Journal of Statistical Research *ISSN 0256 - 422 X*
2016, Vol. 48-50, No. 2, pp. 71-83

BOUNDS OF THE SURVIVAL FUNCTIONS WITH INFORMATIVE DROP-OUTS USING FGM COPULAS

MD MONIR HOSSAIN
Cincinnati Children's Hospital Medical Center, Cincinnati, OH 45229, USA
Email: md.hossain@cchmc.org

TOSHIYA SHUN SATO
Kyoto University School of Public Health, Kyoto 606-8501, Japan
Email: shun@pbh.med.kyoto-u.ac.jp

SUMMARY

The underpinning assumption of independence of failure time and drop-out time is not well supported in many clinical or epidemiological studies. As a consequence, the marginal survival functions are not identifiable. In such a situation, many authors have proposed bounds for the survival functions to check the sensitivity of the estimates to the independence assumption. In this paper, we propose an alternative methodology by adopting an underlying selection process to account for this dependency. We use the Farlie-Gumbel-Morgenstern (FGM) bivariate family for the joint distribution of failure time and the selection variable. Subsequently we derive the conditional distribution of the survival time, given that it is observed, and show how, given the association parameter, the survival functions can be estimated. We compare the proposed estimates with the Copula-Graphic estimator for a real home haemodialysis data and simulated datasets for various proportions of drop-outs.

Keywords and phrases: Informative drop-out, Normal selection process, survival function, Farlie-Gumbel-Morgenstern bivariate distribution, sensitivity analysis

1 Introduction

It is common practice to estimate the marginal survival function to summarize the failure time data, e.g., the Kaplan-Meier (K-M) or the product-limit estimate in a nonparametric case or the Weibull estimate in a parametric case. In either case of parametric or nonparametric approach, the validity of the estimates depends on the appropriateness of the independence assumption of the failure event (i.e. the event of interest) and the events causing the censored times.

In many clinical studies, the events causing the censored times can be grouped into two categories: (a) censored due to random causes (e.g. end of the study period or moving to a new location) and (b) censored due to non-random causes (e.g. too ill to adhere the follow-up schedule, changed to a different therapy or no need to come to a clinic because of the complete recovery from the disease

under study). It is the second group where the assumption of independence deserves special scrutiny, because the event of interest is related to the cause of censoring. For example, if the study objective is to demonstrate whether the experimental treatment is more effective in prolonging the survival time than the existing treatment, the second group of censoring contains information that is related to the study objective whereas the first group can be regarded as independent. In this paper, we name the first group (a) as "random censoring" and the second group (b) as "informative drop-out", defined as the patient who withdraws from the study before the end-of-study for any non-random cause.

The extent of dependency between the failure time and the drop-out time varies according to the cause of drop-out. In one extreme, if the cause of drop-out is *too ill to adhere to the follow-up schedule* then it is very likely that the dropped-out patient failed immediately after the drop-out time. The observed drop-out time can be considered as an approximate failure time. In another extreme, if the drop-out cause is *complete recovery from the disease under study* then it is very likely that the dropped-out patient would survive longer (at least to the end-of-study time) and the patient can be considered as censored at the end-of-study time. Besides these two extremes, other causes of informative drop-out could be *change to another treatment due to lack of effectiveness of the experimental treatment or due to adverse treatment effects*.

From the observed data, it is impossible to get an estimate for the association parameter because we observe either the failure time or the drop-out time for each patient, but not both. To circumvent this problem, a well-known approach is to conduct a sensitivity analysis, which proposes a set of alternative survival functions for a set of plausible values of the association parameter. The upper and lower bounds can be derived for the limits of the association parameter. Fisher and Kanarek (1974) are the first authors who proposed a sensitivity analysis of the assumption of independence for various values of scale parameter. They distinguished the two types of censoring: end-of-study censoring and lost-to-follow-up censoring. Their model is based on the assumption that the remaining lifetime for lost-to-follow-up cases are shortened or expanded by a fixed quantity associated with the scale parameter.

