Biostatistics (2022), **0**, 0, *pp.* **??**–8
doi:10.1093/biostatistics/multiglmm

A multinomial generalized linear mixed model for clustered competing risks data

HENRIQUE APARECIDO LAUREANO*

Instituto de Pesquisa Pelé Pequeno Príncipe, Curitiba, Brasil

RICARDO HASMUSSEN PETTERLE

Departamento de Medicina Integrada, Universidade Federal do Paraná, Curitiba, Brasil

GUILHERME PARREIRA DA SILVA, PAULO JUSTINIANO RIBEIRO JUNIOR,

WAGNER HUGO BONAT

Laboratório de Estatística e Geoinformação, Departamento de Estatística, Universidade Federal do Paraná, Curitiba, Brasil

henriqueaparecidolaureano@gmail.com

Summary

Clustered competing risks data are a complex failure time data scheme. Its main characteristics are the cluster structure, which implies a latent within-cluster dependence between its elements, and its multiple variables competing to be the one responsible for the occurrence of an event, the failure. To handle this kind of data, we propose a full likelihood approach, based on generalized linear mixed models instead the usual complex frailty model. We model the competing causes in the probability scale, in terms of the cumulative incidence function (CIF). A multinomial distribution is assumed for the competing causes and censorship, conditioned on the latent effects that are accommodated by a multivariate Gaussian distribution. The CIF is specified as the

*To whom correspondence should be addressed.

product of an instantaneous risk level function with a failure time trajectory level function. The estimation procedure is performed through the R package TMB (Template Model Builder), an `C++` based framework with efficient Laplace approximation and automatic differentiation routines. A large simulation study was performed, based on different latent structure formulations. The model fitting was challenging and our results indicated that a latent structure where both risk and failure time trajectory levels are correlated is required to reach reasonable estimation.

Key words: Cause-specific cumulative incidence function; Within-cluster dependence; Template Model Builder; Laplace approximation; Automatic differentiation.

1. Introduction

Competing risks data, and more generally failure time data, can be modeled in two possible scales: the hazard and the probability scale, with the former being the most popular. A competing risks process can be seen as the multivariate extension of a failure time process, having multiple causes competing to be the one responsible for the desired event occurrence, properly, a failure. In Figure 1 a visual aid is provided considering m competing causes, where zero represents the initial state.

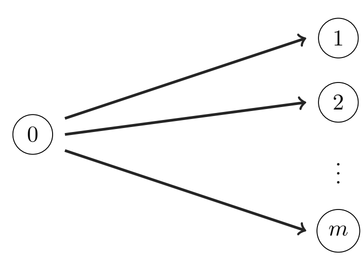

Fig. 1. Illustration of competing risks process.

Failure time data is the branch of Statistics responsible to handle random variables describing the time until the occurrence of an event, a failure (Kalbfleisch and Prentice, 2002; Hougaard, 2000). The time until a failure is called survival experience, and is the modeling object. To accommodate the number of possible causes for a failure there is the competing risks data scheme. More specifically, its clustered version with groups of elements sharing some non-observed latent dependence structure.

When this framework is applied in real-world situations, we have to be able to handle with the nonoccurrence of the desired event, by any of the competing causes, for, let us say, *logistic reasons* (short-time study and outside scope causes are some examples). This, generally noninformative, nonoccurrence of the event is called censorship.

When the elements under study are organized in clusters (a family, e.g.), it opens space to what is called *family studies*. In family studies, the goal is to accommodate the non-observed latent dependence and try to understand the relationship between the family elements. In other words, how the occurrence of an event in a subject affects the survival experience for the same or similar event.

The survival experiences is usually modeled in the hazard (failure rate) scale, and with the latent within-cluster dependence accommodation we have what is called a frailty model (Clayton, 1978; Valpel *and others*, 1979; Liang *and others*, 1995; Petersen, 1998). The use of frailty models implies in complicated likelihood functions and inference routines done via elaborated and slow EM algorithms (Nielsen *and others*, 1992; Klein, 1992) or inefficient MCMC schemes (Hougaard, 2000). With multiple survival experiences, the general idea is the same but with even more elaborated likelihoods (Prentice *and others*, 1978; Therneau and Grambsch, 2000) or mixture model approaches (Larson and Dinse, 1985; Kuk, 1992).

