CohortSurvival:
an R package for survival analysis using the OMOP CDM

Kim López-Güell[1], Marti Català[1], Danielle Newby[1], Ian Koblbauer[1], Xintong Li[1], Berta Raventós[2,3], Maria de Ridder[4], Talita Duarte-Salles[2,4], Dani Prieto-Alhambra[1,4], Edward Burn[1]

[1] Pharmaco- and Device Epidemiology, Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences (NDROMS), University of Oxford, UK; [2] Fundació Institut Universitari per a la recerca a l'Atenció Primària de Salut Jordi Gol i Gurina (IDIAPJGol), Barcelona, Spain; [3] Universitat Autònoma de Barcelona, Bellaterra (Cerdanyola del Vallès), Barcelona, Spain; [4] Department of Medical Informatics, Erasmus University Medical Center, Rotterdam, The Netherlands

Background

Survival analysis refers to a collection of statistical methods to study the time until a certain event of interest occurs. These procedures allow us to analyse, for instance, how many individuals survive after a certain time, the rate at which they experience the outcome of interest or how certain characteristics might be related to their survival [1]. Many questions in medical research involve time-to-event settings. Traditionally, one single outcome event is considered, but competing events or recurring events to be considered as well.

The Observational Medical Outcomes Partnership (OMOP) standard data model (CDM) provides a common structure for health data and numerous healthcare databases. The main advantage of OMOP CDM is that it allows the same analytic code to be efficiently run against multiple data sources. Hence each data source owner can use the same standardised analytic code for its analysis and the results can be comparable, without sharing any patient-level data. Developing reliable and transparent analytic packages can simplify and increase the analysis reliability while minimising errors from code sharing across different data sources.

There exist multiple packages to conduct survival studies in R. However, none provides an easy way to interact with data in an OMOP-formatted data. Given the abundance of survival studies in health research, a package supporting the analysis of survival data mapped to the OMOP CDM is essential to ensure reliability and reproducibility of the research.

Here we aimed to develop an R package to extract and summarise survival data using the OMOP common data model, to provide reliable and reproducible survival analysis capabilities in line with the DARWIN EU Catalogue of Standard Analytics.

Methods

CohortSurvival is written in R (version 4.2.1) and built around the established R package survival [2] for survival data and on top of the CDMConnector and PatientProfiles R packages that provide functionality for working with data mapped to the OMOP CDM [3, 4]. It allows the user to easily extract survival estimates and plot the respective Kaplan-Meier curves, for pre-defined cohorts of interest in OMOP CDM format. The package allows for both single event and competing events analyses. The survival estimates outputs can also be stratified by variables of interest. Age groups or sex stratification can be easily implemented using the specific function arguments, but any other variable pre-defined by the user for stratification is also possible.

As well as providing survival probabilities, the package also generates summary survival information such as restricted mean survival and median survival. In addition, the characteristics of the population of interest can also be summarised.