To date, a number of authors have proposed alternative methodologies to estimate the bounds for the survival function (Peterson, 1976; Williams and Lagakos, 1977; Lagakos and Williams, 1978; Slud and Rubinstein, 1983; Zheng and Klein, 1995). Robins and Finkelstein (2000) also proposed a correction to the survival function for noncompliance and dependent censoring. In Bayesian framework, Kaciroti et al. (2012) developed a product limit method where estimates of the cumulative incidence curves were expanded to include informative censoring by pattern mixture models. The sensitivity analysis was conducted on the prior distribution for the selectivity parameter specified in a probabilistic range. Shardell et al. (2007) extended the standard approach of estimating survival functions for the informatively interval-censored data in a Bayesian framework. The estimates were produced by mixing over a distribution of assumed censoring mechanisms. Yuan et al. (2012) proposed a model-based method to estimate progression-free survival time when the progression status of disease for some patients is unknown. The method specified a joint distribution for the marginal distributions of the time to progression and the time from progression to death by using a Clayton copula. In relevant area of conducting sensitivity analysis for assessing treatment effect

in a follow-up study when patients initiated to a nonrandomized therapy or dropped out, Rotnitzky et al. (2001) proposed semiparametric cause specific selection models where the magnitude of selection bias varies with the cause over a plausible range of selection parameter. Todem et al. (2010) used parametric and semiparametric models with shared random effects to explain the dependence between the measurement and the drop-out processes. A part of the shared random effects was considered for the sensitivity analysis.

The main objective of this paper is to propose a new methodology to calculate the bounds for the survival function when the drop-outs are informative. We compare these bounds with the copula-graphic estimator proposed by Zheng and Klein (1995). To explain the drop-out mechanism, we assume a hidden selection process (Copas and Li, 1997) to explain the non-randomness of the drop-out data. To specify the joint distribution for observed failure time and the underlying selection process, we adopt a bivariate family of distribution known as Farlie-Gumbel-Morgenstern (FGM) copulas (Morgenstern, 1956; Schucany et al., 1978). Then, we derive the conditional distribution of the survival probability given that it is observed.

In the next section, we explain the basic ideas behind our formulation of sensitivity analysis for informative drop-out. In Section 2.1, we propose a modified survival function for sensitivity analysis based on the underlying selection process. A brief description of Copula-Graphic estimates proposed by Zheng and Klein (1995) is given in Section 3. In Section 4, we illustrate the proposed method with a real dataset and compare with the Copula-Graphic estimates. In Section 5, with the simulated data we compare our estimates with the Copula-Graphic estimates for various drop-out proportions. The final section includes all the concluding remarks.

2 Methodology

Let $F, C,$ and D be the random variables indicating the failure time, random censoring time and informative drop-out time, respectively, for N subjects. In practice, we observe only $T = \min(F, C, D)$ and a vector of covariates X for each subject. We introduce the notation, $T^\star = \min(F, C^\star)$, where $C^\star$ combines C and D with the assumption that all the censored events are independent of the cause of event of interest. The informative drop-out time, D, can be considered as a potential drop-out time which is assigned to each individual after his/her onset time and it is observed only when $D_i \leq T_i^\star$, for $i = 1, 2, \ldots, N$.

To explain the drop-out mechanism, we use the concept of unobserved (or, hidden) selection variables. Let Z be a hidden selection variable with dimension N defined as $Z_i = D_i - T_i^\star$ for $i = 1, 2, \ldots, N$. We assume that a patient is observed to drop-out with drop-out time D_i whenever $Z_i \leq 0$ and to experience the failure time or random censoring time at $T_i^\star$ whenever $Z_i > 0$. Thus, it can be assumed that the main model for observed time, T_i, is supplemented by a selection variable, $Z_i > 0$, and whenever the drop-outs are informative, the model of interest is the conditional model of T_i given $Z_i > 0$. In other words, the model of interest is the model of failure time conditioning on the dropped-out subjects did not drop-out, i.e., the model we observe.

For the model for log-transformed failure time, $Y\big(= \log(T)\big)$, we consider an accelerated failure-time model and a linear representation with the covariates $X = (X_1, X_2, \ldots, X_p)^T$. Thus,

the regression model for Y yields

$$y = \beta^T x + \sigma\varepsilon,$$

where $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)^T$ and σ is a scale parameter, also known as an acceleration parameter. The probability distribution for the residual, $\varepsilon \in (-\infty, \infty)$, will follow either a standard extreme value distribution with mean $-E_g$ (Euler gamma ≈ 0.5772) and variance π^2/σ or a standard normal distribution. It is well known that standard extreme value distribution for the residual yields an exponential distribution. (when $\sigma = 1$) or a Weibull distribution (when $\sigma \neq 1$) and standard normal distribution yields a log-normal distribution for the failure time, T.

Let the selection equation with the same set of covariates $X = (X_1, X_2, \ldots, X_p)^T$ can be assumed as

$$z = \gamma^T x + \xi,$$

where the residual of selection equation, ξ has standard normal density.

