Nonparametric estimation of Spearman's rank correlation with bivariate survival data

Svetlana K. Eden[1], **Chun Li**[2], **Bryan E. Shepherd**[1]

[1]Department of Biostatistics, Vanderbilt University Medical Center, Nashville, Tennessee

[2]Department of Preventive Medicine, University of Southern California, Los Angeles, California

Abstract

We study rank-based approaches to estimate the correlation between two right-censored variables. With end-of-study censoring, it is often impossible to nonparametrically identify the complete bivariate survival distribution, and therefore it is impossible to nonparametrically compute Spearman's rank correlation. As a solution, we propose two measures that can be nonparametrically estimated. The first measure is Spearman's correlation in a restricted region. The second measure is Spearman's correlation for an altered but estimable joint distribution. We describe population parameters for these measures and illustrate how they are similar to and different from the overall Spearman's correlation. We propose consistent estimators of these measures and study their performance through simulations. We illustrate our methods with a study assessing the correlation between the time to viral failure and the time to regimen change among persons living with HIV in Latin America who start antiretroviral therapy.

Keywords

bivariate survival; HIV; nonparametric; Spearman's correlation; viral failure

1 | INTRODUCTION

In many medical studies, researchers are interested in measuring the correlation between two time-to-event variables. For example, in studies of HIV/AIDS, there is interest in studying the correlation between the time from antiretroviral therapy (ART) initiation to viral failure and the time from ART initiation to regimen change. These variables should be highly correlated, and it might be important to know if the observed correlation is weaker than expected. Bivariate survival data may also come from paired subjects. For example, researchers might be interested in assessing the correlation between the time to cardiovascular disease for a patient and that for his/her parents, or between times to events in twins.

Correspondence: Svetlana K. Eden, Department of Biostatistics, Vanderbilt University Medical Center, Nashville, TN. svetlana.k.eden.1@vumc.org.

Right censoring is a defining element of bivariate survival data: One or both of the times to event may not be observed. When looking at correlation between times to events that occur in a single subject (e.g., time to viral failure and time to regimen change), the censoring time may be the same time for both outcomes. However, if the times to events are in paired subjects (e.g., times to events in twins), then the censoring times may differ. Our interest is in both scenarios, but we do not consider the setting where an event occurring in one variable causes censoring of the other, that is, competing risks.

Several different methods have been proposed to measure and test the correlation between two right-censored time-to-event variables. Clayton (1978) introduced a bivariate hazard ratio or a *cross ratio* as a single number summary of correlation in the context of a frailty model. Oakes (1982, 1989) suggested a test for independence based on the cross ratio, showed its relationship with Kendall's tau, and extended its definition to a larger class of models. Fan et al. (2000) used a weighted average of the inverse of the cross ratio and a limited region Kendall's tau. Cuzick (1982) proposed a model of correlation and several test statistics, one of which resembles Spearman's rank correlation for censored data. Dabrowska (1986) derived generalized statistics to test the null hypothesis that the joint survival distribution is equal to the product of the marginals. Under certain assumptions, one of these statistics is related to a censored version of Spearman's correlation, and another corresponds to a log-rank test based on martingale residuals. Shih and Louis (1996) developed two additional statistics based on martingale residuals. Shih and Louis (1995) also suggested a two-stage estimation procedure to evaluate the correlation in bivariate data using copulas (Nelsen, 2007). Other approaches for measuring correlation using copulas for bivariate survival data have been considered by Carriere (2000), Romeo et al. (2006), and Schemper et al. (2013).

Spearman's rank correlation is ubiquitous in biomedical research because of its simple interpretation, robustness, and ability to capture nonlinear correlations. It is used much more frequently in practice than Kendall's tau, perhaps because it closely approximates Pearson's correlation under normality (Kruskal, 1958), and it is much easier to compute and interpret than a cross ratio. In the absence of censoring, Spearman's correlation is simply the correlation of the ranked data. However, despite related work by Cuzick (1982), Dabrowska (1986), and Oakes (1989), there is no nonparametric estimator of Spearman's correlation for bivariate survival data. Schemper et al. (2013) proposed a semiparametric iterative multiple imputation (IMI) method to estimate Spearman's correlation (denoted as ρ_{IMI} throughout this paper). Their method transforms bivariate survival data into a Gaussian dependency structure using a normal copula, multiply imputes censored observations from this induced bivariate distribution, and approximates Spearman's correlation using Pearson's correlation of the normal deviates. This approach is semiparametric because it does not require any assumptions about the marginal distributions. However, it uses a Gaussian dependency structure, which may lead to bias due to misspecification.