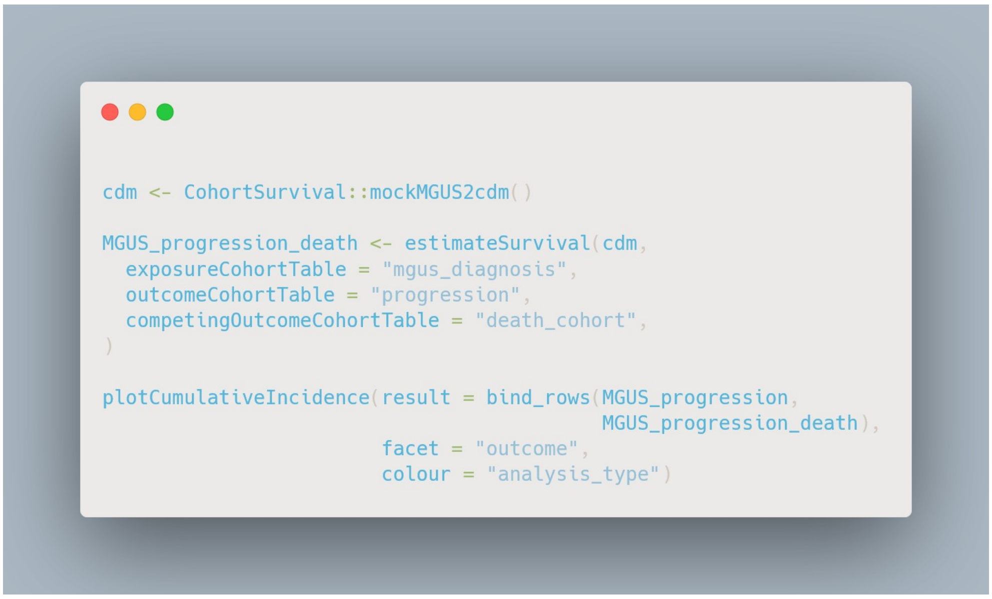
The package has been developed in collaboration with target users to provide the required functionality and intuitive interfaces. The package was developed with thorough testing, with tests of input/output formats, various possible combinations of use cases, logical checks, expected true and expected false outputs and edge cases. This development followed the Agile principles and DARWIN EU Quality Assurance of Software Development.

Results

The package is freely available under the Apache License (Version 2.0) and can be obtained from GitHub, https://github.com/darwin-eu-dev/CohortSurvival. Detailed vignettes on the package's functionality are also freely available online from the GitHub repository. We expect to have released the package on CRAN by the time of the OHDSI conference.

Example code for the package can be found below, applied against the MGUS2 data adapted from the survival package to the OMOP CDM format. This data contains survival data of 1341 sequential patients with monoclonal gammopathy of undetermined significance. In this example we have data on the diagnosis, progression and death of patients with MGUS.

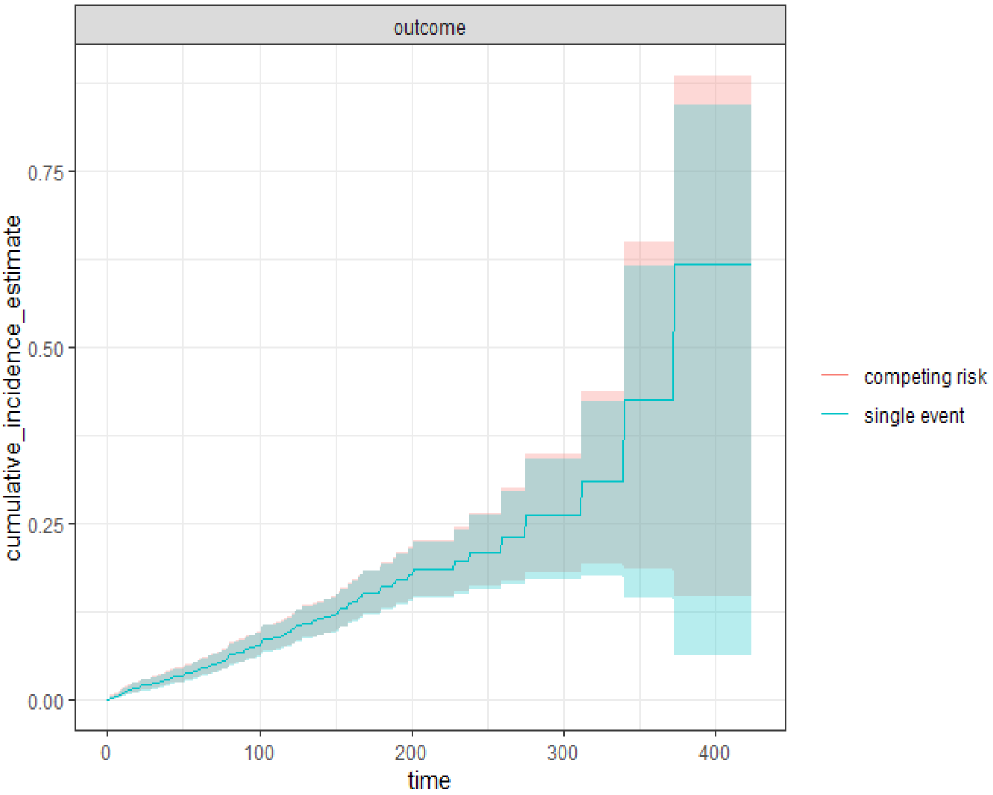
Example code snippet: survival estimates for progression following diagnosis, with competing event of risk of death