In order to calculate the conditional density $f(y \,|\, x, z > 0)$, now the problem reduces to define a joint distribution for the given marginals. We adopt a joint distribution of ε and ξ proposed by Morgenstern (1956), which is given by

$$F(\varepsilon, \xi) = F_M(\varepsilon) F_S(\xi) \big[1 + \theta\big(1 - F_M(\varepsilon)\big)\big(1 - F_S(\xi)\big)\big],$$

where $F_M(\varepsilon)$ and $F_S(\xi)$ are the respective distribution functions for ε and ξ. The above joint distribution is also known as Farlie-Gumbel-Morgenstern (FGM) copulas.The subscripts S and M stand for the selection process and the measurement process, respectively. The covariance parameter $\theta \in [-1, 1]$ is proportional to the product-moment correlation, $\rho \in [-1/3, 1/3]$, between the two residuals, i.e. $\rho = k\theta$. After applying a method proposed by Schucany et al. (1978) for the known marginals, the value of k is determined to be equal to 0.3049. This smaller value of k indicates that the limit of the product-moment correlation, ρ , for our choice of marginals, is even tighter than $(-1/3 \leq \rho \leq 1/3)$. Hereafter, we use covariance parameter θ instead of correlation coefficient ρ as a sensitivity parameter. Because of the notion of hidden selection process, θ is also known as a selectivity parameter.

The tighter bound of correlation coefficient for FGM copulas has restricted the wider application of this joint distribution. But for our purpose, this restriction comes as an advantage, since the main objective of this paper is to show the sensitivity of the survival function for a small violation of the independence assumption. The other reason for choosing this particular form of joint distribution is because the conditional survival function can be expressed in a simple form, as we will see in the next section.

After some simple algebraic operation, the conditional distribution function of Y is given by

$$F_M(y \,|\, x, z > 0) = F_M(y \,|\, x) \big[1 - \theta\big(1 - \Phi(\boldsymbol{\gamma}^T x)\big)\big(1 - F_M(y \,|\, x)\big)\big],$$

where Φ indicates standard normal distribution. In Appendix A, we provide the log-likelihood function and illustrate how the MLE of $\hat{\beta}(\theta)$ and $\hat{\sigma}(\theta)$ can be obtained.

Interpretation of θ. The covariance parameter, θ, is indexing the extent of association. A useful interpretation of θ can be given in parallel to Copas and Li (1997) and Rosenbaum (1987, 1988). When $\theta > 0$, it implies that the subjects with $Y_{z>0}$ (observed log-failure time) would survive longer and the subjects with smaller survival times tend to drop out, i.e., the unobserved failure times for the informative drop-out subjects are likely to be smaller and only the subjects with longer survival times are observed. For $\theta < 0$, the implication is reversed, i.e., the unobserved failure times for the informative drop-out subjects are likely to be longer and only the subjects with shorter survival times are observed. Clearly, for $\theta = 0$, the drop-out is noninformative and the usual analysis, based on the assumption that any type of censoring events are independent of failure events, is acceptable.

2.1 Survival function

After changing some notation for a better presentation, the conditional distribution function of Y can be written in the form of conditional survival distribution

$$S(y \,|\, x, \theta) = S(y \,|\, x, \theta = 0)\big[1 - \theta\big(1 - \Phi(\gamma^T x)\big)\big(S(y \,|\, x, \theta = 0) - 1\big)\big], \tag{2.1}$$

where $S(y \,|\, x, \theta = 0)$ is the usual survival function based on the assumption that the failure event and any type of censoring events are independent. It is possible to prove that for any fixed value of θ, the conditional survival distribution $S(+\infty \,|\, x, \theta) = 0$ and $S(-\infty \,|\, x, \theta) = 1$, since $S(+\infty|\, x, \theta = 0) = 0$ and $S(-\infty|\, x, \theta = 0) = 1$. This implies that the proposed conditional survival function has the same limit as the usual survival function has. For a simple representation for comparison of the survival functions, we ignore all the explanatory variables in equation (2.1). With the replacement of X by a vector of 1's in equation (2.1), the function $\big(1-\Phi(\gamma^T x)\big)$ represents $f_s(z \le 0)$, a distribution for drop-out. Thus at time t, $f_s(z_t \le 0)$ can be replaced by an empirical estimate $n(t)/N$, where $n(t)$ is the total number of dropped out in the interval $(0, t]$. The estimate $n(t)/N$ is the probability of drop-out at time t, thus $n(t)/N$ increases for increasing number of drop-out with time t.

After all these replacements, the conditional survival function in (2.1) reduces to

$$\hat{S}(y \mid \theta) = \hat{S}\big(y\big)\Big[1 - \frac{\theta\, n(t)}{N}\big(\hat{S}(y) - 1\big)\Big], \tag{2.2}$$

where the sensitivity parameter $\theta \in [-1, 1]$ is a function of ρ, $\hat{S}(y)$ is an estimate of the usual survival function based on the assumption that the failure time and any type of censoring time are independent. Thus the modified survival function, $S(y \mid \rho)$ tends to be sensitive to the value of θ as the proportion of drop-out increases.

