Long-Term Survival for Competing Risk Data with Masked Causes

Ronny Westerman[1]
[1]Institute of Medical Sociology and Social Medicine,
Medical School & University Hospital
Philipps-University of Marburg
Karl-von-Frisch-Str.4, 35043 Marburg, Germany

Abstract

Competing Risks Models have a various field of application in medical and public health studies. A challenging clue for applying cause-specific survival models yield on the problem of missing and misclassification in cause of death.
The masked cause of death is related to incomplete or only partial identifiable information of death certificates. Different Bayesian approaches e.g. the mixture cure model are proposed to account for that problem. Another question is related to adequate estimates for long-term survival in respect to the limitation of lifetime among all risks. As a new parametric distribution the long-term exponential distribution (LEG) introduced by Roman et al. 2012 can be considered. The main purpose of this work is to compare the LEG with alternative parametric versions like Weibull distribution, or the simple Exponential distribution for long-term survival estimates. Data analysis will be realized with Cancer Register Data (SEER) and R Statistical Software. As on remarkable conclusion one would expect the best fitting of the LEG for the long-term survival regarding to Weibull and Exponential distribution.

Key Words: Competing Risks, Masked Causes, Long-term Survival
.

1. Background

Many different sophisticated statistical approaches have been applied for Competing Risk data most in last 20 years (i.e. Gasbarra and Karia, 2000, Salinas-Tores et al. 2002, Craiu and Duchesne, 2004 or Pintilie, 2006). Every single approach has preluded some methodological advances and also maintains initial drawbacks in case of masking causes and which could not being exactly answered until right now. In terms of cause-specific survival each subject being exposed to many competing risks, but only one will be caused the failure.
A masking cause situation turns out if the cause of the event for some of units or individuals not exactly identified or recorded (Flehinger et al. 2002 , Craiu and Lee, 2005, Lu and Liang, 2008 Sen et al. 2010, Roman et al. 2012).
For partial masking event the cause is narrowed down but not exactly identified (Basu et al. 2003). Also the reasons for misclassification are manifold.
The information needed for attributing the cause of failure may be not collected, or the cause of diseases for some patients may be difficult to determine.
Two common situations should exemplify the problems concerning the misclassification problem.

First aetiological problems for specific diseases or not exactly clarified determination for disease-specific symptoms could result in misclassification.
As an example, cardioembolic stroke (Leary and Caplan, 2008, Abro and Alio, 2010) occurs when the heart pumps unwanted materials into the brain circulation, resulting in the occlusion of a brain blood vessel and damage to the brain tissue.
Cardioembolic strokes are diagnosed in 3-8% stroke patients, but in various current stroke registries, approximately 10-20% with CS have not maximal symptoms at the onset of their stroke (Leary and Caplan, 2008). In that situation stroke cases only fulfilling a few symptoms will be often excluded from the stroke classification system or will be often set to unknown causes. This not even rare process will be often justify or yield with arguments for simplifying diagnostic criteria, but it will not provide the exact numbers.
Second misclassification will be also proven by stage migration in case of the improved detection of illness leading to movement of people from the set of healthy people to the set of unhealthy people known as Will-Rogers phenomena (Feinstein et al. 1985). For different types of cancer specifically for breast cancer changes in the interpretation of classification schemes can alter the apparent distribution of cancer stage or grade in the absence of a true biologic change (Albertsen et al. 2005).
To overcome that specific problems many different approach were applied to account for.

Multiple imputations are widely used for the analysis of incomplete data or uncertainty in reliability for partial information. As an option MI are also likely used for modelling competing risk with missing cause of failure. (Lu and Tsaitis, 2001, Bakoyannis et al. 2010, Lee at al. 2011). Excepting all the limitations of that procedures MI are appropriated if the assumed baseline functions are not proportional and the considered missing causes can be treated as missing at random. Otherwise for high-morbid cases or multiple-specific mortality risks e.g. in elderly population the assumption for missing at random should be beyond the reality.
Second-stage analysis (Flehinger et al. 2002) uses the information from the second autopsy in case of missing information from the first one to provide the identification of the cause being responsible for the failure. Reliability will be hit by assuming independence between the probabilities of choosing the definite identification for the masked case at second stage and compared to the observed information of first stage. In order that the data requirements need compelling qualitative information for the cause of failure or cause of death often only can be determine by an expert or good trained pathologist.
These examples should illustrate the advances of modeling competing risks in case of masked causes. Mixture long-term survival models are sophisticated to account for that specific problem, but also their methodological limitations are still on discussion.
For simplification these models consider one specific clue. If the information of the responsible component failure is missing, only a minimum of lifetime among all risks can observed, because a part of the population is not susceptible to the event of interest.