The goal of this paper is to propose and study nonparametric estimators of Spearman's correlation for right-censored data. Our methods use a nonparametric bivariate survival surface estimator. A challenge with estimating a bivariate survival surface, however, is that it may be nonparametrically estimable only within a certain region, for example, due to

end-of-study censoring. This motivates us to propose two correlation estimators: one that is defined only within the restricted region and another one that implicitly assigns values outside of this region as having the highest rank.

In Section 2, we express Spearman's correlation for time-to-event data and describe target parameters of interest. In Section 3, we address estimation and inference. In Section 4, we evaluate the performance of our estimators with several sets of simulations. In Section 5, we apply our methods to an HIV study by examining the correlation between times from treatment initiation to viral failure and regimen change. In Section 6, we discuss our methods and future directions. We have implemented our methods in the survSpearman R package (Eden et al., 2021).

2 | POPULATION PARAMETERS

2.1 | Notation and definitions

We are interested in estimating correlation between two time-to-event variables denoted as (T_X, T_Y) defined on $[0, \infty) \times [0, \infty)$. Variables T_X and T_Y can be observed on a single subject or on a pair of subjects. Each time to event can be censored. We denote time-to-censoring variables as (C_X, C_Y) and assume independence between (T_X, T_Y) and (C_X, C_Y), but C_X and C_Y can be dependent. If T_X and T_Y are observed on a single subject, then it is likely that $C_X = C_Y$. If $C_X = C_Y$ with probability one, then we call this *univariate* censoring, otherwise censoring is *bivariate*. In most studies, the follow-up period is bounded. We denote the maximum follow-up times, respectively, as τ_X and τ_Y, and consider these to be fixed by study design. Censoring, due to the end of study, has been referred to as type I censoring. With type I censoring, no events will be observed beyond the region $\Omega = [0, \tau_X) \times [0, \tau_Y)$, or equivalently, $C_X \leq \tau_X$ and $C_Y \leq \tau_Y$. For our presentation, we distinguish between *strict* type I censoring, where censoring occurs only at τ_X and τ_Y (i.e., $C_X = \tau_X$, $C_Y = \tau_Y$), and *generalized* type I censoring (Klein and Moeschberger, 1997), where censoring may also occur prior to τ_X and τ_Y (i.e., $C_X \leq \tau_X$, $C_Y \leq \tau_Y$). Strict type I censoring is rarely observed in practice, but will be helpful for explaining concepts; generalized type I censoring is quite common in practice, where follow-up time is bounded due to the length of the study, while subjects may start the study at different times or may drop out before the end of study. When the follow-up time is unbounded ($\tau_X = \infty$ and $\tau_Y = \infty$), we refer to the censoring as *unbounded*.

As a result of censoring, we only observe $X = \min(T_X, C_X)$, $Y = \min(T_Y, C_Y)$, and event indicators $\Delta_X = \mathbb{1}(T_X \leq C_X)$, $\Delta_Y = \mathbb{1}(T_Y \leq C_Y)$. We denote marginal and joint distribution functions of T_X and T_Y as $F_X(x) = \Pr(T_X \leq x)$, $F_Y(y) = \Pr(T_Y \leq y)$, $F(x, y) = \Pr(T_X \leq x, T_Y \leq y)$, and marginal and joint survival functions as $S_X(x) = \Pr(T_X > x)$, $S_Y(y) = \Pr(T_Y > y)$, $S(x, y) = \Pr(T_X > x, T_Y > y)$. We define $F_X(x^-) = \lim_{t \uparrow x} F_X(t)$ and $F(x^-, y) = \lim_{t \uparrow x} F(t, y)$; $F_Y(y^-)$ and $F(x, y^-)$ are defined similarly.

2.2 | Spearman's rank correlation

As shown by Liu et al. (2018), in the absence of censoring, the population parameter for Spearman's correlation between T_X and T_Y can be defined as:

$$\rho_S = \text{Cor}\left\{\frac{F_X(T_X) + F_X(T_{\bar{X}})}{2}, \frac{F_Y(T_Y) + F_Y(T_{\bar{Y}})}{2}\right\}. \tag{1}$$

