Comparison of competing risks models based on cumulative incidence function in analyzing time to cardiovascular diseases

Minoo Dianatkhah[(1)], **Mehdi Rahgozar**[(2)], **Mohammad Talaei**[(3)], **Masoud Karimloua**[(2)],
Masoumeh Sadeghi[(4)], **Shahram Oveisgharan**[(5)], **Nizal Sarrafzadegan**[(6)]

Original Article

Abstract

BACKGROUND: Competing risks arise when the subject is exposed to more than one cause of failure. Data consists of the time that the subject failed and an indicator of which risk caused the subject to fail.

METHODS: With three approaches consisting of Fine and Gray, binomial, and pseudo-value, all of which are directly based on cumulative incidence function, cardiovascular disease data of the Isfahan Cohort Study were analyzed. Validity of proportionality assumption for these approaches is the basis for selecting appropriate models. Such as for the Fine and Gray model, establishing proportionality assumption is necessary. In the binomial approach, a parametric, non-parametric, or semi-parametric model was offered according to validity of assumption. However, pseudo-value approaches do not need to establish proportionality.

RESULTS: Following fitting the models to data, slight differences in parameters and variances estimates were seen among models. This showed that semi-parametric multiplicative model and the two models based on pseudo-value approach could be used for fitting this kind of data.

CONCLUSION: We would recommend considering the use of competing risk models instead of normal survival methods when subjects are exposed to more than one cause of failure.

Keywords: Competing Risks, Cumulative Incidence Function, Fine and Gray Model, Binomial Approach, Pseudo-value Approach, Cardiovascular Diseases

Introduction

Problems involving competing risks are common in medical researches, where (K > 0) competing causes of failure may occur. Occurrence of any of the risks causes failure or death and precludes the occurrence of other competing risks.[1,2] For such data one observes only the failure time and a cause of failure for each subject in the study. Methods for estimating the probability of failure for events that are subject to competing risks are not new. It is still quite common to see inappropriate methods used to estimate such probabilities for endpoints that suffer from competing risks.[1]

Generally, two types of analysis can be performed when competing risks are present; modeling cause-specific and sub-distribution hazard or cumulative incidence function.[3,4] The Cox regression modeling for each event is an example of the first type. In such a model a subject who has failed in other competing risks is treated as a censored subject. This method is valid if the censoring distributions are independent.[5] Multi-state models that do not require the existence of potential failure times and Aalen additive hazards model are

1- Department of Statistics and Computer Sciences, University of Social Welfare and Rehabilitation Sciences, Tehran AND Isfahan Cardiovascular Research Center, Isfahan Cardiovascular Research Institute, Isfahan University of Medical Sciences, Isfahan, Iran
2- Assistant Professor, Department of Statistics and Computer Sciences, University of Social Welfare and Rehabilitation Sciences, Tehran, Iran
3- Isfahan Cardiovascular Research Center, Isfahan Cardiovascular Research Institute, Isfahan University of Medical Sciences, Isfahan, Iran AND Saw Swee Hock School of Public Health, National University of Singapore, Singapore, Singapore
4- Associate Professor, Cardiac Rehabilitation Research Center, Isfahan Cardiovascular Research Institute, Isfahan University of Medical Sciences, Isfahan, Iran
5- Department of Neurology, Tehran University of Medical Sciences, Tehran, Iran
6- Professor, Isfahan Cardiovascular Research Center, Isfahan Cardiovascular Research Institute, Isfahan University of Medical Sciences, Isfahan, Iran
Correspondence to: Mehdi Rahgozar, Email: m_rahgozar2003@yahoo.com

other examples of the first type of modeling.[6,7] Klein modeled covariate effects using this method.[7] For the second type, we can find the Fine and Gray[8] method, the binomial approach suggested by Scheike and Zhang,[9] and the pseudo-value approach suggested by Klein and Andersen.[10,11] These approaches are introduced in section 3. We fitted these three methods to cardiovascular diseases (CVD) data of the Isfahan Cohort Study (ICS) introduced in section 2.4.[12,13] In section 3 We present the results, and in section 4 findings are discussed in brief.