When in the hazard scale, the interpretations are in terms of hazard rates. A less usual scale but with a more appealing interpretation is the probability scale. For competing risks data, the work on the probability scale is done by means of the cumulative incidence function (CIF) (Andersen *and others*, 2012), with the main modeling approach being the subdistribution (Fine and Gray, 1999).

For clustered competing risks data there are some available options but with a lack of predominance. The options vary in terms of likelihood specification, with its majority being designed for bivariate CIFs, where increasing the CIF's dimension is a limitation. Some of the existing options are (i) nonparametric approaches (Cheng *and others*, 2007, 2009); (ii) linear transformation models (Fine, 1999; Gerds *and others*, 2012); (iii) semiparametric approaches based on composite likelihoods (Shih and Albert, 2009; Cederkvist *and others*, 2019), estimating equations (Scheike and Sun, 2012; Cheng and Fine, 2012), copulas (Scheike *and others*, 2010), or mixtures (Naskar *and others*, 2005; Shi *and others*, 2013).

Besides the interpretation, by modeling the CIF it is possible to specify complex within-cluster dependence structures. We follow Cederkvist *and others* (2019) and work with a CIF specification based on its decomposition in instantaneous risk and failure time trajectory functions, with both being cluster-specifics and possible correlated. As a modeling framework, we use a generalized linear mixed model (GLMM) specification. Through a GLMM we have a straightforward full likelihood specification, easy to virtually extend to any number of competing causes, and capable to allow for complex CIF structures. To make the estimation and inferential process the most efficient as possible we take advantage of state-of-art computational libraries and efficiently implemented routines under the `TMB` (Kristensen *and others*, 2016) package of the R (R Core Team, 2021) statistical software.

The class of generalized linear models (GLMs) (Nelder and Wedderburn, 1972) is probably the most popular statistical modelling framework. Despite its flexibility, the GLMs are not suitable for dependent data. For the analysis of such data, Laird and Ware (1982) proposed the random effects regression models for longitudinal/repeated-measures data, and Breslow and Clayton (1993) presented the GLMMs for the analysis of non-Gaussian outcomes. In this framework, we can accommodate all competing causes of failure and censorship under a multinomial probability distribution. The latent within-cluster dependence is accommodated via a multivariate normal distribution, and the cause-specific CIFs via the model's link function.

The main goal of this work is to propose a GLMM approach to handle clustered competing risks data with a flexible within-cluster dependence structure. The model specification and the inferential routine are much simpler than the usually used approaches, increasing its practical relevance. The latent effects, the key complicator factor, are handled out by means of an efficient Laplace approximation and automatic differentiation routines. The main contributions of this article are: (i) introducing the modeling of cause/cluster-specific CIFs of clustered competing risks data into an efficient implementation of the GLMMs framework; (ii) performing a extensive simulation study to check the properties of the maximum likelihood estimator to learn the cause-specific CIF forms and the feasibility of the within-cluster dependence structure.; (iii) providing R code and `C++` implementation for the used GLMMs.

References

Andersen, P. K., Geskus, R. B., de Witte, T. and Putter, H. (2012). Competing risks in epidemiology: possibilities and pitfalls. *International Journal of Epidemiology* **31**(1), 861–870.

Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association* **88**(421), 9–25.

Cederkvist, L., Holst, K. K., Andersen, K. K. and Scheike, T. H. (2019). Modeling the cumulative incidence function of multivariate competing risks data allowing for within-cluster dependence of risk and timing. *Biostatistics* **20**(2), 199–217.

Cheng, Y. and Fine, J. P. (2012). Cumulative incidence association models for bivariate competing risks data. *Journal of the Royal Statistical Society, Series B (Methodological)* **74**(2), 183–202.

Cheng, Y., Fine, J. P. and Kosorok, M. R. J. (2007). Nonparametric Association Analysis of Bivariate Competing-Risks Data. *Journal of the American Statistical Association* **102**(480), 1407–1415.

Cheng, Y., Fine, J. P. and Kosorok, M. R. J. (2009). Nonparametric Association Analysis of Exchangeable Clustered Competing Risks Data. *Biometrics* **65**(1), 385–393.