2. Methods

2.1 Mixture Long-term Survival Model and LEG-distribution

Assuming units of individuals may be not susceptible to an certain event of interest, it's possible to apply a two components mixture model (Maller and Zhou, 1996, Roman et al. 2012), in the sense that one component will representing the survival time or failure of susceptible individuals (in risk individuals-IR), then the other component will representing the not susceptible individuals to the event (out of risk individuals – OR), under the condition of infinite survival times for this group.
The model formulation can be described as following.
Let Y be a random variable representing the time until the event of interest is likely to occur. For the considered population there exists a probability of cure p , the population survival can be formulated by following Maller and Zhou (1996),
$S(y) = pS_{OR}(y) + (1-p)S_{IR}(y)$ where $S_{OR}(y)$ and $S_{IR}(y)$ are survival function of the individuals OR and IR.
The event of interest shall not occur in the group OR, the failure times are infinite, so $S_{OR}(y) = P(Y > y|OR) = 1, \forall y \geq 0$. Then $S(y)$ can be rewritten as,

$$S(y) = p + (1-p)S_{IR}(y) \quad (1)$$

All susceptible individuals IR shall present the event of interesting at the same time, that is $\lim_{y\to\infty} S_{IR} = 0$, then for the not susceptible individuals the event of interesting should not occur with $\lim_{y\to\infty} S_{OR} = p$. The survival function is not conditional and correspond to the individual proportion OR. Consequently one should follow the latent competing risk scenario, because the event of interest is also caused by an unknown competing cause of failure (Louzada-Neto, 1999, Roman et al. 2012).
For the unobserved number of causes of the event M the probability mass function is
$P(M = m)$ (2)

where $m = 1,2,\dots,M$, with M on in infinite range and $T_m, m = 1,\dots,M$as the time for the j^{th} cause to produce the event of interest. The T_j are independent but conditional on M and identically distributed with the survival function $S_0(t)$. Y is a random variable given by $Y = \min\left(T_{1,}T_{2,\dots,}T_M\right)$.
The survival function of susceptible individuals IR is given by

$$S_{IR}(y) = \sum_{m=1}^{\infty} S_0(y)^m \, P[M = m] \quad (3)$$

In Adamidis and Loukas (1998), M is geometrically distributed and T exponentially distributed, then $S_{IR}(y)$ the survival function of an EG distributed random variable is definded as

$$S_{IR}(y) = \frac{(1-\theta)e^{-\lambda y}}{1-\theta e^{-\lambda y}} \quad (4)$$

Finally the survival function of an LEG distributed nonnegative random variable can be considered by the definition given in (1) and (4)

$$S(y) = \frac{p+(1-\theta-p)e^{-\lambda y}}{1-\theta e^{-\lambda y}} \quad (5)$$

where, $y > 0, \lambda > 0, \alpha > 0, 0 < \theta < 1,$ and $0 < p < 1$.

Its pdf is considered as $f(y) = -dS(y)/dy$ and is given by

$$f(y) = \frac{\lambda e^{-\lambda y}(1-\theta-p+p\theta)}{\left(1-\theta e^{-\lambda y}\right)^2}, \quad (6)$$

where, λ is scale parameter, θ is shape parameter and p is the long-term parameter

2.2 Computation

Fitting of the LEG for the Mixture long-term Survival model will be computed with R software package 'optimx' as a replacement and extension of the optim() function, which was approved by Nash and Varadhan, 2012. With an identifiable optimization wrapper function the general purpose is to estimate the best existing optim()function. The optim() procedure allows Nelder-Mead, quasi-Newton and conjugate-gradient algorithm as well as box-constrained optimization via L-BFGS-B.
Nash and Varadhan (2011, 2012) also note that optimx is only working well for one-dimensional minimization, so the argument optimize with constrOptim or spg should be also considered for computation.
The program code for Nelder-Mead, BFGS and CG bases orginally on Pascal code Nash (1990), that is also available.
The L-BFGS-B method based on Fortran code by Zhu, Byrd, Lu-Chen and Nocedal, associated to Netlib (file 'opt/lbfgs_bcm.shar' another version is in 'toms/778').