Materials and Methods

The most common model for competing risks is in terms of potential failure times, where K is competing risks denoted by $D_1,\ldots,D_k$, and for each risk there is a potential failure time of X_i, $i=1,\ldots,K$. One observes $T = min(X_1,\ldots,X_k)$ and a variable $\varepsilon = j$, $j = 1,\ldots,K$,

Where $T = X_j$ defines which of the risks caused the event to occur. Competing risk probabilities can be summarized by cumulative incidence function for the jth competing risk. This function is defined as probability of experiencing risk j prior to time t in the presence of all competing risks. This quantity depends on all the cause-specific hazard rates ($h_j(t) = 1,\ldots,k$), not just the crude hazard rate of cause of interest.[1]

(1)

$$F_j(t) = P[T \leq t, \varepsilon = j\,] = \int_0^t h_j(x) \exp\{-\sum_{i=1}^k \int_0^x h_i\,(u)$$

When there is a covariate, it is common in medical sciences to study the effect on competing risks quantities.[14-18] One solution is a direct regression modeling of cumulative incidence function. Here, we discuss three approaches that focus on this topic.

Fine and Gray Model

The first approach suggested by Fine and Gray[8] is a proportional sub-distribution hazards model with:

(2) $\gamma(t, Z) = \gamma_0(t)\exp[\beta'Z]$

Where γ and γ_0 are hazard and baseline hazard of the sub-distribution, Z and β are vectors of covariates and coefficients, respectively. The partial likelihood is given by:

(3) $L(\beta) = \prod_{i=1}^r \left(\frac{exp(\beta z_i)}{\sum_{j \in R_i} \omega_{ij} exp(\beta z_j)} \right)$

$R_i = \{j; t_j \geq t_i$ or $(t_{j \leq} t_i$ and the subject had competing risk event)$\}$

The risk set R_i is formed of those who did not experience an event by time t and those who experienced a competing risk event by time t. Thus, those who experienced other types of events remain

in the risk set all the time. The weights are defined as:

(4) $\omega_{ij} = \frac{\hat{G}(t_i)}{\hat{G}(min(t_i,t_j))}$

Where $\hat{G}$ is the Kaplan-Meier estimate of survivor function of the censoring distribution.[3] This model is valid if the proportionality assumption is established.

Binomial Approach

The second method is the direct binomial approach suggested by Scheike and Zhang[9] which models cumulative incidence function by a general class of models given by:

(5) $h\{F_1(t,z)\} = g\{\eta(t),\beta,z\}$

Where h and g are the known link and regression functions, respectively, $\eta(t)$ is the unknown regression function and β is the vector of regression parameters. We use the semi-parametric multiplicative model:

(6) $c1n1n\{1-F_1(t;x,z)\} = \eta(t)'x+\beta'z$

Where X is a $(p+1)$ –dimensional ($X = (1,x_1,\ldots,x_p)$), and Z a q-dimensional covariate. These flexible models allow covariate X to have time-varying effects and the covariate Z to have constant effects:

(7) $E\left(\frac{\Delta_i N_i(t)}{G(T_i)}\right) = F_1(t; X_i, Z_i)$

The model suggests testing the hypothesis that a specific covariate x_j has a constant effect over time and define hypothesis H_0: $\eta_j(t) \equiv \eta$. This leads to a very useful goodness-of-fit test for model validation. The test shows exactly where non-proportionality is present. This approach is to start out with a model where all effects initially have parametric or non-parametric effects, and then reduce model complexity by successive testing to find an appropriate semi-parametric model that fits the data. In brief, for this approach, the model is chosen according to proportionality assumption.

