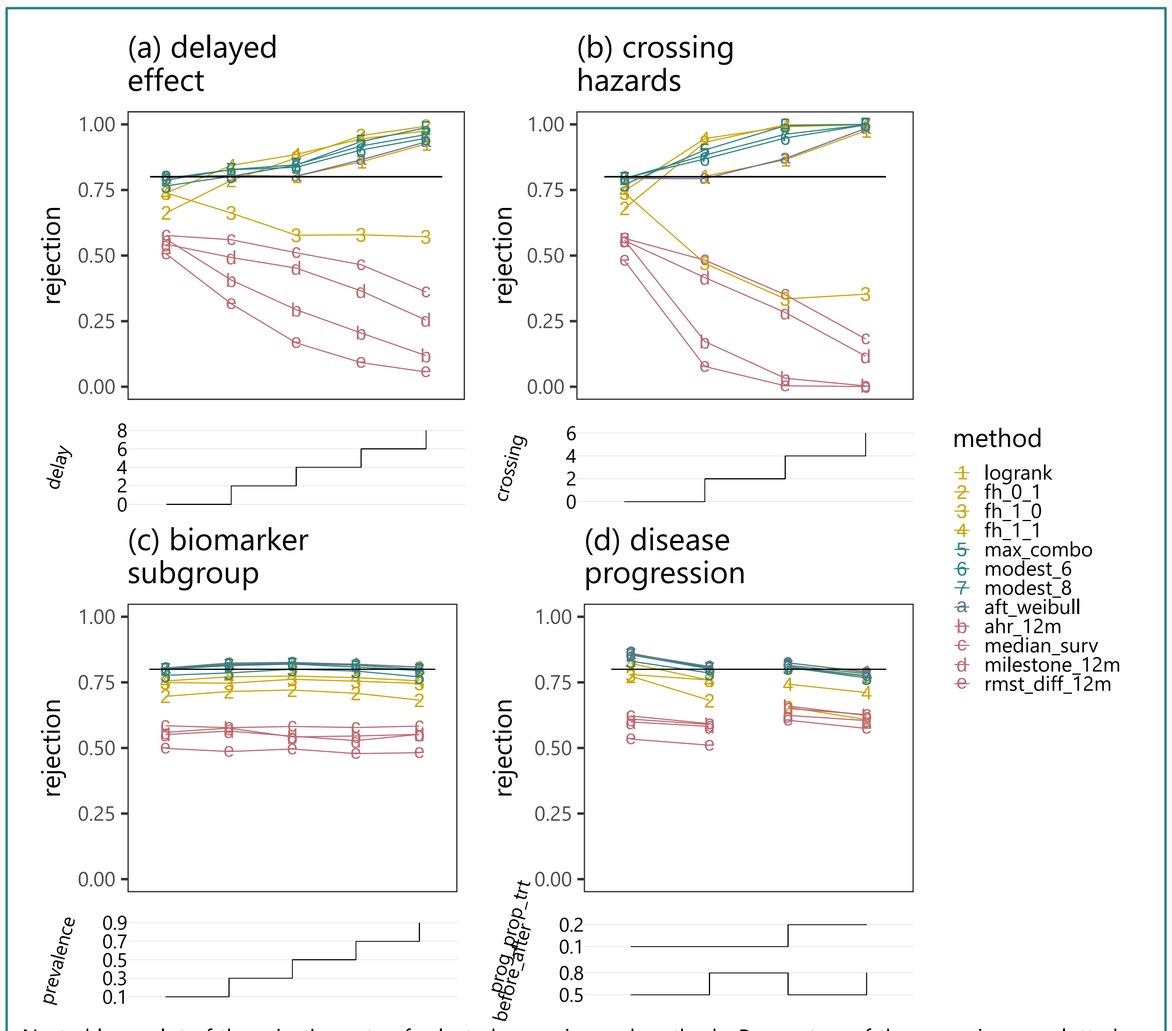
Hazard functions of the simulated scenarios. (a) delayed effect: the treatment hazard is the same as the control hazard up to a certain time and then lower. (b) crossing hazards: the treatment hazard is higher than the control hazard up to a time and then lower (c) biomarker subgroup: the treatment arm consists of two groups with different constant hazard (d) disease progression: hazards for death before disease progression as well as hazard for disease progression are different in the arms, hazard for death after disease progression is equal.

