Published in final edited form as:

Scand Stat Theory Appl. 2015 March 1; 42(1): 87–103. doi:10.1111/sjos.12096.

Marginal and Conditional Distribution Estimation from Double-Sampled Semi-Competing Risks Data

Menggang Yu and

Department of Biostatistics & Medical Informatics, University of Wisconsin - Madison

Constantin T Yiannoutsos

Department of Biostatistics, Indiana University Purdue University Indianapolis

Abstract

Informative dropout is a vexing problem for any biomedical study. Most existing statistical methods attempt to correct estimation bias related to this phenomenon by specifying unverifiable assumptions about the dropout mechanism. We consider a cohort study in Africa that uses an outreach program to ascertain the vital status for dropout subjects. These data can be used to identify a number of relevant distributions. However, as only a subset of dropout subjects were followed, vital status ascertainment was incomplete. We use semi-competing risk methods as our analysis framework to address this specific case where the terminal event is incompletely ascertained and consider various procedures for estimating the marginal distribution of dropout and the marginal and conditional distributions of survival. We also consider model selection and estimation efficiency in our setting. Performance of the proposed methods is demonstrated via simulations, asymptotic study, and analysis of the study data.

Keywords

copula model; double sampling; informative dropout; semi-competing risks

1 Introduction

Competing risk data are common for time-to-event outcomes (Klein & Moeschberger 1997). In particular, the usual censored failure time data can be viewed as competing risk data when censoring is regarded as a competing cause of failure. More generally, study subjects may experience a number of distinct failure types. It is well-known that, with competing risk data, nonparametric estimation of the marginal and conditional distributions of the competing outcomes is generally infeasible (Tsiatis 1975).

Menggang Yu, Department of Biostatistics & Medical Informatics, University of Wisconsin, K6/446 CSC 600 Highland Ave., Madison, WI 53792-4675, USA, meyu@biostat.wisc.edu.

Supporting Information: Additional supporting material may be found in the online version of this article:

Appendix. Asymptotic properties of the estimator $\hat{\alpha}^{\pi}$.

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

Semi-competing risk data are related to two types of events, a terminal event and a non-terminal event, where a terminal event such as death is always observed but the non-terminal event (such as disease progression or treatment interruption) may be censored by the terminal event. Semi-competing risk data are therefore an enriched form of competing risk data in the sense that the non-terminal event is not a competing cause of the terminal event but not vice versa. This special structure allows estimation of the marginal distribution of the terminal event. Using copula models for the joint distribution of the two types of events, the marginal distribution of the non-terminal event can also be identified, and the correlation between the two events can be modeled explicitly (Fine *et al.* 2011, Wang 2003, Lakhal *et al.* 2008).

Semi-competing risk data are frequently encountered. A typical example is the illness-death situation (Fix & Neyman 1951, Sverdrup 1965, Xu *et al.* 2010), where death can censor the observation of illness if it occurs prior to illness, but occurrence of the illness may not prevent further follow-up on death. There are other semi-competing risk examples that do not fit into the illness-death framework. In AIDS studies, for example, time to first virologic failure and treatment discontinuation can be considered as the non-terminal and terminal events respectively (Jiang *et al.* 2005). In oncology, they can be times to local and distant recurrences respectively (Dignam *et al.* 2007). Other interesting examples exist in biomarker studies (Ghosh 2009, Day *et al.* 1997).

In this article, we consider the case when it is difficult or infeasible to ascertain all terminal events, especially in large cohort studies. This is the case of our motivating example, a cohort study of 8,977 adults who were enrolled between January 1, 2005 and January 31, 2007 in the Academic Model Providing Access To Healthcare (AMPATH) program (Mamlin *et al.* 2004). AMPATH, the Nobel Peace Prize-nominated HIV care and treatment program, is a partnership between the Indiana University and Moi University Schools of Medicine in western Kenya. A key component in the evaluation of the effectiveness of this program is the marginal survival as well as the conditional survival (on being retained in care) distributions for patients under care. However, estimation of the distributions are severely complicated by the fact that there is a very high rate of patient loss to follow-up (Wools-Kaloustian *et al.* 2008). Moreover, there was evidence that the individuals lost to follow-up were generally sicker than those who remained on observation (An *et al.* 2008, Yiannoutsos *et al.* 2008). These factors have the potential of introducing a significant bias if estimation is based only on data derived from patients under care.

Fortunately, AMPATH has instituted a major campaign to locate patients who are lost to follow-up. The AMPATH patient outreach program uses location information available on all patients in an effort to ascertain the whereabouts of missing patients and attempt to persuade them to return to care. The resource constraints enables only outreach to a subsample of these patients. In the process, the program records, among other data, the vital status of all patients sought and successfully located. Using outreach data and theory on double sampling (Frangakis & Rubin 2001), estimation of the marginal survival distribution is possible (An *et al.* 2008). In this article, by viewing dropout as the non-terminal event and death as the terminal event, we revisit this estimation and consider more efficient estimators within the semi-competing risk data framework. In addition, we consider estimation of the

conditional survival distributions and of the marginal distributions of the dropout in rural and urban areas that is informative for future policy making in providing care. Note that our considerations are applicable to the usual semi-competing risk situation (i.e. with 100% outreach), which is a special case of our setting.

Double sampling in time-to-event data is advocated as an important strategy to deal with dependent censoring (Baker & Kramer 2008). By ascertain some of the outcomes of the informatively censored subjects, the dependent censoring issues can be tackled while the costs of the studies can still be under control. In competing risks setting, double sampling for failure cause ascertainment is also advocated (Flehinger *et al.* 1998).

The rest of the article is organized as follows. In Section 2 we consider estimation for the dependency between the terminal and non-terminal events using copula models. In Section 3 we discuss estimation of marginal distributions for non-terminal and terminal events. In particular, we investigate how to improve the efficiency of marginal distribution estimation for the terminal event by using the non-terminal event information. In Section 4, we consider estimation of the conditional survival distribution for the terminal event, conditioning on the non-terminal event. In Section 5, we consider model selection among possible models. In Section 6 we assess the performance of our revised copula estimators through simulations, using realistic parameters from our experience in this setting and perform a reanalysis of the AMPATH database (An *et al.* 2008 and Yiannoutsos *et al.* 2008). We conclude with a brief discussion of the implications of this methodological development in Section 7.