Usage
optimx(par, fn, gr=NULL, hess=NULL, lower=-Inf, upper=Inf,
method=c("Nelder-Mead","BFGS"), itnmax=NULL, hessian=FALSE,
control=list(),
...)

For further detail information, see aslo Nash and Varadhan, 2012

2.3 Simulation Data

The mixture long-term survival approach will be applied for Breast Cancer Data provided by the SEER Cancer Statistic Data Base National Cancer Institute, DCCPS, Surveillance Research Program, and Cancer Statistics Branch was released in April 2012. Information on the incidence by race, gender and age for different period of time are available.
We use cause-specific mortality data including all cancer. In period of 1992-The SEER public use dataset on survival of breast cancer patients from 1992-2009 includes (n=69,990 in Situ).
The general purpose is to study which distribution will be providing the best performance and fitness to the SEER Breast Cancer Data. In that way the proposed LEG distribution will be compared with the long-term Exponential (LE) (the particular case of LEG) and also the long-term Weibull (LW). The Weibull distribution is comparatively supposed to fit Breast Cancer Survival Data, and it also used for many biomedical applications.
For the competing risk setting the objective consider Survival function of $S_{IR}(y)$ consider all susceptible and identified breast cancer cases, and the survival function of $S_{OR}(y)$ consider the not susceptible breast cancer including all masked cases.

3. Results

Table 1: MLEs and the standard errors for SEER Breast Cancer Data

Distribution	λ	θ	φ	p
LEG	0.0032 (0.00294)	0.9868 (0.00759)	-	0.2348 (0.1329)
LW	0.0149 (0.0225)	-	0.6249(0.1231)	0.2761 (0.1569)
LE	0.0142(0.0211)	-	-	0.3435 (0.0986)

The results data simulation with parameter estimates are provided in Table 1.
The information of AIC and BIC criterion show evidence for LEG, but in general the differences are quite low, so also the LE or LW can be used as well for the application. These results are comparative to the findings of Roman et al. applied for the model for Myelomatosis and Leukaemia Data.

Table 2:

Model	ℓ(.)	AIC	BIC
LEG	-44.09170	96.67375	102.0963
LW	-45.42845	96.37691	101.7556
LE	-45.89798	97.13586	100.8188

4. Conclusion:

In general the Survival Cure Rate Model is reliable for competing risk scenarios. With the data simulation it was shown that The LEG distribution as an extension of LE proposed by Adamidis and Loukas (1998) is sophisticated to latent competing risk setting, when only the information for a minimum of lifetime among all risks is available. On major problem need to account for further applications: The survival function of the not susceptible individuals will be treated as infinite; this assumption seems not realistic and practicable for the model application. Also this model relies only for univariate survival data.
In case of multi- or bivariate cause-specific survival data different dependence structures between variables can be suited with different copula functions (Lo and Wilke, 2009, Louzada et al. 2012). There are two main methodical aspects for the marginal distributions need to account for: first the maximum of flexibility and second the application in case of masked causes. A bivariate mixture long-term model based on the Farlie-Gumbel-Morgenstern (FGM) copula was applied by Louzada et al. (2012) this need further investigation.

Acknowledments

This work was funded by in part by Von-Behring-Röntgen- Foundation which supports and promotes the medical faculties of the Justus-Liebig-University of Giessen and the Philipps-University of Marburg in their network of life sciences and other academic field.

References

Adamidis and Loukas (1998). A lifetime distribution with decreasing failure rate. Statistics probability Letters 39, 35-42.

Albertsen P.C., Hanley J.A., Barrows G.H., Penson D.F., Kowalczyk P.D., Sanders M.M., Fine J. (2005). Prostate cancer and the Will Rogers phenomenon. Journal of National Cancer Institute

Arboix A. and Alió J. (2010). Cardioembolic stroke: clinical features, specific cardiac disorders and prognosis. Curr Cardiol Rev. 6(3):150-61.