In the above expression, the survival time, Y, is in logarithmic scale. Changing it to original scale, T, does not change the expression (2.2). For example, suppose that the survival time T follows Weibull distribution, then the survival function for Y given ρ is expressed as

$$\hat{S}(y \mid \theta) = \exp\Big(-\exp\big((y - \hat{\beta})/\hat{\sigma}\big)\Big)\Big[1 - \frac{\theta\, n(t)}{N}\Big\{\exp\Big(-\exp\big((y - \hat{\beta})/\hat{\sigma}\big)\Big) - 1\Big\}\Big].$$

On the original scale, the survival function for T given ρ is given by

$$\hat{S}(t \mid \theta) = \exp\big(-t^{1/\hat{\sigma}}\exp(-\hat{\beta}/\hat{\sigma})\big)\Big[1 - \frac{\theta\, n(t)}{N}\Big\{\exp\Big(-t^{1/\hat{\sigma}}\exp(-\hat{\beta}/\hat{\sigma})\Big) - 1\Big\}\Big].$$

The survival function in (2.2) can also be written in the form of Kaplan-Meier or product limit estimate as

$$\hat{S}(t \mid \theta) = \prod_{t_i<t} \big(1 - (-d_i/n_i)\big)\Big[1 - \frac{\theta\, n(t)}{N}\Big\{ \prod_{t_i<t} \big(1 - (-d_i n_i)\big) - 1\Big\}\Big], \tag{2.3}$$

where d_i is the number of deaths at time t_i and n_i is the number of subjects at risk prior to time t_i. Note that, the expressions (2.2) and (2.3) reduce to the common survival function (when failure time and censoring time are independent) for $\theta = 0$.

Expected Lifetime. The mean log survival time, μ_p, when the failure time follows a Weibull distribution, is given by

$$\mu_\theta = E\big(Y \mid \theta\big) = \beta + \sigma\big[- E_g + \theta\,(n/N)\log(2)\big], \tag{2.4}$$

where β and σ are the mean and scale parameter of log failure time, respectively, and n is the total number of drop-outs. Thus, for any value of θ, the average survival time is approximately increased (decreased) by $2^{\sigma\theta(n/N)}$. In other words, the average survival time for a specific θ is $2^{\sigma\theta(n/N)}$ times longer (shorter) than that of under the independence assumption.

Confidence Interval. Two-sided approximate $100(1-\alpha)\%$ confidence limits for $S(y \mid \rho)$ are given by

$$S_L\big(y \mid \theta\big) = S_L(y)\Big[1 - \frac{\theta\, n(t)}{N}\big(S_L(y) - 1\big)\Big]$$
$$S_U\big(y \mid \theta\big) = S_U(y)\Big[1 - \frac{\theta\, n(t)}{N}\big(S_U(y) - 1\big)\Big].$$

Suffixes L and U are for lower and upper bounds, respectively. Approximate limits are obtained by following Nelson (1982) for large sample as:

$$S_L(y) = \exp\big[- e^{\varepsilon_L}\big] \text{ and } S_U(y) = \exp\big[- e^{\varepsilon_U}\big].$$

For 95% confidence interval

$$\varepsilon_L = \big(\frac{y - \hat{\beta}}{\hat{\sigma}}\big) + 1.96\text{se}\big(\frac{y - \hat{\beta}}{\hat{\sigma}}\big) \text{ and } \varepsilon_U = \big(\frac{y - \hat{\beta}}{\hat{\sigma}}\big) - 1.96\text{se}\big(\frac{y - \hat{\beta}}{\hat{\sigma}}\big),$$

where

$$\text{se}\Big(\frac{y - \hat{\beta}}{\hat{\sigma}}\Big) = \sqrt{\frac{1}{\hat{\sigma}^2}\Big[\text{var}(\hat{\beta}) + \big(\frac{y - \hat{\beta}}{\hat{\sigma}}\big)^2 \text{ var}(\hat{\sigma}) + 2\Big(\frac{y - \hat{\beta}}{\hat{\sigma}}\Big) \text{ cov}(\hat{\beta}, \hat{\sigma})\Big]}.$$