2 Modeling the dependency between X and Y

Let X be the time to the non-terminal event, Y the time to the terminal event, and C the administrative censoring time. In semi-competing risk data, X can be censored by Y if $Y < X$, but is observable if $Y \geq X$. However both X and Y can be censored by C . Therefore the observable quantities are: $R = Y \wedge C$, $\delta_R = 1(Y < C)$, $S = X \wedge Y \wedge C = X \wedge R$, and $\delta_S = 1(X < R)$ where $\wedge$ is the minimum operator. Now suppose that, among all subjects with $\delta_S = 1$, we only observe R in a subset. Therefore, we have an indicator variable η for each subject such that R is only observed if $\eta = 1$. The probability of p for locating the dropouts is known because it is based on the outreach program. Denote $\pi \triangleq P(\eta = 1) = \delta_S p + (1 - \delta_S)$. Our objective is to estimate the marginal survival distributions $F_X(x) = P(X > x)$ and $F_Y(y) = P(Y > y)$, and the conditional survival distributions $P(Y > y | X = x, Y > t)$ and $P(Y > y | X > x, Y > t)$ for $y > t > x$.

We assume the following copula model for the joint survival distribution of X and Y (Oakes 1989):

$$F(x, y) = P(X > x, Y > y) = C_\alpha\{F_X(x), F_Y(y)\}$$

We are particularly interested in a class of copulas indexed by a single parameter α . A possible choice for C_α is the well known Archimedean copula with generator φ_α (Fine *et al.* 2001, Wang 2003),

$$C_\alpha(u, v) = \phi_\alpha^{-1}\{\phi_\alpha(u) + \phi_\alpha(v)\}, 0 \leq u, v \leq 1$$

Popular choices of φ_α include the Clayton copula (Clayton 1978): $\varphi_\alpha(t) = t^{1-\alpha} - 1$ and the Gumbel copula: $\varphi_\alpha(t) = (-\log t)^\alpha$, $\alpha \geq 1$. Because $p > 0$ and $\pi > 0$, the estimation function is consistent to the setting with fully observed semi-competing risks data, the identifying conditions of Ding (2010) are satisfied. The Archimedean copula with analytic generating function leads to identifiable double sampled semi-competing risks data.

A quantity which is useful in the effort to explicitly model the dependency between the terminal and non-terminal event is the cross-ratio function (Oakes 1989):

$$\theta_\alpha^*(x, y) = \frac{P\{(X_i - X_j)(Y_i - Y_j) > 0 | X_i \wedge X_j = x, Y_i \wedge Y_j = y\}}{P\{(X_i - X_j)(Y_i - Y_j) < 0 | X_i \wedge X_j = x, Y_i \wedge Y_j = y\}}. \quad (1)$$

This function is related to Kendall's tau (Wang 2003, Lakhal *et al.* 2008). Under the Archimedean copulas, θ_α depends on x and y only through their joint survival distribution $F(x, y)$ and takes the form

$$\theta_\alpha^*(x, y) = -F(x, y) \frac{\phi_\alpha''(F(x, y))}{\phi_\alpha'(F(x, y))} \triangleq \theta_\alpha F(x, y), \quad (2)$$

where $\theta_\alpha(v) = -v\phi_\alpha''(v)/\phi_\alpha'(v)$ for any v .

Utilizing (2), Lakhal *et al.* (2008) proposed the following estimating equation for general Archimedean copulas:

$$\Psi_n(\alpha) = \binom{n}{2}^{-1} \sum_{i < j} W(\tilde{S}_{ij}, \tilde{R}_{ij}) Z_{ij} \left\{ \Delta_{ij} - \frac{\theta_\alpha\{\hat{F}(\tilde{S}_{ij}, \tilde{R}_{ij})\}}{1 + \theta_\alpha\{\hat{F}(\tilde{S}_{ij}, \tilde{R}_{ij})\}} \right\} \quad (3)$$

where $\tilde{S}_{ij} = S_i \wedge S_j$, $\tilde{R}_{ij} = R_i \wedge R_j$, and $\tilde{C}_{ij} = C_i \wedge C_j$; $\Delta_{ij} = 1\{(S_i - S_j)(R_i - R_j) > 0\}$; $Z_{ij} = 1(\tilde{S}_{ij} < \tilde{R}_{ij} < \tilde{C}_{ij})$; and W is a weighting function. Because

$$F(x, y) = \frac{P(S > x, R > y)}{P(C > y)}, \quad (4)$$

$F(x, y)$ is estimated by

$$\hat{F}(x, y) = \frac{n^{-1} \sum_{i=1}^n 1(S_i > x, R_i > y)}{\hat{G}(y)}$$

with

$$\hat{G}(y) = \prod_{i: C_i \leq y} \left\{ 1 - \frac{1 - \Delta_{R_i}}{\sum_{k=1}^n 1(R_k \geq y)} \right\}$$

being the Kaplan-Meier estimator for the censoring distribution. Z_{ij} is needed in (3) because ij is evaluable only when $Z_{ij} = 1$ (Lakhal et al. 2008, Fine et al. 2001) and in this case ij equals the concordance between (X_i, Y_i) and (X_j, Y_j) , that is, $1\{(X_i - X_j)(Y_i - Y_j) > 0\}$. The weighting function may be taken as 1 but a preferred form is given by (Fine et al. 2001)

$$W_{a,b}(x, y) = \frac{n}{\sum_{i=1}^n 1\{S_i \geq a \wedge x, R_i \geq b \wedge y\}} \quad (5)$$

where a and b are constants and may be selected to down weight for ‘large’ x and y .

In our case, because for subjects with $S_i < R_i$, R_i is observable only when $\eta_i = 1$, we propose the following estimating function when double sampling data are available:

$$\Psi_n^\pi(\alpha) = \binom{n}{2}^{-1} \sum_{i < j} \frac{\eta_i \eta_j}{\pi_i \pi_j} W^\pi(\tilde{S}_{ij}, \tilde{R}_{ij}) Z_{ij} \left\{ \Delta_{ij} - \frac{\theta_\alpha \{ \hat{F}^\pi(\tilde{X}_{ij}, \tilde{Y}_{ij}) \}}{1 + \theta_\alpha \{ \hat{F}^\pi(\tilde{X}_{ij}, \tilde{Y}_{ij}) \}} \right\}. \quad (6)$$

We can take the weighting function W^π as 1, or corresponding to (5), use

$$W_{a,b}^\pi(x, y) = \frac{n}{\sum_{i=1}^n \frac{\eta_i}{\pi_i} 1\{S_i \geq a \wedge x, R_i \geq b \wedge y\}} \quad (7)$$

$F^{\hat{\pi}}(x, y)$ estimates $F(x, y)$ and again it takes the following form because of (4),

$$\hat{F}^\pi(x, y) = \frac{\hat{P}(S > x, R > y)}{\hat{P}(C > y)} \quad (8)$$

When double sampled data are available, various estimators can be used for the numerator and denominator in (8). For $\hat{P}(S > x, R > y)$, we can use

$$\hat{P}^{\pi,1}(S > x, R > y) = n^{-1} \sum_{i=1}^n \frac{\eta_i}{\pi_i} 1(S_i > x, R_i > y) \quad (9)$$

which inversely weighs the observations because of double sampling. An alternative choice is

$$\hat{P}^{\pi,2}(S > x, R > y) = n^{-1} \sum_{i=1}^n \left\{ 1(S_i > y) + \frac{\eta_i}{\pi_i} 1(S_i > x, S_i \leq y, R_i > y) \right\} \quad (10)$$

where inverse weighting is only used for the subjects that need double sampling to ascertain the indicator $1(S_i > x, R_i > y)$. Because (10) utilizes more information, we use it in this article.