Pseudo-value Approach

The third method of direct modeling of the cumulative incidence function is based on a pseudo-value approach.[11] For this model a grid of time points $\tau_1,\ldots,\tau_M$ is selected. At each grid point, the estimated cumulative incidence function is computed based on the complete data set $\hat{F}(\tau_h)$ and the estimated cumulative incidence function based on the sample of size n-1 obtained by deleting the ith observation $\hat{F}^{(i)}(\tau_h)$ then the pseudo-value for the ith subject at time τ_h is defined as:

(8) $\hat{\theta}_{ih} = n\hat{F}(\tau_h) - (n-1)\hat{F}^{(i)}(\tau_h)$, $i = 1, \ldots, n$, $h = 1$, There are the pseudo-values known from jack-knife techniques. $n\hat{F}(t)$ is the number of events of type of interest occurring prior to t, When there

is no censoring. In this case $\hat{\theta}_i = (\hat{\theta}_{ih}, h = 1, ..., M) = (I(T_i \leq \tau_1, \varepsilon_i = 1), ..., I(T_i \leq \tau_M, \varepsilon_i = 1))$ and $\hat{\theta}_i's$ are independent. When we have censoring, because pseudo-values are close to the indicators they are approximately independent. This allows us to make use of results from generalized linear models to model the effects of covariates.

(9) $g(\theta_{ih}) = \alpha_h + \gamma'Z_i = \beta'Z_{ih}$, $i = 1, ..., n$, $h = 1, ..., M$

Where $g(0)$ is a link function. The possible choices could be the logit link $g(x) = \log(x/(1-x))$, or complementary log-log function $g(x) = -\log(-\log(1-x))$ on x. Unlike the Fine and Gray model, this approach does not need to establish proportionality assumption. To select the appropriate link function, one crude way, when the factor is categorical, is to look at plots of differences in transformed estimates of the cumulative incidence functions for each category from the baseline category.

For two categorical factors, the cumulative incidence functions for two groups (ignoring other covariates), is estimated separately. Then, $g(F_{1h}(t)) - g(F_{10}(t))$ is plotted, here $F_{10}(t)$ and $F_{1h}(t)$ are the estimated cumulative incidence function for baseline and other categories, respectively, and $g(0)$ is either the logit or complementary log-log transforms. If the link chosen for the plot is correct, then the curves should approximately be horizontal.

Data

To compare these three approaches, we used the data of the Isfahan Cohort Study. The ICS is a community-based, ongoing longitudinal study on 6504 adults aged 35 and older at baseline, aiming at Iranian cardiovascular disease risk chart. Participants lived in both urban and rural areas of three cities and their associated district villages in central Iran (Isfahan, Arak, Najafabad). Several risk factors for cardiovascular disease, like smoking status, lipids, blood pressure, and anthropometric measurements, were measured at baseline. They were followed for 5 years from January 1997 to September 2001. End of study for each subject was confirmed if one of the cardiovascular disease events (CVD) (non-fatal myocardial infarction, fatal myocardial infarction, non-fatal stroke, fatal stroke, sudden cardiac death, and unstable angina) occurred or the subject experienced unrelated CVD death. Finally, data of 5515 participants who had at least one follow-up time after baseline were included in analysis. There is one competing risk of CVD event (event of interest), and it has occurred when the subject experienced unrelated CVD death.[12-19]

Results

From 5515 (2815 females and 2700 males) cases in ICS data, 5.13% had one of the mentioned CVD and 1.5% experienced unrelated CVD death. The study consisted of patients with non-fatal myocardial infarction (n = 52), fatal myocardial infarction (n = 19), sudden cardiac death (n = 46), non-fatal stroke (n = 40), fatal stroke (n = 14), and unstable angina (n = 112). Moreover, 2133 subjects were 35 to 44 years old, 2449 between 45 to 64, and 933 were 65 and older at baseline.

To fit ICS data with R software, the 3 Fine and Gray, binomial, and pseudo-value competing risks approaches, which are directly based on cumulative incidence function were used.[3,5,20,21] As is common in medical literature, parametric models have been studied first. Table 1 shows the results. The Fine and Gray model has maximum number of significant covariates (8) and the lowest variances. On the contrary, multiplicative models have minimum number of significant covariates (6) and the most variances, and 7 covariates are significant in logit and complementary log-log models. In the Fine and Gray model, except for abdominal obesity (P = 0.76) and high low-density lipoprotein cholesterol (high LDL-C) (P = 0.20), other covariates are significant (P < 0.05). For the multiplicative model, age, abdominal obesity, hypertension, diabetes mellitus, and current smoking status are significant (P < 0.05). In logit and complementary log-log models, age, hypertension, high LDL-C, low high-density lipoprotein cholesterol (low HDL-C), diabetes mellitus, and current smoking status are significant (P < 0.05). Slight differences among the models are seen for parameter estimates. In addition, for the Fine and Gray logit and multiplicative models, we can interpret $\exp(\hat{b})$ as the odds in favor of the categories of a factor relative to the baseline category. Table 2 shows the results of fitting of non-parametric multiplicative model. These models differ from parametric models, because their coefficients have time-varying effects. This table also shows the results of testing goodness-of-fit or constant effect test. Age (65 years and older), abdominal obesity, and diabetes mellitus are significant (P < 0.05). This implies that Fine and Gray, parametric and non-parametric multiplicative models are not appropriate, because the proportionality assumption is violated. Therefore, fitting the semi-parametric model is necessary and allows the covariates with constant and non-constant effects to be presented simultaneously in