3 Copula-Graphic Estimator

Zheng and Klein (1995) proposed a copula-graphic estimator for the marginal survival function when failure and censoring times are dependent. In their work, censoring means both random censoring and informative drop-out. Their estimates are based on assumed copula function, but did

not propose any closed form expression for the survival estimates. Rivest and Wells (2001) gave a closed form expression for the copula-graphic estimator when the assumed copula is Archimedean. The estimate is given as

$$\hat{S}(t) = \omega^{-1}\Big[- \sum_{t_i<t,\delta_i=1} \big\{\omega\big(\hat{\pi}(t_i)\big) - \omega(\hat{\pi}(t_i) - 1/N)\big\}\Big],$$

where $\hat{\pi}(t) = \sum I(t_i \geq t)/N$ and $\omega(\cdot)$ is a decreasing convex function defined on $(0, 1]$ satisfying $\omega(1) = 0$ and δ is a censoring indicator defined as $\delta_i = 1$ if $T_i = F_i$ and 0 otherwise.

When comparing the proposed survival function in (2.2) (or, in (2.3)) with its closest contender copula-graphic estimator, it is important to note that the FGM copulas are not Archimedean, but it is a first-order approximation of Frank copulas which are Archimedean. For this reason, we consider the copula-graphic estimator for Frank copulas in the comparison. The Frank copulas has the form

$$C(u, v) = -\frac{1}{\theta} \ln\Big(1 + \frac{(e^{-\theta u} - 1)(e^{-\theta v} - 1)}{(e^{-\theta} - 1)}\Big),\ \theta \in (-\infty, \infty)\backslash\{0\}.$$

For which the ω-function (Rivest and Wells, 2001) is given as

$$\omega(t) = -\ln\Big(\frac{e^{-\theta t} - 1}{e^{-\theta} - 1}\Big),\ \theta \in (-\infty, \infty)\backslash\{0\}.$$

4 Example: Home Dialysis Data

Lynn et al. (2002) conducted a retrospective cohort study of 168 home dialysis patients to investigate whether blood pressure (BP) was an independent risk factor for survival. All patients were from a regional dialysis unit, beginning dialysis treatment within a certain time period. The treatments were home haemodialysis (HD) and continuous ambulatory peritoneal dialysis (CAPD). The primary outcome was death from any cause. Out of 168 patients, 71 died before the end of the study and 97 were censored. Censoring occurred due to kidney transplantation, change of treatment modality (HD to CAPD, or CAPD to HD), return of renal function, transfer to another centre, and still being on their initial dialysis modality at the end of the study. In this paper, we consider the HD treated group because of moderate drop-out size (12.1%) and longer survival compared to the other dialysis modality treated patients (Delano, 1996) to show how the survival estimates vary for informative drop-outs.

After careful scrutiny on all the censoring events, it is reasonable to suspect that some of the censoring events were related to the effects of dialysis. For example, "return of renal function" or "changed treatment modality" may indicate, respectively, that the patient had recovered from the treatment or the patient had problem with the ongoing treatment. The other censoring events, "transplantation", "transfer to another centre" and "still being on their initial dialysis modality at the end of the study" can be considered as unrelated to any treatment effects. "Transplantation" occurred based on waiting time and availability of kidneys conditional to tissue matching. "Still on their initial dialysis modality" indicates end-of-study censoring and "transfer to another centre" had no relation to treatment. Based on these assumptions, we classify all censoring events into two

categories: random censoring and informative drop-out. Table 1 summarizes the data by treatment group and censoring type.

Table 1: Summary of home dialysis data by treatment group and censoring type

		Random censoring			Informative drop-out		
Treatment group	Dead	A	B	C	D	E	Total
HD	41	57	7	4	2	13	124
CAPD	30	7	2	2	0	3	44

A (transplanted patients), B(still on their initial dialysis)

C (transferred to another center), D (renal function returned)

E (changed treatment modality)

For home HD treated group, the Weibull survival curves for $\theta \in \{1.0, 0.0, -1.0\}$ from top to bottom are given in Figure 1. For the HD treated group, $\hat{\beta}_{\theta=0} = 7.8986$ and $\hat{\sigma}_{\theta=0} = 0.7042$. This yield an average survival time of $\hat{\mu}_{\theta=0} = 1793.88$ days based on independent assumption. When $\theta = 0.5(-0.5)$ the average survival time is $1.03(0.97)$ times longer than that obtained based on the independence assumption. Similarly, for $\theta = 1.0(-1.0)$, it is $1.06(0.94)$ times longer, i.e. approximately 6% increase (decrease). For other values of θ, we can also measure the relative change in the average survival time.