There are also two choices to estimate the censoring distribution in (8) under the double sampled data. One is

$$\hat{G}^{\pi}(y) = \prod_{i: R_i \leq y} \left\{ 1 - \frac{\eta_i}{\pi_i} \frac{1 - \Delta_{Ri}}{\sum_{k=1}^n \frac{\eta_k}{\pi_k} 1(R_k \geq y)} \right\},$$

which utilizes inverse weighting. The other is

$$\tilde{G}(y) = \prod_{i: S_i \leq y} \left\{ 1 - \frac{(1 - \Delta_{Si})(1 - \Delta_{Ri})}{\sum_{k=1}^n 1(S_k \geq y)} \right\}.$$

which uses $X \wedge Y$ as a censoring variable for C and does not involve inverse weighting. We use $\tilde{G}(y)$ in (8) because greater efficiency gains were observed in simulations.

By recognizing that the weighted estimator is related to U-statistics, we derive the asymptotic properties of the resulting estimator $\hat{\alpha}^{\pi}$ from (6) as Supporting information online. Because the asymptotic variance is complicated, we use a bootstrap procedure to obtain standard error estimates as in Lakhal *et al.* (2008).

Finally, we note that for the Clayton copula, $\theta_{\alpha}^* = \alpha$ is not related to either x or y and

$P(\Delta_{ij}=1) = \frac{\alpha}{1+\alpha}$ for any $i < j$ (Oakes 1989, Fine et al. 2001). Then it is easily seen that an explicit solution for α exists from (3):

$$\hat{\alpha}^{\pi} = \frac{\sum_{i < j} \frac{\eta_i \eta_j}{\pi_i \pi_j} W^{\pi}(\tilde{S}_{ij}, \tilde{R}_{ij}) Z_{ij} \Delta_{ij}}{\sum_{i < j} \frac{\eta_i \eta_j}{\pi_i \pi_j} W^{\pi}(\tilde{S}_{ij}, \tilde{R}_{ij}) Z_{ij} (1 - \Delta_{ij})}.$$

When $\pi \equiv 1$ in the semi-competing risk data, this reduces to the estimator from Fine *et al.* (2001).

3 Estimation of marginal distributions

3.1 Estimation of the marginal survival distribution of Y

Adopting the counting process notation (Kalbfleisch and Prentice 2002), we estimate the cumulative hazard function of Y by

$$\hat{\Lambda}_Y^{\pi,1}(t) = \int_0^t \frac{\sum_{i=1}^n (1 - \delta_{S_i}) dN_i(u) + \sum_{i=1}^n \frac{\eta_i}{\pi_i} \delta_{S_i} dN_i(u)}{\sum_{i=1}^n (1 - \delta_{S_i}) 1(R_i \geq u) + \sum_{i=1}^n \frac{\eta_i}{\pi_i} \delta_{S_i} 1(R_i \geq u)}$$

where $N_i(u) = 1(Y_i \leq u, C_i > Y_i)$. This estimator was proposed in Frangakis and Rubin (2001). The corresponding survival function estimate is then

$$\hat{F}_Y^{\pi,1}(t) = \exp \left\{ - \sum_{i=1}^n \hat{\Lambda}_{Y_i}^{\pi,1}(t) \right\}. \quad (11)$$

An alternative way is to estimate the cumulative hazard function of Y by

$$\hat{\Lambda}_Y^{\pi,2}(t) = \int_0^t \frac{\sum_{i=1}^n (1 - \delta_{S_i}) dN_i(u) + \sum_{i=1}^n \frac{\eta_i}{\pi_i} \delta_{S_i} dN_i(u)}{\sum_{i=1}^n (1 - \delta_{S_i}) 1(R_i \geq u) + \sum_{i=1}^n \delta_{S_i} 1(S_i \geq u) + \sum_{i=1}^n \frac{\eta_i}{\pi_i} \delta_{S_i} 1(S_i < u) 1(R_i \geq u)}.$$

This estimator has been mentioned in Robins, Rotnitzky, and Bonetti (2001). Notice that the calculation of the at-risk set at any time u uses the similar idea as in the estimation of $P^{\hat{\pi},2}(S > x, R > y)$ in (10) where inverse weighting is only used for the subjects that need double sampling to ascertain the at risk status. The corresponding survival function estimate is then

$$\hat{F}_Y^{\pi,2}(t) = \exp \left\{ - \sum_{i=1}^n \hat{\Lambda}_{Y_i}^{\pi,2}(t) \right\}. \quad (12)$$

Asymptotics for both (11) and (12) can also be established in a similar fashion as in Yu & Nan (2010).

Neither approach utilizes the correlation between X and Y . We therefore consider a model-based estimator. In particular, by partitioning $P(Y > t)$ as

$$P(Y > t) = P(Y \wedge C > t) + P(Y > t, C \leq t, X < C) + P(Y > t, C \leq t, X > C),$$

we propose the following estimate

$$\begin{aligned}
\hat{F}_Y^{\pi, M}(t) &= \frac{1}{n} \sum_{i=1}^n 1(R_i > t) \left\{ (1 - \delta_{S_i}) + \frac{\eta_i}{\pi_i} \delta_{S_i} \right\} \\
&\quad + \frac{1}{n} \sum_{i=1}^n 1(R_i \leq t) (1 - \delta_{R_i}) \frac{\eta_i}{\pi_i} \delta_{S_i} \hat{P}(Y > t | X = S_i, Y > R_i) \\
&\quad + \frac{1}{n} \sum_{i=1}^n 1(R_i \leq t) (1 - \delta_{R_i}) \hat{P}(Y > t | X > S_i, Y > S_i).
\end{aligned} \tag{13}$$

Here the $P(\hat{Y} > t | X = S_i, Y > R_i)$ and $P(\hat{Y} > t | X > S_i, Y > S_i)$ are conditional estimates of the corresponding probabilities. The estimates will be discussed in Section 4.