the model. We use this model later to predict cumulative incidence function for specific subjects. Table 3 shows semi-parametric model results. For this model, age (65 years and older), abdominal obesity, and diabetes mellitus do not have parameter estimates, because of their non-constant effects in time. Figure 1 shows goodness-of-fit plot for hypertension with two logit and complementary log-log transforms. The two plots are approximately horizontal; meaning that both are suitable. Because of differences in variance estimation between these two models, the complementary log-log model is preferred.

Table 1. Results of fitting parametric models on Isfahan Cohort Study (ICS) data

Covariate			Fine and Gary model	Logit model	Complementary log-log model on 1-F_1(t)	Multiplicative model
Sex[**]		B	0.482	0.320	0.265	0.186
		SE (b)	0.146	0.191	0.180	0.221
		P	(0.001)[*]	(0.094)	(0.142)	(0.400)
Age[***]	45-64	B	0.828	0.790	0.770	1.190
		SE (b)	0.188	0.252	0.246	0.276
		P	(< 0.001)[*]	(0.002)[*]	(0.002)[*]	(< 0.001)[*]
	≥ 65	B	1.475	1.438	1.372	1.900
		SE (b)	0.198	0.259	0.251	0.278
		P	(< 0.001)[*]	(< 0.001)[*]	(< 0.001)[*]	(< 0.001)[*]
Abdominal obesity		B	-0.04	-0.165	-0.168	-0.460
		SE (b)	0.151	0.200	0.188	0.244
		P	(0.760)	(0.409)	(0.372)	(0.050)[*]
Hypertension		B	0.980	1.154	1.099	1.190
		SE (b)	0.129	0.158	0.150	0.202
		P	(< 0.001)[*]	(< 0.001)[*]	(< 0.001)[*]	(< 0.001)[*]
High LDL-C		B	0.455	0.412	0.381	0.194
		SE (b)	0.124	0.163	0.154	0.200
		P	(< 0.001)[*]	(0.012)[*]	(0.013)[*]	(0.313)
Low HDL-C		B	0.162	0.376	0.353	0.336
		SE (b)	0.153	0.168	0.157	0.210
		P	(0.200)	(0.025)[*]	(0.024)[*]	(0.109)
Diabetes mellitus		B	0.592	0.600	0.513	0.733
		SE (b)	0.153	0.191	0.177	0.225
		P	(< 0.001)[*]	(0.002)[*]	(0.004)[*]	(0.001)[*]
Hypertriglyceridemia		B	0.340	0.253	0.233	0.119
		SE (b)	0.137	0.177	0.167	0.224
		P	(0.013)[*]	(0.153)	(0.163)	(0.597)
Smoking		B	0.391	0.585	0.533	0.607
		SE (b)	0.153	0.198	0.184	0.233
		P	(0.010)[*]	(0.003)[*]	(0.003)[*]	(0.009)[*]

[*] Significant at α = 0.05 level; [**] Females are reference group; [***] Age between 35 and 44 are reference group
SE: Standard error; LDL-C: Low-density lipoprotein cholesterol; HDL-C: High-density lipoprotein cholesterol

