Survival analysis of biobank data: some methodological aspects and examples

Krista Fischer and Kristi Läll

University of Tartu, Estonia, Krista.Fischer@ut.ee, Kristi.Lall@ut.ee

Keywords: survival analysis, nested case-control studies

Recent decades have seen a considerable increase in the availability of data from large prospective biobank cohorts. As the follow-up time increases, analysis of overall survival or disease incidence becomes more feasible and may provide valuable new information on disease and mortality risks in general population. While in conventional epidemiological cohort studies one is mainly interested in the effects of various lifestyle and/or clinical characteristics, the biobank-based studies are often focused on *omics*-based parameters, aiming for discovery of novel predictive biomarkers.

In such studies, several methodological aspects need to be considered. The first issue is the timescale to be used in the analysis. For instance, the DNA-based variables (data on single nucleotide polymorphisms, SNPs) are fixed before birth of the individual and may influence the entire lifespan. As the biobank cohorts have recruited people of different ages, the data is both right-censored and left-truncated. When the DNA-based markers are combined with other potential predictors that have measured at recruitment and, contrary to the DNA, vary in time, the question of proper adjustment arises. We discuss different options and use simple simulations to illustrate how the results would differ when different timescale is used in such analysis.

The second issue is related to the large size of the datasets. The use of Cox proportional hazards model in genome-wide analysis of several million markers in large cohorts ($n > 10000$) is complicated due to the relative slowness of the conventional fitting algorithm. We propose a two-stage algorithm based on martingale residuals to overcome this problem and implement this to identify survival-related genetic variants in the UK Biobank cohort [1].

Finally, we discuss the potential of nested case-control design in cases where molecular data can only be obtained for a limited subset of the cohort, using some examples based on the Estonian Biobank data.

References

[1] Joshi, P.K., Fischer, K., Schraut, K.E., Campbell, H., Esko, T., Wilson, J.F. (2016). Variants near CHRNA3/5 and APOE have age- and sex-related effects on human lifespan. *Nature Communications* **7**, 11174.