As we demonstrate in our numerical section, compared to (11) and (12), $\hat{F}_Y^{\pi, M}(t)$ improves efficiency noticeably when there is moderate to high correlation between X and Y . Such improvement is important especially when a large number of subjects experience X and double sampling for Y is costly (i.e., p is small).

3.2 Estimation of the marginal distribution of X

We adapt the copula graphic (CG) estimator for the marginal distribution $F_X(x)$ from Lakhal *et al.* (2008) in our double sampled data. The CG estimator utilizes the following relationship (Zheng & Klein 1995):

$$F_X(X) = \phi_\alpha^{-1} \{ \phi_\alpha[F(x, x)] - \phi_\alpha[F_Y(x)] \}$$

where $F(x, x) = P(X \wedge Y > x) \triangleq F_{X \wedge Y}(x)$ and F_Y is the marginal distribution of Y . By replacing unknown parameters with estimates, the resulting CG estimator is

$$\hat{F}_X^{CG}(x) = \phi_\alpha^{-1} \left\{ \sum_{S_i \leq x, \delta_{S_i} = 1} \phi_\alpha[\hat{F}_{X \wedge Y}(S_i)] - \phi_\alpha[\hat{F}_{X \wedge Y}(S_i -)] \right\}. \tag{14}$$

Here $\hat{F}_{X \wedge Y}$ is the Kaplan-Meier estimator for $X \wedge Y$ which needs no inverse weighting because observation of $X \wedge Y$ is unaffected by double sampling. Note that because F_Y is discrete and jumps only at observed failure times of Y , the term involving F_Y disappears in (14) (Rivest & Wells 2001, Lakhal *et al.* 2008). The CG estimator is therefore a discrete function that jumps only at observed failure times of X .

In the presence of double sampling, because S_i are always observed and therefore are used in the same way to estimate $F(x, x)$ and because F_Y is not involved in (14), the CG estimator has exactly the same format as in (14), i.e.,

$$\hat{F}_X^{\pi, CG}(x) = \phi_{\hat{\alpha}^\pi}^{-1} \left\{ \sum_{S_i \leq x, \delta_{S_i}=1} \phi_{\hat{\alpha}^\pi} [\hat{F}_{X \wedge Y}(S_i)] - \phi_{\hat{\alpha}^\pi} [\hat{F}_{X \wedge Y}(S_i -)] \right\}. \quad (15)$$

The asymptotic distribution of this estimator is therefore exactly the same as in the complete ascertainment case, except that we replace $\hat{\alpha}$ with $\hat{\alpha}^\pi$ given as a solution of (6), taking into consideration the asymptotic normality of $\hat{\alpha}^\pi$ from the Supporting information online.

4 Estimation of Conditional Distributions

The conditional survival function of a patient that has not dropped out at time x is $P(Y > y | X > x; Y > x)$. Note here that the ability to evaluate this conditional probability is particularly relevant in our context, since it is useful in directly addressing the question of what the survival distribution is for a patient who is retained in care up to time $x > x^*$, where x^* is a clinically meaningful threshold (e.g., three or six months from start of antiretroviral therapy, a critical period in the care and treatment of HIV-infected patients initiating therapy).

Just like in the estimation of the marginal distribution of Y , two different estimators can be used for $P(Y > y | X > x; Y > x) = F(x, y)/F(x, x)$. One is nonparametric:

$$\hat{P}^\pi(Y > y | X > x; Y > x) = \frac{\hat{F}^\pi(x, y)}{\hat{F}_{X \wedge Y}^\pi(x)}$$

where $\hat{F}^\pi$ is defined in (8). The other is model-based:

$$\hat{P}^M(Y > y | X > x; Y > x) = \frac{C_{\hat{\alpha}^\pi} \{ \hat{F}_X^{\pi, CG}(x), \hat{F}_Y^{\pi, 2}(y) \}}{C_{\hat{\alpha}^\pi} \{ \hat{F}_X^{\pi, CG}(x), \hat{F}_Y^{\pi, 2}(x) \}} \quad (16)$$

with $\hat{F}_X^{\pi, CG}$ from (15) and $\hat{F}_Y^{\pi, 2}$ from (12). Of course $\hat{F}_Y^{\pi, 1}$ from (11) can also be used in (16)

with some loss of efficiency. Obviously another more natural choice is $\hat{F}_Y^{\pi, M}$ from (13).

Notice that because (13) involves both $P(Y > t | X > S_i, Y > S_i)$ and $P(Y > t | X = S_i, Y > R_i)$, the

estimation of $P(Y > t | X > S_i, Y > S_i)$ will be iterative if we replace $\hat{F}_Y^{\pi, 2}$ with $\hat{F}_Y^{\pi, M}$ in (16).

From our simulation, we did not observe any noticeable improvement in efficiency. On the other hand, computation of the iterative estimator became more burdensome. We therefore only report results based on (16) in Section 6.

The conditional survival function of a patient who drops out at time x and is alive at x is $P(Y > y | X = x; Y > x)$ for $y > x$. It satisfies (Lakhal *et al.* 2008)

$$P(Y > y | X = x; Y > x) = \frac{\phi'_\alpha \{F(x, x)\}}{\phi'_\alpha \{F(x, y)\}} = \frac{\phi'_\alpha \{F_{X \wedge Y}(x)\}}{\phi'_\alpha \{F(x, y)\}}.$$

with $\phi'_\alpha(t)$ the derivative of $\phi_\alpha(t)$ with respect to t . We can then obtain an estimator by plugging in estimators for α , $F_{X \wedge Y}(x)$, and $F(x, y)$. Similar to estimation of $P^{\hat{M}}(Y > y | X > x; Y > x)$, we can use a nonparametric estimator,

$$\hat{P}^\pi(Y > y | X = x; Y > x) = \frac{\phi'_{\hat{\alpha}^\pi}[\hat{F}^\pi(x, x)]}{\phi'_{\hat{\alpha}^\pi}[\hat{F}_{X \wedge Y}^\pi(x)]}$$

or a model-based estimator,

$$\hat{P}^M(Y > y | X = x; Y > x) = \frac{\phi'_{\hat{\alpha}^\pi}[C_{\hat{\alpha}^\pi}\{\hat{F}_X^{\pi, CG}(x), \hat{F}_Y^{\pi, 2}(y)\}]}{\phi'_{\hat{\alpha}^\pi}[C_{\hat{\alpha}^\pi}\{\hat{F}_X^{\pi, CG}(x), \hat{F}_Y^{\pi, 2}(x)\}]} \quad (17)$$

The performance of all these semi-parametric estimators of the conditional survival distributions is explored through simulations.