Table 2. P-values for non-parametric model on Isfahan Cohort Study (ICS) data

Covariate		Multiplicative Model	
		H_0: η(t)=0	H_0: Constant effect
Sex		0.358	0.264
Age	45-64	< 0.001[*]	0.280
	> = 65	< 0.001[*]	< 0.001[*]
Abdominal obesity		0.002[*]	0.016[*]
Hypertension		< 0.001[*]	0.096
High LDL-C		0.170	0.508
Low HDL-C		0.118	0.490
Diabetes mellitus		< 0.001[*]	0.024[*]
Hypertriglyceridemia		0.240	0.578
Smoking		0.012[*]	0.084

[*] Significant at α = 0.05 level
LDL-C: Low-density lipoprotein cholesterol; HDL-C: High-density lipoprotein cholesterol

**Table 3.** Results of fitting semi-parametric model on Isfahan Cohort Study (ICS) data

Covariate		Multiplicative Model		
		b	SE (b)	P
Sex		0.142	0.225	0.527
Age	45-64	1.090	0.225	< 0.001[*]
	≥ 65	-	-	< 0.001[*]
Abdominal obesity		-	-	< 0.001[*]
hypertension		1.190	0.201	< 0.001[*]
High LDL-C		0.213	0.202	0.292
Low HDL-C		0.375	0.225	0.081
Diabetes mellitus		-	-	< 0.001[*]
Hypertriglyceridemia		0.110	0.236	0.640
Smoking		0.635	0.234	0.006[*]

[*] Significant at α = 0.05 level;
SE: Standard error; LDL-C: Low-density lipoprotein cholesterol; HDL-C: High-density lipoprotein cholesterol
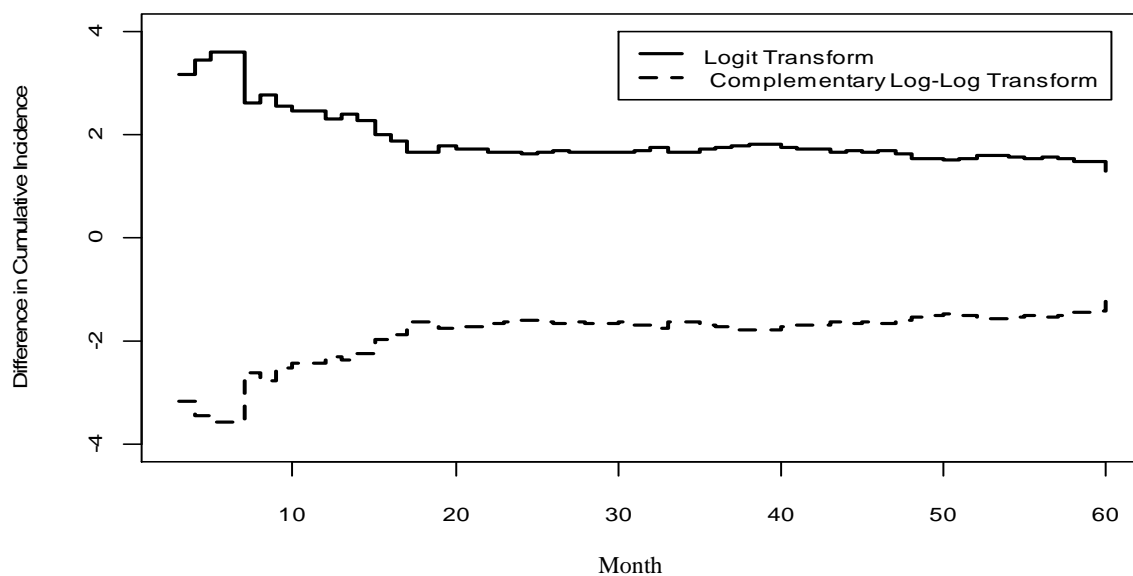
Figure 1. Difference in cumulative incidence function for logit and complementary log-log transform in hypertension
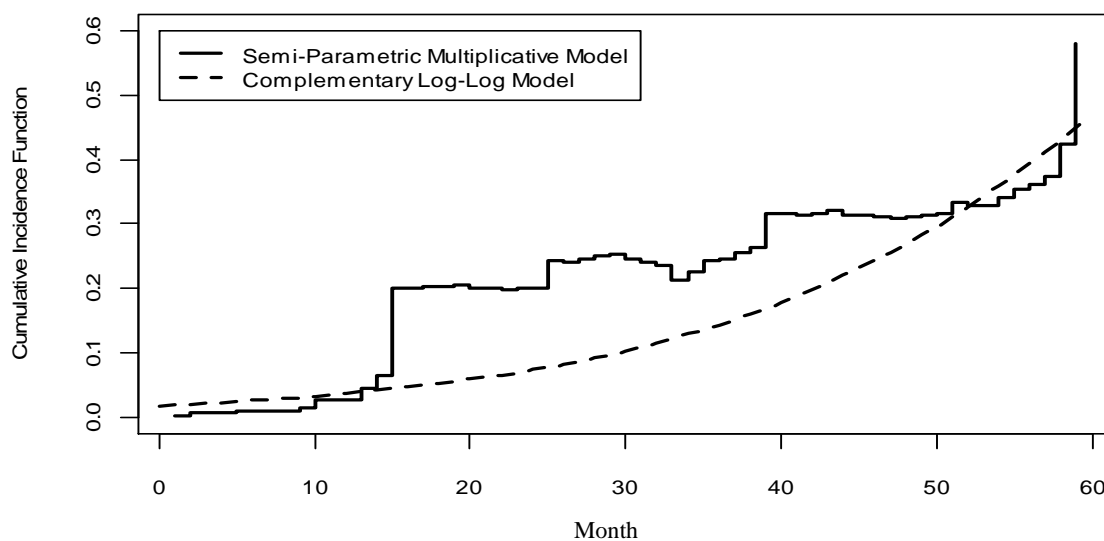

