

Comparing Time to Event Models for Sparse High Dimensional Medical Questionnaire Data of Rheumatoid Arthritis.

Gustaf Randén*

June 2026

Abstract

In this study, three survival models were used to predict the likelihood to develop Rheumatoid Arthritis. The methods compared were the Cox regression, Weibull regression and Random Survival Forest. They were tested on a Questionnaire study answered by 79 individuals. The answers were stored in a matrix with 475 variables for each individual. Pre-processing the data included usage of principal component analysis. The prediction models were validated by the Concordance index, Brier's score, a Confusion matrix and the Receiver Operating Characteristic curve. The results indicated that Cox and Weibull models could more accurately predict if a specific person would develop rheumatoid arthritis. However the Random Forest had the highest specificity, which means that it could predict which patient would not develop arthritis. The Random Forest model could more accurately rank who is more likely to develop the disease in three years even if their probability estimates were worse.

*Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden.
E-mail: gus.randen@gmail.com. Supervisor: Chun-Biu Li.