5 Model selection

The idea is to choose a quantity that is both estimable by the data (nonparametrically) and by the proposed models. Among the proposed models, the model which produces the closest estimate to the nonparametrically estimated quantity is deemed the best. Note that the quantity, which is estimated by each candidate model, should not have been used already in the estimating procedure and be sensitive to different models. One such quantity is $F(x, y; \delta_S = \delta_R = 1) \triangleq P(X > x, Y > y, S = 1, R = 1)$, which is estimable nonparametrically by

$$\hat{F}(x, y; \delta_S = \delta_R = 1) \triangleq n^{-1} \sum_{i=1}^n \frac{\eta_i}{\pi_i} 1(S_i > x, R_i > y, \Delta_{S_i} = 1, \Delta_{R_i} = 1).$$

Under an Archimedean copula model,

$$F(x, y; \delta_S = \delta_R = 1) = \int_y^\infty \left\{ \frac{\phi'_\alpha[F_Y(t)]f_Y(t)}{\phi'_\alpha[F(x, t)]} - \frac{\phi'_\alpha[F_Y(t)]f_Y(t)}{\phi'_\alpha[F(t, t)]} \right\} G(t) dt$$

Therefore it can also be estimated by

$$\tilde{F}(x, y; \delta_S = \delta_R = 1) = \int_y^\infty \left\{ \frac{\phi'_{\hat{\alpha}^\pi}[\hat{F}_Y^{\pi,2}(t)]}{\phi'_{\hat{\alpha}^\pi}[\hat{F}^\pi(x, t)]} - \frac{\phi'_{\hat{\alpha}^\pi}[\hat{F}_Y^{\pi,2}(t)]}{\phi'_{\hat{\alpha}^\pi}[\hat{F}^\pi(t, t)]} \right\} \hat{G}(t) d\hat{F}_Y^{\pi,2}(t).$$

The maximum distance, $\mathcal{D} \triangleq \max_{0 < x < y} |F(\hat{x}, y; \delta_S = \delta_R = 1) - \tilde{F}(\hat{x}, y; \delta_S = \delta_R = 1)|$ is then used as a criterion to choose different models. In other words, among possible AC models, the model with the smallest difference measure $\mathcal{D}$ is chosen. This builds on the model checking method of Hsieh *et al.* (2008).

Of course, it is likely that no model among the proposed ones provides adequate fit to the data. In this case, calibration by p-values may be desirable for model selection purposes (Hsieh *et al.* 2008). A similar parametric bootstrap approach was taken to approximate the null distribution of the difference measure. Specifically, samples of (X^*, Y^*) were generated from the estimated copula models $\phi_{\hat{\alpha}}^{-1}\{\phi_{\hat{\alpha}}(\hat{F}_X(x)) + \phi_{\hat{\alpha}}(\hat{F}_Y(y))\}$ with $\hat{\alpha}$, $\hat{F}_X$, and $\hat{F}_Y$ estimated from our proposed methods. The corresponding samples of the censoring variable C^* were generated from $\hat{G}$. The bootstrap samples were then censored according to the semi-competing risk framework. The resulting observed values were $R^* = Y^* \wedge C^*$, $\delta_{R^*} = 1(Y^* < C^*)$, $S^* = X^* \wedge Y^* \wedge C^*$, and $\delta_{S^*} = 1(X^* < R^*)$. The model fitting measures were then calculated for all the bootstrap samples to create the null distribution.

6 Numerical examples

In this section we assess the performance of various proposed estimators and the model selection procedure. We also analyze the data from the AMPATH study.

6.1 Simulations

In generating data, the marginal distribution of Y is taken to be unit exponential and the marginal distribution of X exponential with rate 0.8. This leads to about 40 percent dropout which is close to the rate of 39.3% reported in An, Yiannoutsos, Frangakis *et al.* (2008). The censoring distribution is uniform within the interval (0, 4). To save space, we present results using the Clayton and Gumbel copulae with various α for the association between X and Y . We used the bootstrap procedure for standard error estimation in all analyses.

In the following sections, we present results for estimation of the association between X and Y in Table 1; for the marginal distributions of X and Y in Table 2; and for the conditional distributions of Y given X in Table 3. A total of 400 data sets were simulated in each scenario with various correlations between X and Y . The sample sizes were set to be 1000 and 3000 in each setting. The double sampling probability p is set as 0.2.

In Table 1, we see that our estimation procedure performs satisfactorily in all cases. The standard error decreases with larger sample size $n = 3000$ as compared with $n = 1000$. In Table 2, we try to estimate the marginal distribution of X at one year (i.e., $t = 1$) and the marginal distribution of Y at 2 years ($t = 2$). The CG estimates are all close to the true value $\exp(-1) = 0.368$ and the estimates for $F_Y(2)$ are all close to the true value $\exp(-0.8 * 2) = 0.202$. Among the standard error estimates of the three estimators of F_Y , we see that $\hat{F}_Y^{\pi,1}$ is

the least efficient and $\hat{F}_Y^M$ the most efficient among the three. The improvement of efficiency is actually related to the strength of correlation between X and Y . Figure 1 illustrates the efficiency comparison among these marginal distribution estimators under the two copulae. Here the x-axis represents the association between X and Y based on Kendall's tau. The results are based on the same setting as in Table 1 but with 50,000 simulated data.

We see that $\hat{F}_Y^{\pi,2}$ is in general more efficient than $\hat{F}_Y^{\pi,1}$ except in the case of very high associations. When the association is small, the model-based estimator can be inefficient. However, the efficiency improves noticeably as association increases. This implies that the model-based estimator may be preferred if there is moderate to high associations.