Clayton, D. G. (1978). A model for association in bivariate life tables and its application in epidemiological studies of familial rendency in chronic disease incidence. *Biometrika* **65**(1), 141–151.

Fine, J. P. (1999). Analysing competing risks data with transformation models. *Journal of the Royal Statistical Society, Series B (Methodological)* **61**(4), 817–830.

Fine, J. P. and Gray, R. J. (1999). A proportional hazards models for the subdistribution of a competing risk. *Journal of the American Statistical Association* **94**(446), 496–509.

Gerds, T. A., Scheike, T. H. and Andersen, P. K. (2012). Absolute risk regression for competing risks: interpretation, link functions and prediction. *Statistics in Medicine* **31**(29), 3921–3930.

Hougaard, P. (2000). *Analysis of Multivariate Survival Data*. New York: Springer-Verlag.

Kalbfleisch, J. D. and Prentice, R. L. (2002, February). *The Statistical Analysis of Failure Time Data*, Second Edition edition. Hoboken, New Jersey: John Wiley & Sons, Inc.

Klein, J. P. (1992). Semiparametric estimation of random effects using cox model based on the em algorithm. *Biometrics* **48**(1), 795–806.

Kristensen, K., Nielsen, A., Berg, C. W., Skaug, H. J. and Bell, B. M. (2016). TMB: Automatic Differentiation and Laplace Approximation. *Journal of Statistical Software* **70**(5), 1–21.

KUK, A. Y. C. (1992). A semiparametric mixture model for the analysis of competing risks data. *Australian Journal of Statistics* **34**(2), 169–180.

LAIRD, N. M. AND WARE, J H. (1982). Random-effects models for longitudinal data. *Biometrics* **38**(4), 963–974.

LARSON, M. G. AND DINSE, G. E. (1985). A Mixture Model for the Regression Analysis of Competing Risks Data. *Journal of the Royal Statistical Society, Series C (Applied Statistics)* **34**(3), 201–211.

LIANG, K. Y., SELF, S., BANDEEN-ROCHE, K. J. AND ZEGER, S. L. (1995). Some recent developments for regression analysis of multivariate failure time data. *Lifetime Data Analysis* **1**(1), 403–415.

NASKAR, M., DAS, K. AND IBRAHIM, J. G. (2005). A Semiparametric Mixture Model for Analyzing Clustered Competing Risks Data. *Biometrics* **61**(3), 729–737.

NELDER, J. A. AND WEDDERBURN, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society, Series A* **135**(3), 370–384.

NIELSEN, G. G., GILL, R. D., ANDERSEN, P. K. AND SØRENSEN, T. I. A. (1992). A Counting Process Approach to Maximum Likelihood Estimation in Frailty Models. *Scandinavian Journal of Statistics* **19**(1), 25–43.

PETERSEN, J. H. (1998). An Additive Frailty Model for Correlated Life Times. *Biometrics* **54**(1), 646–661.

PRENTICE, R. L., KALBFLEISCH, J. D., PETERSON JR, A. V., FLOURNOY, N., FAREWELL, V. T. AND BRESLOW, N. E. (1978). The analysis of failure times in the presence of competing risks. *Biometrics* **1**(1), 541–554.

R Core Team. (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/.

Scheike, T. and Sun, Y. (2012). On cross-odds ratio for multivariate competing risks data. *Biostatistics* **13**(4), 680–694.

Scheike, T., Zhang, Y. Sun M. and Jensen, T. K. (2010). A semiparametric random effects model for multivariate competing risks. *Biometrika* **97**(1), 133–145.

Shi, H., Cheng, Y. and Jeong, J. H. (2013). Constrained parametric model for simultaneous inference of two cumulative incidence functions. *Biometrical Journal* **55**(1), 82–96.

Shih, J. H. and Albert, P. S. (2009). Modeling Familial Association of Ages at Onset of Disease in the Presence of Competing Risk. *Biometrics* **66**(4), 1012–1023.

Therneau, T. M. and Grambsch, P. M. (2000). *Modeling Survival Data: Extending the Cox Model*. New York: Springer-Verlag.

Valpel, J. W., Manton, K. G. and Stallard, E. (1979). The impact of heterogeneity in Individual Frailty on the Dynamics of Mortality. *Demography* **16**(1), 439–454.

[*Received March 18, 2022; revised ; accepted for publication*]