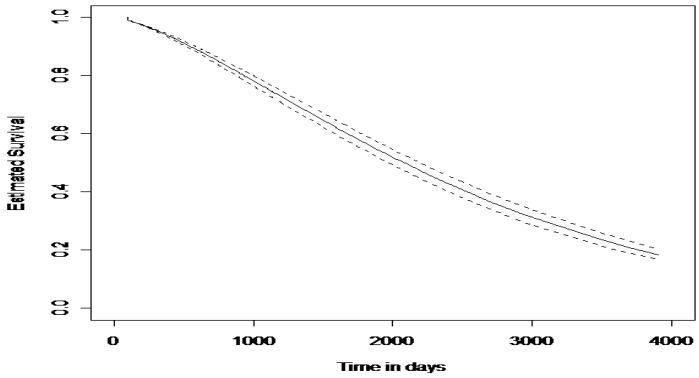

Figure 1: The Weibull sensitivity analysis for HD treated group. The middle solid curve is for $\theta = 0$, the upper and lower dashed curves are for $\theta = 1$ and $\theta = -1$, respectively.

In Figure 2, the survival estimates of Kaplan-Meier (K-M) for $\theta \in \{1.0, 0.0, -1.0\}$ are compared with the Copula-Graphic estimates. The middle dotted line is K-M estimates based on independent

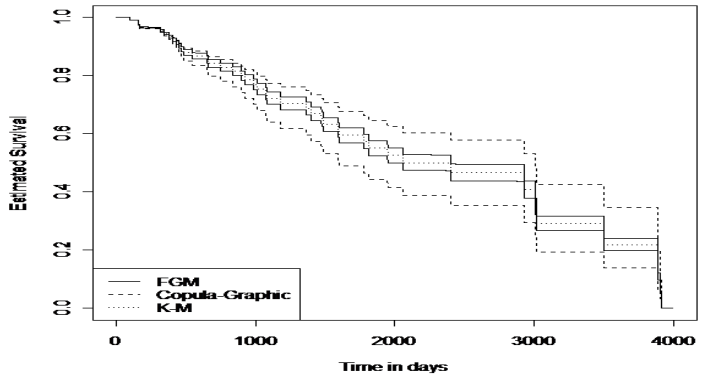

Figure 2: Survival estimates for home dialysis data. FGM estimates (solid lines) for $\theta = 1$ (top) and $\theta = -1$ (bottom), Copula-Graphic estimates for Frank copula (dashed lines) for $\theta = 2.1$ (top) and $\theta = -2.1$ (bottom), and K-M estimates (dotted line).

assumption. The two solid lines are the K-M estimates for $\theta \in \{1.0, -1.0\}$ (from top to bottom) calculated by equation (2.3). The two dashed lines are the Copula-Graphic estimates based on Frank copulas. It clearly shows that the proposed survival estimates based on FGM copulas give much narrower bounds than the Copula-Graphic estimates.

5 Simulation and Results

We conduct a simulation experiment to check the relative bias of our estimates in comparison to Copula-Graphic estimates. The relative performance has been checked for varying rates of informative drop-outs. The comparisons are made for 30%, 15% and 5% drop-out. For all the drop-out scenarios, we have assumed a constant rate of random censoring, 20%.

We generate 1000 data sets of 100 samples each from a gamma frailty copula (Oakes, 1989) with exponential marginal distributions. The value of covariance parameter for gamma frailty copula is assigned to be 1.5 for which Kendall's $\tau = 0.2$ (Zheng and Klein, 1995). The particular reason for this choice is the covariance parameter of FGM copulas $\theta \in [-1, 1]$ the Kendall's $\tau \in [-2/9, 2/9]$. The values for the parameters of the exponential marginal distributions are chosen in a way to attain 30%, 15%, or 5% drop-out. The censoring times are generated from an exponential distribution where the parameter is defined to ensure 20% censoring. The steps involved in generating the data sets are given in Appendix B.

The relative bias was defined as the bias of the estimates from the true, divided by the true, and averaged over 1000 replicates. The K-M estimates are assumed to be the true. In order to ensure that Kendall's $\tau = 0.2$; the covariance parameters for the Copula-Graphic estimates for Frank copula was set to 1.9 and for the FGM estimates was set to 0.9. Figure 3 depicts the relative bias for the FGM estimates in solid lines and for the Copula-Graphic estimates in dashed lines. From top