Figure 2. Predictions for cardiovascular diseases (CVD) cumulative incidence function for Isfahan Cohort Study (ICS) data using semi-parametric multiplicative and complementary log-log model

Sometimes it is important to get an idea of the cumulative incidence probability for specific patients. Therefore, computing the predicted cumulative incidence function for a given set value of covariates is very popular.[22,23] For example, suppose that physicians want to know the value of cumulative incidence function for male patients older than 65 with abdominal obesity, hypertension, high LDL-C, low HDL-C, diabetes mellitus, hypertriglyceridemia, and smoking. Figure 2 shows the predicted cumulative incidence function during 60 months for two appropriate complementary log-log and semi-parametric multiplicative models. The predicted values for the first model are less than the second model for about 35 months (between the 15th-58th months).

Discussion

Data from studies with competing risks outcomes present challenges to the data analyst. Some articles analyze such data with normal survival models. A criticism that can be leveled at these models is the assumption that upon removal of one cause of failure, the risk of failure from remaining causes is unchanged. In human studies this assumption is rarely true.[3-5] Here we have used three approaches (Fine and Gray, binomial, and pseudo-value approaches) which are based directly on the cumulative incidence function and their validity depends on proportionality assumption. This collection of models gives a rich variety, from which a user can choose an appropriate model for analyzing the data.

We saw that the Fine and Gray, parametric multiplicative model was not able to describe the cumulative incidence function for ICS data. This model's lacking flexibility was found using the goodness-of-fit approach. This showed that its non-proportionality can primarily be attributed to the effect of covariates. A similar conclusion was reached for the non-parametric multiplicative model. The semi-parametric multiplicative model could be a good choice for this data. With the pseudo-value approaches, two link functions were used in GLM model (logit or complementary log-log function). Unlike the Fine and Gray and multiplicative models, this is more flexible so that we do not need to assume proportionality. Goodness-of-fit plots showed that both link functions are suitable for hypertension groups, but they were different in variance estimation. Moreover, it seems the complementary log-log function is more appropriate. Predictions plot for

ICS data using semi-parametric multiplicative and complementary log-log models were quite similar during 5 years, but slight differences in parameters regression were found between the two models.

Conclusion

Inappropriate statistical methods are not rare in binomial literature.[5] The competing risk problem is a critical issue in survival analysis. We would recommend considering competing risk models instead of simply using normal survival methods when subjects are exposed to more than one cause of failure. In future studies like ICS, using competing risks models is suggested, because a large number of unrelated CVD deaths will occur during years of follow-up and the use of normal survival functions can lead to incorrect or at least imprecise estimates. As we described, the two appropriate semi-parametric multiplicative and complementary log-log models are proposed for fitting of such data.