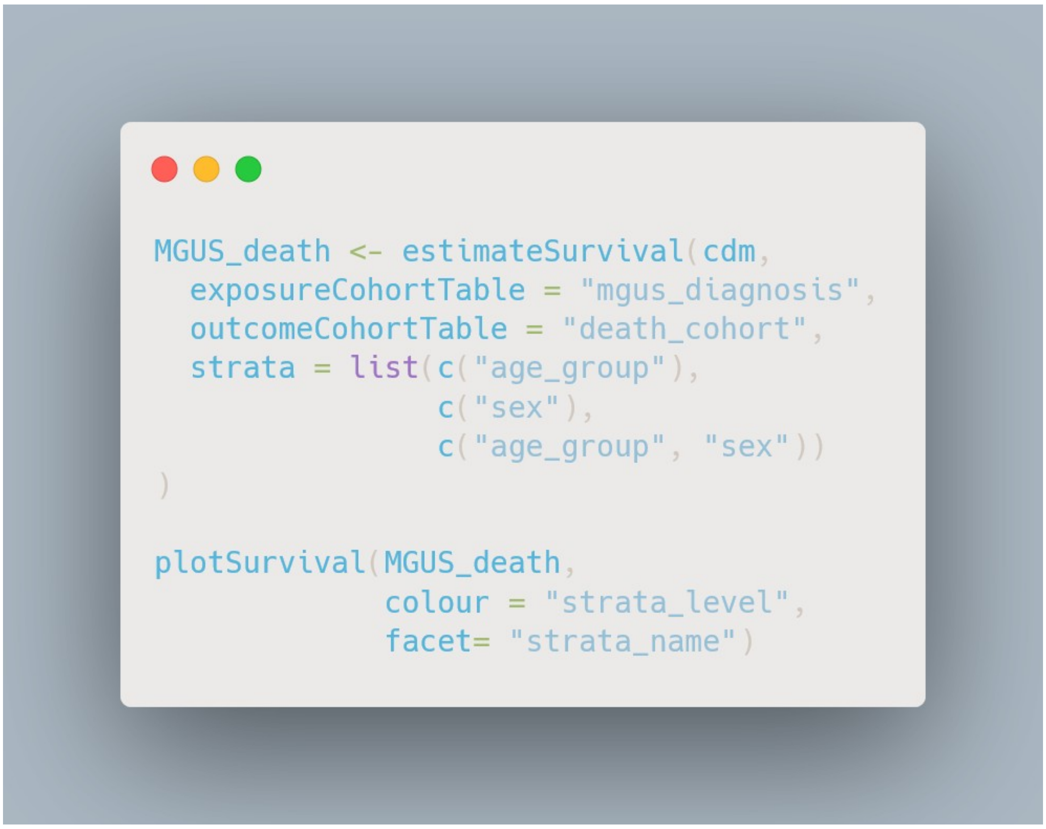
```
cdm <- CohortSurvival::mockMGUS2cdm()

MGUS_progression_death <- estimateSurvival(cdm,
  exposureCohortTable = "mgus_diagnosis",
  outcomeCohortTable = "progression",
  competingOutcomeCohortTable = "death_cohort",
)

plotCumulativeIncidence(result = bind_rows(MGUS_progression,
                                           MGUS_progression_death),
                        facet = "outcome",
                        colour = "analysis_type")
```