When both T_X and T_Y are continuous, the above definition becomes

$$\rho_S = \text{Cor}\{F_X(T_X), F_Y(T_Y)\},$$

the grade correlation (Kruskal, 1958). Notice that if F_X and F_Y are estimated with their respective empirical distributions, then $F_X(T_X)$ and $F_Y(T_Y)$ are simply estimated as the ranks of T_X and T_Y, respectively, divided by the number of points in the sample, corresponding to the well-known Spearman's correlation estimator:

$$\hat{\rho}_S = \text{Cor}\{\text{rank}(T_{X,i}), \text{rank}(T_{Y,i})\},$$

where $(T_{X,i}, T_{Y,i})$ for $i = 1, \ldots, n$ are independent and identically distributed (iid) draws of (T_X, T_Y). Liu et al. (2018) have shown that Equation (1) can be presented as:

$$\begin{aligned}
\rho_S / c_\rho &= E_{T_X, T_Y}[\{F_X(T_X) + F_X(T_{\bar{X}}) - 1\} \\
&\quad \{F_Y(T_Y) + F_Y(T_{\bar{Y}}) - 1\}] \\
&= \int_0^\infty \int_0^\infty \{F_X(x) + F_X(x^-) - 1\} \\
&\quad \{F_Y(y) + F_Y(y^-) - 1\} F(dx, dy),
\end{aligned} \tag{2}$$

where $c_\rho = [\text{Var}\{F_X(T_X) + F_X(T_{\bar{X}}) - 1\}\text{Var}\{F_Y(T_Y) + F_Y(T_{\bar{Y}}) - 1\}]^{-1/2}$; and $c_\rho = 3$ when T_X and T_Y are continuous. The right-hand side of (2) is the covariance of probability-scale residuals (PSRs) proposed and studied by Li and Shepherd (2012) and Shepherd et al. (2016) and defined as:

$$\begin{aligned}
r_X(t_X, F_X) &= E\{\text{sign}(t_X - T_X)\} = \Pr(T_X < t_X) \\
&\quad - \Pr(T_X > t_X) = F_X(t_{\bar{X}}) + F_X(t_X) - 1,
\end{aligned}$$

where $\text{sign}(t_X - T_X)$ is -1, 0, and 1 for $t_X < T_X$, $t_X = T_X$, and $t_X > T_X$, respectively. We can rewrite definition (2) in terms of survival functions:

$$\begin{aligned}
\rho_S / c_\rho &= \int_0^\infty \int_0^\infty \{1 - S_X(x) - S_X(x^-)\} \\
&\quad \{1 - S_Y(y) - S_Y(y^-)\} S(dx, dy).
\end{aligned} \tag{3}$$

Right censoring causes serious challenges for nonparametric estimation of (3). First, nonparametric estimation of $S(x, y)$ is challenging due to nonunique solutions of the nonparametric likelihood in the presence of censoring and the fact that even consistent estimators of $S(x, y)$ may have negative mass for some x and y (Dabrowska, 1988; Pruitt, 1991). Second, nonparametric estimation of $S_X(x)$, $S_Y(y)$, and $S(x, y)$ beyond the maximum

follow-up time(s) is not possible. To overcome the latter challenge, one could focus on estimating Spearman's correlation inside Ω. Another possible approach is to focus on Spearman's correlation for an altered but estimable joint distribution. Population parameters for these two nonparametric approaches are presented in the next two subsections. Section 2.5 contains some examples.

2.3 | Spearman's rank correlation in a restricted region

Suppose a researcher is interested in a rank correlation only inside a restricted region, denoted Ω_R. This correlation can be computed as Spearman's correlation defined conditionally on Ω_R, which we denote as $\rho_{S \mid \Omega_R}$. With failure time data, others have proposed and advocated the use of estimators in restricted regions including restricted mean survival times (Royston and Parmar, 2013) and limited region Kendall's tau (Fan et al., 2000).

A natural choice is $\Omega_R = \Omega = [0, \tau_X) \times [0, \tau_Y)$, to estimate the restricted rank correlation over the region for which estimation is possible; to avoid introducing new notation, in this section we will use $\Omega_R = \Omega$. However, investigators can vary Ω_R depending on their research question as long as $\Omega_R \subseteq \Omega$; presumably Ω_R would generally be a rectangle that includes the origin (0,0). The probability of double failure happening in this rectangle is

$$
\begin{aligned}
P_R &= \Pr(x < \tau_X, y < \tau_Y) = F(\tau_{\bar{X}}, \tau_{\bar{Y}}) \\
&= 1 - S_X(\tau_{\bar{X}}) - S_Y(\tau_{\bar{Y}}) + S(\tau_{\bar{X}}, \tau_{\bar{Y}}).
\end{aligned} \tag{4}
$$