Acknowledgements

Conflict of Interests

Authors have no conflict of interests.

References

1. Gooley TA, Leisenring W, Crowley J, Storer BE. Estimation of failure probabilities in the presence of competing risks: new representations of old estimators. Stat Med 1999; 18(6): 695-706.
2. Klein JP. Competing risks. Wiley Interdisciplinary Reviews: Computational Statistics 2010; 2(3): 333-9.
3. Pintilie M. Competing Risks: A Practical Perspective. New Jersey, NJ: John Wiley & Sons; 2006.
4. Pintilie M. Analysing and interpreting competing risk data. Stat Med 2007; 26(6): 1360-7.
5. Putter H, Fiocco M, Geskus RB. Tutorial in biostatistics: competing risks and multi-state models. Stat Med 2007; 26(11): 2389-430.
6. Andersen PK, Abildstrom SZ, Rosthoj S. Competing risks as a multi-state model. Stat Methods Med Res 2002; 11(2): 203-15.
7. Klein JP. Modelling competing risks in cancer studies. Stat Med 2006; 25(6): 1015-34.
8. Fine JP, Gray RJ. A Proportional Hazards Model for the Subdistribution of a Competing Risk.

Journal of the American Statistical Association 1999; 94(446): 496-509.

9. Scheike TH, Zhang MJ. Flexible competing risks regression modeling and goodness-of-fit. Lifetime Data Anal 2008; 14(4): 464-83.

10. Andersen PK, Klein JP, Rosthøj S. Generalised Linear Models for Correlated Pseudo-Observations, with Applications to Multi-State Models. Biometrika 2003; 90(1): 15-27.

11. Klein JP, Andersen PK. Regression modeling of competing risks data based on pseudovalues of the cumulative incidence function. Biometrics 2005; 61(1): 223-9.

12. Sarrafzadegan N, Talaei M, Sadeghi M, Kelishadi R, Oveisgharan S, Mohammadifard N, et al. The Isfahan cohort study: rationale, methods and main findings. J Hum Hypertens 2011; 25(9): 545-53.

13. Talaei M, Sadeghi M, Marshall T, Thomas GN, Kabiri P, Hoseini S, et al. Impact of metabolic syndrome on ischemic heart disease - a prospective cohort study in an Iranian adult population: Isfahan Cohort Study. Nutr Metab Cardiovasc Dis 2012; 22(5): 434-41.

14. Cox DR, Oakes D. Analysis of survival data. London, UK: Chapman & Hall; 1984.

15. Crowder MJ. Classical competing risks. New York, NY: Taylor & Francis; 2001.

16. David HA, Moeschberger ML. The Theory of Competing Risks. London, UK: Griffin Publishing Group; 1978.

17. Prentice RL, Kalbfleisch JD, Peterson AV, Flournoy N, Farewell VT, Breslow NE. The analysis of failure times in the presence of competing risks. Biometrics 1978; 34(4): 541-54.

18. Klein JP, Moeschberger ML. Survival Analysis: Techniques for Censored and Truncated Data. New York, NY: Springer; 2003.

19. Sarraf-Zadegan N, Sadri G, Malek AH, Baghaei M, Mohammadi FN, Shahrokhi S, et al. Isfahan Healthy Heart Programme: a comprehensive integrated community-based programme for cardiovascular disease prevention and control. Design, methods and initial experience. Acta Cardiol 2003; 58(4): 309-20.

20. Crawley MJ. The R Book. New Jersey, NY: John Wiley & Sons; 2007.

21. Klein JP, Gerster M, Andersen PK, Tarima S, Perme MP. SAS and R functions to compute pseudo-values for censored data regression. Comput Methods Programs Biomed 2008; 89(3): 289-300.

22. Hyun S, Sun Y, Sundaram R. Assessing cumulative incidence functions under the semiparametric additive risk model. Stat Med 2009; 28(22): 2748-68.

23. Zhang MJ, Zhang X, Scheike TH. Modeling cumulative incidence function for competing risks data. Expert Rev Clin Pharmacol 2008; 1(3): 391-400.