Bakoyannis G., Siannis F., Touloumi G. (2010) Modelling competing risks data with missing cause of failure. Statistics in Medicine 29(30):3172-85

Basu, S., Sen, A., and Bannered, M. (2003). Bayesian analysis of competing risks with-partially masked cause of failure. *Applied Statistics* 52, 77-93.

Conway, D.A. (1983). Farlie-Gumbel-Morgenstern distributions. In Encyclopedia of Statistical Sciences (Edited by Kotz and N.L. Johnson), Volume 3, 28-31. Wiley, New York

Craiu, R. V. and Duchesne, T. (2004). Inference based on the EM algorithm for the competing risk model with masked causes of failure. Biometrika 91, 543-558.

Craiu R.V. and Lee T.C.M. (2005): Model Selection for the Competing-Risks Model with and without masking. Lifetime Data Anal (2006) 12:21–33

Flehinger, B. J., Reiser, B. and Yashchin, E. (2002). Parametric modeling for survival with competing risks and masked failure causes. Lifetime Data Anal. 8, 177-203.

Feinstein A.R., Sosin D.M., Wells C.K. (1985) The Will Rogers phenomenon: stage migration and new diagnostic techniques as a source of misleading statistics for survival in cancer. New Engl J Med;312:1604–8.

Gasbarra D. and Karia S. (2000) Analysis of competing risks by using Bayesian smoothing Scandinavian Journal of Statistics 27 605-617.

Salinas-Torres, V. H., Pereira, C. A. B., and Tiwari, R. C. (2002). Bayesian nonparametric estimation in a series system or a competing-risk model. Journal of Nonparametric Statistics 14, 449-458.

Leary M.C., Caplan L.R. (2008): Cardioembolic stroke: An updated on etiology, diagnosis and management. Annuals Indian Academic Neurology , 11, 52-63

Lee M., Cronin K.A., Gail M.H., Dignam J.J., Feuer E.J. (2011). Multiple imputation methods for inference on cumulative incidence with missing cause of failure. Biometrical Journa l53(6):974-93.

Lo, S.M.S. and Wilke, R.A.(2009). A copula model for dependent competing risks. Discussion Papers in Economics No.09/01 University of Nottingham.

Louzada-Neto F. (1999). Polyhazard model for lifetime data. Biometrics 55, 1281-1285.

Louzada, F., Suzuki, A.K., Cancho, V.G., Prince F.L. and Pereira, G.A. (2012) The Long-term Bivariate. Survival FGM Copula Model: An application to a Brazilian HIV Data. Journal of Data Science 10, 511-535.

Lu K. and Tsiatis A.A. (2001) Multiple imputation methods for estimating regression coefficients in the competing risks model with missing cause of failure. Biometrics 54, 1191-1197.
Lu W. and Liang Y. (2008): Analysis of competing risks data with missing cause of failure under additive hazard model. Statistica Sinica 18, 219-234.
Nash J.C. (1990) Compact Numerical Methods for Computers. Linear Algebra and Function Minimisation. Adam Hilger.
Nash J.C. and Varadhan R. (2011) Unifying Optimization Algorithms to Aid Software System Users: optimx for R., Journal of Statistical Software, 43(9), 1-14., URL http://www.jstatsoft.org/v43/i09/.
Nash J.C. and Varadhan R. (2012): Package 'optimx'. A Replacement and Extension of the optim() Function. (April, 22, 2012)
http://cran.r-project.org/web/packages/optimx/optimx.pdf
Roman M., Louzada F., Cancho V.G.. and Leite J.G. (2012): A New Long-Term Survival Distribution for Cancer Data. Journal of Data Science 10, 242-258
SEER Research Data 1973-2009. Surveillance, Epidemiology, and End Results (SEER) Program (www.seer.cancer.gov) Research Data (1973-2009), National Cancer Institute, DCCPS, Surveillance Research Programm, Surveillance Systems Branch, released April 2012, based on the November 2011 submission.
Sen A. Banerjee M.B., Yun L. and Noone A.M. (2010): A Bayesian approach to competing risks analysis with masked cause of death. Statistics in Medicine, 29, 1681-1695