Based on the above observation, one referee suggested a hybrid estimator $\hat{F}_Y^{\pi,hybrid}$ that equals $\hat{F}_Y^{\pi,M}$ when Kendall's tau between X and Y exceeds certain threshold, τ_0 , and $\hat{F}_Y^{\pi,2}$ otherwise. We redid the simulation to incorporate this suggestion. Based on Figure 1, we set $\tau_0 = 0.4$ for the Clayton copula based analysis and $\tau_0 = 0.5$ for the Gumbel copula based analysis. Indeed the empirical variance of the hybrid estimator improved on $\hat{F}_Y^{\pi,M}$ and $\hat{F}_Y^{\pi,2}$ except in the case when Kendall's tau is around τ_0 . In such case, $\hat{F}_Y^{\pi,hybrid}$ has efficiency in between $\hat{F}_Y^{\pi,M}$ and $\hat{F}_Y^{\pi,2}$. We have observed that ρ_0 can depend on a particular estimand (e.g. whether we are estimating $F_Y(1)$ or $F_Y(2)$). Therefore in practice, the cut-off ρ_0 needs to change accordingly to ensure a noticeable help from the hybrid estimator. Overall, the hybrid estimator can still be quite interesting.

Table 3 gives results for estimation of conditional distributions $F_{1.5|.75} \triangleq P(Y > 1.5|X = 0.75, Y > 0.75)$ and $F_{1.5|.75+} \triangleq P(Y > 1.5|X > 0.75, Y > 0.75)$. Again the results are satisfactory. Finally, we conducted model selection procedures under the same setting with results listed in Table 4. We can see that correct models were selected with very high percentage in all cases. The correct selection percentages are higher with the larger sample size as expected.

6.2 The AMPATH study

The study included 8,977 adults coming from both urban and rural clinics in western Kenya between January 1, 2005 and January 31, 2007. There is a very high rate of patient loss to follow-up (3,528 dropouts). In the outreach program, 621 were double sampled for further follow-up. The initial goal of AMPATH was to establish an HIV care system to serve the needs of both urban and rural patients and to assess the barriers to and outcomes of antiretroviral therapy. It is important to explore differences in dropout patterns and patient survival between the rural and urban areas. For the urban area, there were 6561 subjects among whom 38% dropped out. The outreach rate is 18%. The maximum follow up is 761 days with a total of 136 deaths observed. For the rural area, there were 2416 subjects among whom 43% dropped out. The outreach rate is 23%. The maximum follow up is 761 days with a total of 86 deaths observed.

The analysis results listed in Table 5 are based on the Clayton copula which was selected based on the difference measure proposed in Section 5. The marginal distributions for dropout and survival in rural and urban clinics are plotted in Figure 2. The conditional survival distributions in rural and urban clinics are plotted in Figure 3. The one-year estimate of mortality generated from the estimated marginal distribution is consistent with that in Yiannoutsos *et al.* (2008). Outreach-adjusted one-year mortality from treatment initiation is about 10% (Table 5).

In Table 5, the estimate of the α index of the Clayton copula (s.e.) was 2.865 (0.481) among patients enrolled in the urban clinic and 1.852 (0.419) for rural clinics. These estimates correspond to Kendall's tau of 0.59 and 0.48, which in turn suggests that patient dropout is non-ignorable. We can also see that the rural clinics appear to have inferior outcomes in many respects (Table 5). Patient retention rate at 1 year is 59.8% compared with 62.4% in the urban area. Patients retained in care for 3 or 6 months, also have shorter survival when attending rural clinics versus receiving care from the referral hospital. From the marginal distribution curve of dropout, we find that the dropout rates are similar in the early period (up to about 180 days) and then the dropout rate increases in the rural area (Figure 2).

Our analysis showed the advantages of being able to estimate conditional survival distributions for the terminal event given the non-terminal event. We considered patient dropout at 3 and 6 months from start of therapy, two important landmarks when the hazard of mortality was highest (Yiannoutsos, 2009). Conditioning on being retained on treatment (and thus being alive) at 3 and 6 months after therapy produced estimates of mortality (Figure 3 right panel), which were much lower than overall (marginal) mortality (Figure 2 right panel). In addition, patient retention for at least 6 months is associated with substantial improvement of mortality (both early and late) compared to retention at least for 3 months (Figure 3). These results have important policy implications about interventions which target patients early after initiating antiretroviral therapy. AMPATH for example, has instituted an intense follow-up of patients during the first 3 months after treatment initiation (Braitstein *et al.* 2012). These analyses provide strong supportive evidence for such interventions.

7. Discussion

In this article, we considered various estimation aspects and model selection for mortality and dropout in the context of a large HIV care and treatment program in western Kenya. We developed novel statistical methods in a semi-competing risk framework where we were able to ascertain only a fraction of the terminal events. By utilizing the correlation between the non-terminal and terminal events, we demonstrated that improved estimation for the marginal and conditional distributions of the terminal event is feasible as is the marginal distribution of the non-terminal event. In addition to the asymptotic study, we performed simulations under scenarios closely resembling the real-life settings where these methods will be applied. In all cases, the performance of the estimators was excellent.

In practice, this means that HIV care and treatment program effectiveness evaluation by metrics such as mortality, patient retention, treatment interruption or disease progression rates, all falling under our semi-competing risk framework, can be accomplished by locating

only a relatively small subset of patients who are lost. This ensures the viability and sustainability of ongoing program evaluation efforts in this setting as most programs in low and middle-income countries in sub-Saharan Africa, Asia and Latin America are so large and the loss to follow-up problem so pronounced (Rosen *et al.* 2007) as to make it infeasible to follow-up all patients who discontinue treatment.

While these analyses are intended more as a proof of concept rather than an exhaustive analysis of all possible explanatory factors of dropout and mortality patterns in our setting, these first results underline some important issues and further emphasize the practical utility of these methods. The ability to explicitly model the possible correlation between X and Y allows both the assessment of the strength of the association as well as the comparison of these quantities between groups. In our re-analysis of the AMPATH data, we see a possibly different pattern in the urban versus the rural settings with dropout being associated less strongly with mortality in the rural setting (although, in both cases, dropout is strongly informative). These observations are important because they underscore the necessity of methods which do not rely on independence assumptions between X and Y and the absolute critical nature of the availability of data from outreach and vital status ascertainment among lost patients. In addition, these findings could be used to tailor contextually appropriate interventions depending on the setting (rural versus urban).

While our motivating example was focused on issues arising in the evaluation of large HIV/AIDS care and treatment programs recently proliferating in low and middle-income countries, the applicability of the proposed methods is quite broad. The extent of the application of this methodology is related both to the number of issues in the present context as well as any other chronic disease intervention requiring patient retention and long-term adherence to treatment or prevention regimen.

