Development of prediction models using competing risk models in big healthcare databases

Constantinos Koshiaris[1], Richard Stevens[1], Richard Riley[2], Sarah Lay-Flurrie[1], Kym Snell[2], Lucinda Archer[2], James Sheppard[1]

[1] Nuffield Department of Primary Care Health Sciences, University of Oxford, Radcliffe Primary Care Building, Radcliffe Observatory Quarter, Woodstock Rd, Oxford OX2 6GG
[2] Centre for Prognosis Research, School of Primary, Community and Social Care. Keele University, Staffordshire. ST5 5BG

Background

The Cox proportional hazards model is a commonly used method when developing prognostic prediction models using time-to-event data. In the presence of competing risks - events that might prevent the occurrence of the event of interest – using a Cox model leads to predicted probabilities that are too high. Thus methods that account for competing risks, such as the Fine-Gray model, are preferred. However, fitting this model is computationally complex, particularly when used in combination with multiple imputation and fractional polynomials. This poses a significant challenge when developing prediction models in big databases.

Aims

To describe prediction modelling approaches that minimise computation time, without compromising model validity, in a large dataset of electronic health records.

Data

Data from the Clinical Practice Research Datalink were used to create prognostic models for adverse events related to antihypertensive medication, treating death as a competing event (prevalence 10%). The dataset included 1,773,224 patients and 40 predictors.

Methods/Results

A multivariable competing risk model developed using the stcrreg command in STATA 16 (8 cores) required approximately two weeks to converge using an 8 core 32GB, i9 PC. Computation time was reduced to less than one day when estimating regression coefficients using the R package fastcmprsk, which uses a forward backward scan algorithm: this is more efficient than the Newton-Raphson method. Robust bootstrap confidence intervals were estimated using the percentile method. Fractional polynomial transformations were computationally prohibitive, thus variable transformations were modelled with the use of Cox regression, providing a good approximation of the relationship.

Conclusions

Computational obstacles to correctly account for competing risks in clinical prediction models can be overcome by combining fast algorithms, robust bootstraps and approximate fractional polynomials transformations. We are currently investigating how to optimise the use of the Fine-Gray model in conjunction with multiple imputation.

Keywords

Prediction, competing risks, big data