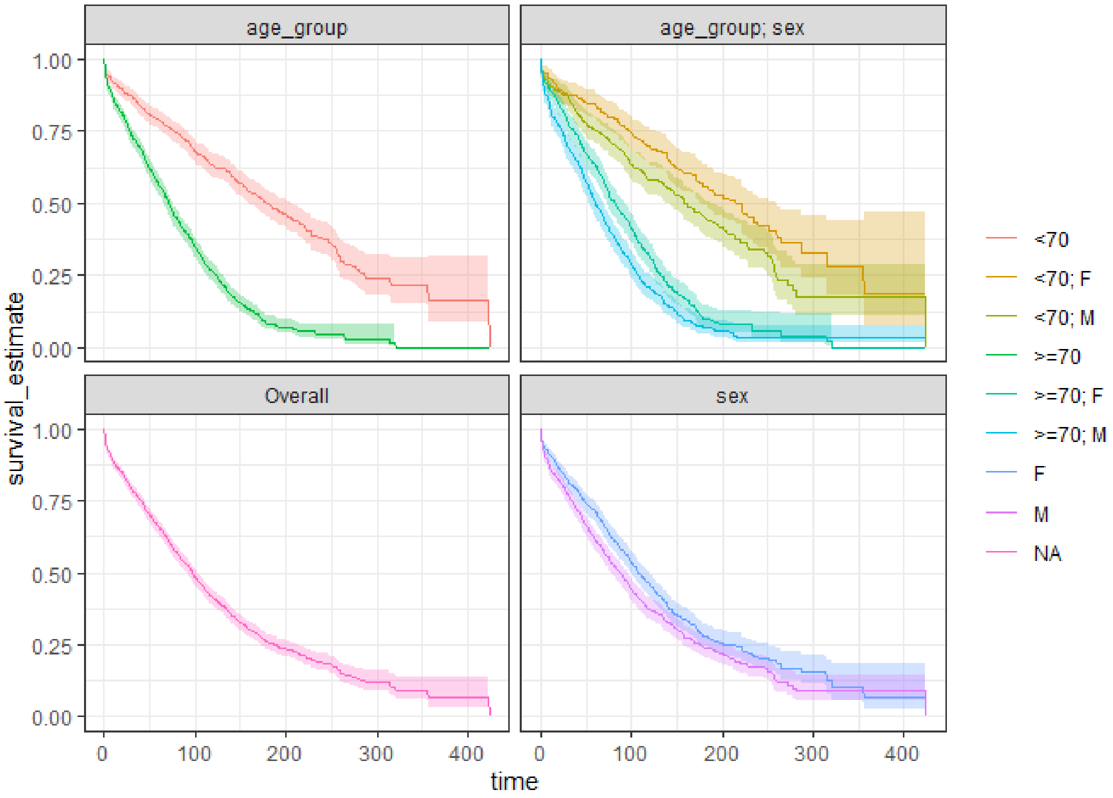


Example code snippet: stratified results by age group and sex for survival estimates for death after diagnosis



```
MGUS_death <- estimateSurvival(cdm,
  exposureCohortTable = "mgus_diagnosis",
  outcomeCohortTable = "death_cohort",
  strata = list(c("age_group"),
                c("sex"),
                c("age_group", "sex"))
)

plotSurvival(MGUS_death,
             colour = "strata_level",
             facet= "strata_name")
```



Conclusions

The R package CohortSurvival provides functionality for generating survival data from cohorts defined in the OMOP standard data model. It is helpful tool for researchers utilising OMOP CDM interested in answering descriptive survival questions and provides flexibility for extension for users interested in performing more complex analyses (e.g., regression modelling, multi-state survival models, and so on). Future releases of the package will include additional functionality, such as log-rank tests.

References

1. Altman D G, Bland J M. Time to event (survival) data BMJ 1998; 317 :468 doi:10.1136/bmj.317.7156.468

2. Therneau T (2023). *A Package for Survival Analysis in R*. R package version 3.5-5, https://CRAN.R-project.org/package=survival.

3. Catala M, Guo Y, Du M, Lopez-Guell K, Burn E (2023). *PatientProfiles: Identify Characteristics of Patients in the OMOP Common Data Model*. R package version 0.2.0, https://darwin-eu-dev.github.io/PatientProfiles/