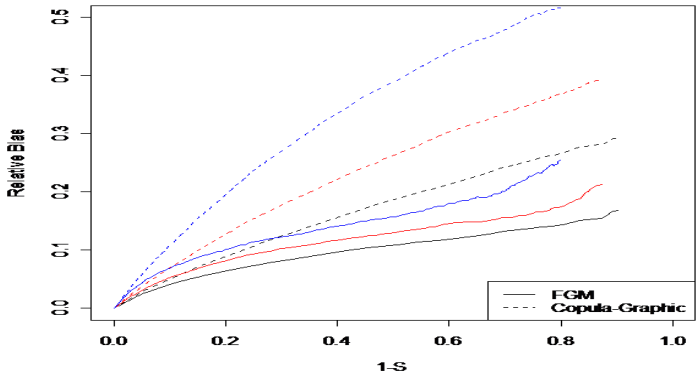

Figure 3: The relative bias of FGM estimates (solid lines), and Copula-Graphic estimates for Frank copula (dashed lines), for Kendalls $\tau = 0.2$. The drop-out rates from top to bottom are 30%, 15%, and 5%.

to bottom are the estimates for 30%, 15% and 5% drop-out, respectively. For all these drop-out scenarios, the FGM estimates have less bias than Copula-Graphic estimates.

6 Conclusion

The lack of identifiability of the competing risk model is a very well-known problem. In order to overcome this problem, we have proposed a sensitivity analysis which depends on unknown, inestimable covariance parameter, θ. We have given a simple interpretation of this parameter. An important advantage of our methodology comes from separating the censoring event into random and non-random components. We illustrated with a real data set how this non-random component distorts the survival estimates for different values of the selectivity parameter. For home dialysis data, where the percentage of drop-out is 12%, the survival curves are clearly distinguishable. The average survival time at maximum is also increased (or decreased) by 6%. The differences would be even more marked in a case of heavy censoring where the proportion of drop-outs is likely to be bigger or in a situation where the hazard of drop-out increases over time.

We showed that the proposed estimates give much narrower bounds than Copula-Graphic estimates. With simulated data for the drop-out proportion 30%, 15% and 5%, we have also demonstrated that the proposed estimates have less relative bias than Copula-Graphic estimates.

Acknowledgements

The authors are grateful to the Nephrology Department and Elisabeth Wells, Christchurch School of Medicine and Health Sciences, New Zealand, for providing the home dialysis data and also to Mohammed Lajmi Lakhal-Chaieb for providing R codes for Copula-Graphic estimator.

References

Copas, J. B. and H. Li (1997). Inference for non-random samples. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 59*(1), 55–95.

Delano, B. G. (1996). Home hemodialysis offers excellent survival. *Advances in Renal Replacement Therapy 3*, 106–111.

Fisher, L. and P. Kanarek (1974). Presenting censored survival data when censoring and survival times may not be independent. *Reliability and Biometry*, 303–326.

Kaciroti, N. A., T. E. Raghunathan, J. M. Taylor, and S. Julius (2012). A Bayesian model for time-to-event data with informative censoring. *Biostatistics 13*(2), 341–354.

Lagakos, S. W. and J. Williams (1978). Models for censored survival analysis: A cone class of variable-sum models. *Biometrika 65*(1), 181–189.

Lynn, K. L., D. O. McGregor, T. Moesbergen, A. L. Buttimore, J. A. Inkster, and J. E. Wells (2002). Hypertension as a determinant of survival for patients treated with home dialysis. *Kidney International 62*(6), 2281–2287.

Morgenstern, D. (1956). Einfache beispiele zweidimensionaler verteilungen. *Mitteilungsblatt für Mathematische Statistik 8*(1), 234–235.

Nelson, W. (1982). *Applied life data analysis*, New York: John Wiley.

Oakes, D. (1989). Bivariate survival models induced by frailties. *Journal of the American Statistical Association 84*(406), 487–493.

Peterson, A. V. (1976). Bounds for a joint distribution function with fixed sub-distribution functions: Application to competing risks. *Proceedings of the National Academy of Sciences 73*(1), 11–13.

Rivest, L.-P. and M. T. Wells (2001). A martingale approach to the copula-graphic estimator for the survival function under dependent censoring. *Journal of Multivariate Analysis 79*(1), 138–155.

Robins, J. M. and D. M. Finkelstein (2000). Correcting for noncompliance and dependent censoring in an aids clinical trial with inverse probability of censoring weighted (ipcw) log-rank tests. *Biometrics 56*(3), 779–788.

Rosenbaum, P. R. (1987). Sensitivity analysis for certain permutation inferences in matched observational studies. *Biometrika 74*(1), 13–26.

Rosenbaum, P. R. (1988). Sensitivity analysis for matching with multiple controls. *Biometrika 75*(3), 577–581.

Rotnitzky, A., D. Scharfstein, T.-L. Su, and J. Robins (2001). Methods for conducting sensitivity analysis of trials with potentially nonignorable competing causes of censoring. *Biometrics 57*(1), 103–113.