Investigated statistical methods were selected based on a systematic literature review conducted by the consortium [4] and the EPAR review. Common statistical methods for time to event data like the log-rank test, the Cox proportional hazards model and accelerated failure time models were included as reference models.

Methods were compared with respect to control of type I error rate and power under different departures form the null hypothesis. Methods estimating a summary statistic were compared with respect to their bias, variance and mean squared error and where applicable confidence interval coverage and length.

Results

To facilitate the simulations an R package [5] was developed. The package includes methods to simulate time to event distributions with piecewise constant hazards, calculate true values of summary statistics for those distributions, functions to apply statistical methods implemented in other packages to the simulated data and functions to summarize and present the simulation results. A Shiny App [6] to visualise results and data generating mechanisms of all simulations was created. A total of 10662 scenarios were simulated with 2500 replications for each scenario. Results were consistent across investigated parameter values.



Nested loop plot of the rejection rate of selected scenarios and methods. Parameters of the scenarios are plotted on the x-axis and rejection rate on the y-axis. The black horizontal lines show the asymptotic rejection rate of the log-rank test for the same median survival in the treatment arm under proportional hazards. The step functions below the plot indicate which parameter level the values above correspond to.

While all methods controlled the type I error rate the methods showed considerably different power. The log-rank test performed well under small deviations from proportional hazards. On larger deviations the MaxCombo test and the modestly weighted log-rank test performed better. The weighted log-rank tests with weights appropriate for the type of PH-violation performed best.

Non-parametric estimates of summary statistics of the survival distribution were adequately unbiased but showed lower power than testing procedures. Corresponding (semi-)parametric estimates showed larger bias but larger power and therefore need to be considered with care.

Conclusions

Method Performance

- Performance of the statistical methods is strongly impacted by strength of PH violation.
- Type I error is controlled for the strong H0 of equal survival curves.
- Methods with higher power have a challenging interpretation under NPH.

There is **no universally best statistical method** to address NPH.

- Investigators need to balance between the most suitable methods and an interpretable estimand
- Single summary measures are unable to capture all aspects of survival functions under NPH. Report multiple suitable summary measures.

Understanding and addressing NPH is a **multi-disciplinary exercise**.

- Methods and summary measures should be chosen taking into account the expected characteristics of the survival functions, such as the timing of possible delayed effects.
- Statistical methods might be the answer, but not in all cases.

Acknowledgements

This work has received funding from the European Medicines Agency (Re-opening of competition EMA/2020/46/TDA/L3.02 (Lot 3)) This document expresses the opinion of the authors of the paper, and may not be understood or quoted as being made on behalf of or reflecting the position of the European Medicines Agency or one of its committees or working parties.

Literature and Links

[1] CONFIRMS consortium, "Review of EMA EPARS where nonproportional hazards were identified" 2022
[2] CONFIRMS consortium, "Protocol: A Simulation Study to Evaluate the Performance Characteristics of Statistical Methods for the Analysis of Time-To-Event Data under Non-Proportional Hazards," 2022. Available: https://www.encepp.eu/encepp/openAttachment/fullProtocol/49769
[3] F. Klinglmüller, "A neutral comparison of statistical methods for time-to-event analyses under non-proportional hazards", 2022
[4] M. F. Bardo et al., "Methods for non-proportional hazards in clinical trials: A systematic review", arXiv (Cornell University), Jun. 2023, doi: https://doi.org/10.48550/arxiv.2306.16858
[5] T. Fellinger, "SimNPH: Simulate Non-Proportional Hazards." https://simnph.github.io/SimNPH/ (accessed Sep. 06, 2023).
[6] T. Fellinger, "Visualise Simulation Results." https://sny.cemsiis.meduniwien.ac.at/~mp314/rsnph/ (accessed Sep. 06, 2023).

www.ages.at