To investigate the influence of model misspecification, we also conducted sensitivity analysis on the Archimedean model assumption. We simulated settings where data were generated from non-Archimedean copula and from the Frank Archimedean copula. For non-Archimedean copula, we generated data from the normal copula where the correlation coefficient ρ for the generating function is taken as 0.3, 0.6, or 0.9. For the Frank copula, we took the association parameter α as 2, 3, or 4. The analysis was still conducted using the Clayton and Gumbel copulae. We observed that for interpretable quantities such as the Kendall's tau, marginal, and conditional survival probabilities, the analysis results appeared to be quite robust to model assumptions, especially for the Gumbel copula based analysis. The relative bias was generally less than 5% for the Gumbel copula analysis and less than 15% for the Clayton analysis.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Dr. Yiannoutsos' work was supported by National Institute of Allergies and Infectious Diseases (NIAID) grants number AI-069911, for the East Africa Regional IeDEA Consortium, and AI102710. The data were collected

through a grant to the USAID-AMPATH Partnership from the United States Agency for International Development as part of PEPFAR. The funders had no role in study design, interpretation of results, or decision to publish.

References

- An MW, Frangakis CE, Musick BS, Yiannoutsos CT. The need for double-sampling designs in survival studies: an application to monitor PEPFAR. *Biometrics*. 2008; 65:301–306. [PubMed: 18479488]
- Baker SG, Karmner BS. Randomized trials for the real world: making as few and as reasonable assumptions as possible. *Stat Methods Med Res*. 2008; 17:243–52. [PubMed: 17925312]
- Braitstein P, Siika A, Hogan JW, Kosgei R, Sang E, Sidle J, Wools-Kaloustian K, Keter A, Mamlin J, Kimaiyo S. A clinician-nurse model to reduce early mortality and increase clinic retention among high-risk HIV-infected patients initiating combination antiretroviral treatment. *J Int AIDS Soc*. 2012; 15:7. [PubMed: 22340703]
- Clayton DG. A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika*. 1978; 65:141–151.
- Day R, Bryant J, Lefkopoulou M. Adaptation of bivariate frailty models for prediction, with application to biological markers as prognostic indicators. *Biometrika*. 1997; 84:45–56.
- Dignam JJ, Wieand K, Rathouz PJ. A missing data approach to semi-competing risks problems. *Stat Med*. 2007; 26:837–56. [PubMed: 16755544]
- Ding AA. Identifiability conditions for covariate effects model on survival times under informative censoring. *Statist Probab Lett*. 2010; 80:911–5.
- Fine JP, Jiang H, Chappell R. On semicompeting risks. *Biometrika*. 2001; 88:907–919.
- Fix E, Neyman J. A simple stochastic model of recovery, relapse, death and loss of patients. *Human Biol*. 1951; 23:205–41. [PubMed: 14880165]
- Flehinger BJ, Reiser B, Yashchin E. Survival with Competing Risks and Masked Causes of Failures. *Biometrika*. 1998; 85:151–64.
- Frangakis CE, Rubin DB. Addressing an idiosyncrasy in estimating survival curves using double sampling in the presence of self-selected right censoring. *Biometrics*. 2001; 57:333–342. [PubMed: 11414553]
- Ghosh D. On assessing surrogacy in a single trial setting using a semicompeting risks paradigm. *Biometrics*. 2009; 65:521–9. [PubMed: 18759839]
- Hsieh JJ, Wang W, Ding AA. Regression analysis based on semicompeting risks data. *J R Stat Soc B*. 2008; 70:320.
- Jiang H, Fine JP, Kosorok MR, Chappell R. Pseudo self-consistent estimation of a copula model with informative censoring. *Scand J Stat*. 2005; 32:1–20.
- Klein, JP.; Moeschberger, ML. *Survival analysis: Techniques for censored and truncated data*. New York: Springer; 1997.
- Kalbfleisch, JD.; Prentice, RL. *The Statistical Analysis of Failure Time Data*. Second. New York: John Wiley & Sons, Inc; 2002.
- Lakhal L, Rivest LP, Abdous B. Estimating survival and association in a semicompeting risks model. *Biometrics*. 2008; 64:180–188. [PubMed: 17645782]
- Mamlin, J.; Kimaiyo, S.; Nyandiko, W.; Tierney, W.; Einterz, R. *Perspectives and Practice in Antiretroviral Treatment*. Geneva: World Health Organization; 2004. Academic institutions linking access to treatment and prevention: Case study.
- Oakes D. Bivariate survival models induced by frailties. *J Amer Statist Assoc*. 1989; 84:487–493.
- Rivest LP, Wells MT. A martingale approach to the copula-graphic estimator for the survival function under dependent censoring. *J Multivar Analysis*. 2001; 79:138–155.
- Robins JM, Rotnitzky A, Bonetti M. Discussion of the Frangakis and Rubin article, “Addressing an idiosyncrasy in estimating survival curves using double sampling in the presence of self-selected right censoring. *Biometrics*. 2001; 57:343–347. [PubMed: 11414554]
- Rosen S, Fox MP, Gill CJ. Patient retention in antiretroviral therapy programs in sub-Saharan Africa: A systematic review. *PLoS Med*. 2007; 4:e298. [PubMed: 17941716]

- Sverdrup E. Estimates and test procedures in connection with stochastic models for deaths, recoveries and transfers between different states of health. *Skand Aktua-rietiidskr.* 1965; 48:184–211.
- Tsiatis A. A nonidentifiability aspect of the problem of competing risks. *Proc Nat Acad Sci USA.* 1975; 72:20–22. [PubMed: 1054494]
- Wang W. Estimating the association parameter for copula models under dependent censoring. *J R Stat Soc B.* 2003; 65:257–273.
- Xu J, Kalbfleisch JD, Tai B. Statistical analysis of illness-death processes and semicompeting risks data. *Biometrics.* 2010; 66:716–725. [PubMed: 19912171]
- Yiannoutsos CT, An MW, Frangakis CE, Musick BS, Braitstein P, et al. Sampling-based approaches to improve estimation of mortality among patient dropouts: Experience from a large PEPFAR-funded program in western Kenya. *PLoS ONE.* 2008; 3:e3843. [PubMed: 19048109]
- Yiannoutsos CT. Modeling AIDS survival after initiation of antiretroviral treatment by Weibull models with changepoints. *J Int AIDS Soc.* 2009; 12:9. [PubMed: 19558677]
- Yu M, Nan B. Regression calibration in semiparametric accelerated failure time models. *Biometrics.* 2010; 66:405–414. [PubMed: 19645700]
- Zheng M, Klein J. Estimates of marginal survival for dependent competing risks based on an assumed copula. *Biometrika.* 1995; 82:127–138.