Schucany, W. R., W. C. Parr, and J. E. Boyer (1978). Correlation structure in Farlie-Gumbel-Morgenstern distributions. *Biometrika 65*(3), 650–653.

Shardell, M., D. O. Scharfstein, and S. A. Bozzette (2007). Survival curve estimation for informatively coarsened discrete event-time data. *Statistics in Medicine 26*(10), 2184–2202.

Slud, E. V. and L. V. Rubinstein (1983). Dependent competing risks and summary survival curves. *Biometrika 70*(3), 643–649.

Todem, D., K. Kim, J. Fine, and L. Peng (2010). Semiparametric regression models and sensitivity analysis of longitudinal data with non-random dropouts. *Statistica Neerlandica 64*(2), 133–156.

Williams, J. and S. Lagakos (1977). Models for censored survival analysis: Constant-sum and variable-sum model. *Biometrika 64*(2), 215–224.

Yuan, Y., P. F. Thall, and J. E. Wolff (2012). Estimating progression-free survival in paediatric brain tumour patients when some progression statuses are unknown. *Journal of the Royal Statistical Society: series C (Applied Statistics) 61*(1), 135–149.

Zheng, M. and J. P. Klein (1995). Estimates of marginal survival for dependent competing risks based on an assumed copula. *Biometrika 82*(1), 127–138.

A Log-Likelihood Function

The log-likelihood function for $i = 1, 2, \ldots, N$ subjects can be written as

$$\ell = \sum_{i=1}^{N} \Big[\delta_i \ln f_M(y_i \mid x_i) + (1-\delta_i) \ln S(y_i \mid x_i) + \delta_i \ln[1 + \theta(2F_M(y_i \mid x_i) - 1)(1 - \Phi(\boldsymbol{\gamma}^T x_i))] \\ + (1-\delta_i) \ln[1 + \theta F_M(y_i \mid x_i)(1 - \Phi(\boldsymbol{\gamma}^T x_i))] + \ln \Phi(\boldsymbol{\gamma}^T x_i)\Big],$$

where the indicator variable, δ_i , equals one for the subject with failure event and zero otherwise. Here ℓ is a function of β, σ, γ and $\theta(= \rho/0.3049)$. For any specific value of θ, the profile likelihood method will give the conditional maximum likelihood estimates as $\hat{\beta}(\theta)$, $\hat{\sigma}(\theta)$, $\hat{\boldsymbol{\gamma}}(\theta)$. The first order approximation of the above log-likelihood function is

$$\ell \cong \sum_{i=1}^{N} \Big[\delta_i \ln f_M(y_i|x_i) + (1-\delta_i) \ln S(y_i|x_i) + \theta[(1+\delta_i)F_M(y_i|x_i) - \delta_i](1 - \Phi(\boldsymbol{\gamma}^T x_i)) \\ + \ln \Phi(\boldsymbol{\gamma}^T x_i)\Big]$$

If the failure time, T, follows Weibull distribution then the log-likelihood reduces to

$$\ell = -r \ln \sigma + \sum_{i \in F} \varepsilon_i - \sum_{i=1}^{N} \Big\{ e^{\varepsilon_i} - \theta\big((1+\delta_i)F_M(y_i|x_i) - \delta_i\big)\big(1 - \Phi(\boldsymbol{\gamma}^T x_i)\big) - \ln \Phi(\gamma^T x_i) \Big\},$$

where $r = \sum_{i=1}^{N} \delta_i$ and F is the set of events with failure time and $\varepsilon_i = (y - \beta^T x_i)/\sigma$. First and second derivatives $w.r.t.\ \beta_l$ and σ would enable us to define the likelihood equation and the observed information matrix for given θ. Now, by using any optimization procedure for maximum likelihood, e.g. Newton-Raphson method or other iterative methods, we can get the estimates of $\hat{\beta}(\theta)$ and $\hat{\sigma}(\theta)$.

B Data Generation

Failure time (y_i) and potential drop-out time (x_i) for $i = 1, \ldots, 100$ have been generated as follows:

1. Generate $u \sim$ Uniform $(0, 1)$, $t \sim$ Uniform $(0, 1)$
2. Calculate $v = 1 - \left[\{(1-t)(1-u)^{\theta}\}^{-\frac{\theta-1}{\theta}} - (1-u)^{-(\theta-1)} + 1\right]^{-\frac{1}{\theta-1}}$
3. Calculate $x = -\dfrac{1}{\lambda_x}\ln(1-v)$
4. Calculate $y = -\dfrac{1}{\lambda_y}\ln(1-u)$

In step 2, θ was set to 1.5. In steps 3 and 4, λ_x and λ_y are selected in a way to ensure a desired proportion of drop-out occurs.