Paper AS-07

# Machine-Learning Approaches to Survival Analysis of Clinical Data

Dorothee Childs, HMS Analytical Software, Heidelberg, Germany
Christoph Bergen, HMS Analytical Software, Heidelberg, Germany

Survival analysis (also referred to as time-to-event analysis) is a specialized area in statistics with a range of applications in clinical and non-clinical use cases. It enables posing supervised learning problems to model the probability of an event over time, such as tumor progression. In clinical studies, patients are monitored for an intended time period, and events observed in this period are recorded. This results in censored data, where unobserved events can mean that either the event actually did not occur, or that the observed instance could not be monitored until the end of the observation period. Consequently, survival analysis requires models that take this unique data characteristic into account.

The standard modeling approach to analyse time-to-event data is Cox-Proportional Hazards (Cox-PH) regression, which is based on a linear model and a hazard function. Strengths of the Cox-PH are the interpretability of the model coefficients and the possibility to assess the influence of individual covariates on overall hazard ratios. However, it assumes linearity and requires estimation of the baseline hazard in order to perform predictions and to derive associated performance measures.

To address these points, several machine learning algorithms have been adapted to survival analysis problems and implemented as open-source tools in recent years. The scikit-survival package is a Python module for survival analysis by a variety of ML models. Our experience with scikit-survival includes models such as:

- Tree-based learners like random survival forests (RSF)
- Kernel methods like survival support vector machines (SSVM)
- Boosted Trees.

The package also provides Cox-PH and regularized Cox-PH (with L1, L2 or elastic net penalty) implementations with the possibility to perform predictions and performance comparisons.

Incorporation into the popular Python machine learning package scikit-learn enables straight-forward implementation of model selection, cross-validation, performance metrics, and an interface to the SHAP-package which is a standard tool for explainable AI techniques. This improves the explainability of black box models, a model property that is especially important in the clinical space.

## REFERENCES

Fasching P.A. et al. Abstract P4-01-05: Machine learning to predict treatment response and tolerability in HR+, HER2– advanced breast cancer: German study AI4ANNA. *Cancer Res 1* March 2023; 83 (5_Supplement): P4–01–05.

James G., Witten D., Hastie T., Tibshirani R., Taylor J. An Introduction to Statistical Learning with Applications in Python. Springer Texts in Statistics (STS), 2023.

Pölsterl S.: scikit-survival: A Library for Time-to-Event Analysis Built on Top of scikit-learn, *Journal of Machine Learning Research 21* 1-6, 2020.

Pedregosa et al.: Scikit-learn: Machine Learning in Python, JMLR 12, pp. 2825-2830, 2011.

## CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the authors at:

Dr. Dorothee Childs
HMS Analytical Software GmbH, Grüne Meile 29, 69115 Heidelberg
Email: dorothee.childs@analytical-software.de

Christoph Bergen
HMS Analytical Software GmbH, Grüne Meile 29, 69115 Heidelberg
Email: christoph.bergen@analytical-software.de

Brand and product names are trademarks of their respective companies.