Figure 1. Efficiency comparison for marginal survival estimation

Figure 2. Marginal distribution estimation

Figure 3. Conditional distribution estimation

Table 1
Simulation results for the estimation of α

	$\alpha = 2$	$\alpha = 3$	$\alpha = 4$
<u>Clayton; $n = 1000$</u>			
Mean Est	2.058	3.038	4.117
SE	.411	.536	.658
Emp SE	.410	.509	.660
95% CP	94.0	96.0	94.5
<u>Clayton; $n = 3000$</u>			
Mean Est	1.993	3.035	4.011
SE	.228	.298	.363
Emp SE	.237	.290	.354
95% CP	92.8	95.2	96.8
<u>Gumbel; $n = 1000$</u>			
Mean Est	2.023	2.935	4.001
SE	.237	.407	.605
Emp SE	.259	.418	.619
95% CP	91.8	92.8	93.2
<u>Gumbel; $n = 3000$</u>			
Mean Est	2.077	2.972	3.965
SE	.135	.234	.335
Emp SE	.135	.251	.328
95% CP	92.8	92.5	93.2

Simulation results for the estimation of marginal distributions at times $t = 1$ (F_X) and $t = 2$ (F_Y). The target quantities are $F_X(1) = \exp(-1) = 0.368$ and $F_Y(2) = \exp(-0.8 \times 2) = 0.202$.

Table 2

$\alpha = 2$													$\alpha = 3$				$\alpha = 4$			
		$\hat{F}_X$	$\hat{F}_Y^{\pi,1}$	$\hat{F}_Y^{\pi,2}$	$\hat{F}_Y^{\pi,M}$	$\hat{F}_X$	$\hat{F}_Y^{\pi,1}$	$\hat{F}_Y^{\pi,2}$	$\hat{F}_Y^{\pi,M}$	$\hat{F}_X$	$\hat{F}_Y^{\pi,1}$	$\hat{F}_Y^{\pi,2}$	$\hat{F}_Y^{\pi,M}$							
Clayton: $n = 1000$																				
Mean Est		.368	.204	.205	.203	.369	.206	.209	.205	.368	.206	.208	.206							
SE1		.02	.030	.029	.028	.020	.030	.029	.027	.018	.030	.030	.026							
Emp SE		.021	.032	.030	.029	.019	.033	.031	.029	.018	.030	.029	.026							
95% CP		95.5	93.5	93.5	93.5	96.0	91.8	92.2	91.8	95.0	95.2	94.2	96.0							
Clayton: $n = 3000$																				
Mean Est		.368	.204	.205	.204	.368	.203	.205	.203	.368	.203	.205	.203							
SE		.012	.018	.017	.016	.011	.018	.017	.016	.011	.018	.017	.016							
Emp SE		.012	.018	.018	.016	.011	.019	.017	.017	.011	.019	.017	.017							
95% CP		94.5	96.0	93.8	95.5	92.8	93.2	95.5	92.8	92.8	93.2	95.5	92.8							
Gumbel: $n = 1000$																				
Mean Est		.372	.208	.209	.208	.372	.205	.206	.206	.369	.208	.209	.207							
SE		.021	.032	.030	.031	.019	.031	.029	.029	.018	.031	.030	.028							
Emp SE		.022	.034	.032	.033	.018	.032	.032	.030	.019	.031	.031	.028							
95% CP		93.0	92.0	91.8	93.8	96.8	93.5	92.5	93.5	91.5	93.2	93.0	93.8							
Gumbel: $n = 3000$																				
Mean Est		.370	.204	.204	.204	.370	.203	.202	.202	.370	.203	.202	.202							
SE		.012	.019	.018	.018	.011	.018	.017	.017	.011	.018	.017	.017							
Emp SE		.013	.020	.019	.020	.011	.019	.017	.017	.011	.019	.017	.017							
95% CP		93.0	93.0	93.8	94.0	93.0	93.8	95.5	94.0	93.0	93.8	95.5	94.0							

Table 3

Simulation results for the estimation of the conditional distributions.

	$\alpha = 2$			$\alpha = 3$			$\alpha = 4$		
	$\hat{F}_{1.5 75}$	$\hat{F}_{1.5 75}$	$\hat{F}_{1.5 75}$	$\hat{F}_{1.5 75}$	$\hat{F}_{1.5 75}$	$\hat{F}_{1.5 75}$	$\hat{F}_{1.5 75}$	$\hat{F}_{1.5 75}$	$\hat{F}_{1.5 75}$
<u>Clayton, $n = 1000$</u>									
Mean Est	.207	.587	.127	.585	.072	.576			
SE	.070	.062	.063	.065	.051	.068			
Emp SE	.074	.066	.065	.070	.048	.069			
95% CP	92.0	92.0	94.8	90.2	95.0	92.2			
<u>Clayton, $n = 3000$</u>									
Mean Est	.207	.588	.116	.581	.064	.569			
SE	.041	.036	.035	.039	.026	.041			
Emp SE	.044	.038	.034	.038	.028	.044			
95% CP	92.0	93.0	94.2	93.5	94.0	92.2			
<u>Gumbel, $n = 1000$</u>									
Mean Est	.363	.533	.246	.545	.162	.556			
SE	.067	.060	.064	.062	.058	.065			
Emp SE	.069	.066	.067	.067	.058	.068			
95% CP	93.5	90.5	91.5	91.5	95.0	92.8			
<u>Gumbel, $n = 3000$</u>									
Mean Est	.355	.526	.233	.540	.150	.545			
SE	.040	.036	.038	.038	.032	.039			
Emp SE	.043	.039	.039	.038	.032	.040			
95% CP	93.0	94.0	94.8	95.5	95.0	94.5			

Table 4

Correct model selection percentage results based on 400 simulated data sets.

	$\alpha = 2$	$\alpha = 3$	$\alpha = 4$
<u>$n = 1000$</u>			
Clayton	96	95	81
Gumbel	93	98	98
<u>$n = 3000$</u>			
Clayton	100	99	89
Gumbel	100	100	100

Table 5

Data analysis results for the AMPATH study.

Estimand	Urban		Rural		P-value (Urban vs. Rural)
	Est	SE	Est	SE	
α	2.865	.481	1.852	.419	.029
$P(X > 1yr)$	.624	.007	.598	.013	.004
$P(Y > 1yr)$	.915	.014	.892	.024	.183
$P(Y > 1yr X = 3mo, Y > 3mo)$	.864	.018	.846	.028	.395
$P(Y > 1yr X > 3mo, Y > 3mo)$	.958	.006	.941	.010	.020
$P(Y > 1yr X = 6mo, Y > 6mo)$	.949	.014	.930	.021	.240
$P(Y > 1yr X > 6mo, Y > 6mo)$	.985	.004	.974	.008	.041