We consider the conditional distribution over Ω_R. An example of a conditional distribution is illustrated in the middle panel of Figure 1. Its probability mass function is

$$
S(dx, dy \mid \Omega_R) = S(dx, dy) / P_R, \tag{5}
$$

and its marginal survival function on the X-axis is

$$
S_X(x \mid \Omega_R) = 1 - F_X(x \mid \Omega_R) = \begin{cases} 0 & x \geq \tau_X \\ 1 - \dfrac{F(x, \tau_{\bar{Y}})}{P_R} & x < \tau_X \end{cases}, \tag{6}
$$

where $F(x, \tau_{\bar{Y}}) = 1 - S_X(x) - S_Y(\tau_{\bar{Y}}) + S(x, \tau_{\bar{Y}})$. The marginal survival function on the Y-axis, $S_Y(y \mid \Omega_R)$, is similarly defined. Spearman's correlation in the restricted region satisfies

$$
\begin{aligned}
\rho_{S \mid \Omega_R} / c_{\rho \mid \Omega_R} = \iint\limits_{\Omega_R} &\left\{ 1 - S_X(x \mid \Omega_R) - S_X(x^- \mid \Omega_R) \right\} \\
&\left\{ 1 - S_Y(y \mid \Omega_R) - S_Y(y^- \mid \Omega_R) \right\} S(dx, dy \mid \Omega_R),
\end{aligned} \tag{7}
$$

where $c_{\rho \mid \Omega_R} = \left[\mathrm{Var}\{ 1 - S_X(x \mid \Omega_R) - S_X(x^- \mid \Omega_R) \} \, \mathrm{Var}\{ 1 - S_Y(y \mid \Omega_R) - S_Y(y^- \mid \Omega_R) \} \right]^{-1/2}$. Note that the population parameter, $\rho_{S \mid \Omega_R}$, depends only on τ_X and τ_Y and is invariant to the censoring distribution within Ω_R.

Note that $\rho_{S\,|\,\Omega_R}$ and ρ_S are different population parameters. Therefore, it is not appropriate to use $\rho_{S\,|\,\Omega_R}$ to estimate ρ_S; rather, $\rho_{S\,|\,\Omega_R}$ can be of interest on its own. For example, Fan et al. (2000) suggested a limited region Kendall's tau to estimate over time the correlation of time to appendicitis in twins; $\rho_{S\,|\,\Omega_R}$ can be used instead of the limited region Kendall's tau for similar purposes.

2.4 | Spearman's rank correlation with highest ranks

Suppose now that a researcher is interested in the overall rank correlation, but the observations are only available within a restricted region. This is a typical situation in studies with a bounded follow-up time or when measurements have an upper detection limit. In this case, $S_X(x)$, $S_Y(y)$, and $S(x, y)$ are only nonparametrically estimable inside the region $\Omega = [0, \tau_X) \times [0, \tau_Y)$, where $\tau_X < \infty$ and $\tau_Y < \infty$. Since we are interested in a rank correlation, one approach would be to define any observation censored at τ_X as receiving the highest rank value for T_X and any observation censored at τ_Y as receiving the highest rank value for T_Y, which is the same as setting $T_X = \min(T_X, \tau_X)$ and $T_Y = \min(T_Y, \tau_Y)$. Such an approach is sensible because these censored observations at τ_X and τ_Y do have the highest rank values and there is no information to distinguish between these highest rank values without parametric modeling assumptions. Mathematically, this approach replaces $S(dx, dy)$ with a probability mass function $S^H(dx, dy)$, which is, $S(dx, dy)$ inside Ω and the left-over probability mass outside of Ω. That is,

$$
S^H(dx, dy) = \begin{cases}
S(dx, dy) & x < \tau_X \text{ and } y < \tau_Y, \\
S(\tau_X^-, dy) & x = \tau_X \text{ and } y < \tau_Y, \\
S(dx, \tau_Y^-) & x < \tau_X \text{ and } y = \tau_Y, \\
S(\tau_X^-, \tau_Y^-) & x = \tau_X \text{ and } y = \tau_Y, \\
0 & x > \tau_X \text{ or } y > \tau_Y.
\end{cases}
\tag{8}
$$

The new probability mass function $S^H(dx, dy)$ is depicted in the right panel of Figure 1. The part of $S^H(dx, dy)$ inside Ω is the same as $S(dx, dy)$ but its part outside of Ω is concentrated on the borders of Ω and at point (τ_X, τ_Y). The corresponding population parameter for the rank correlation of this new distribution is ρ_S^H, which satisfies:

$$
\rho_S^H/c_\rho^H = \int_0^{\tau_X} \int_0^{\tau_Y} \left\{1 - S_X^H(x) - S_X^H(x^-)\right\} \\
\left\{1 - S_Y^H(y) - S_Y^H(y^-)\right\} S^H(dx, dy),
\tag{9}
$$

where $S_X^H(x)$ and $S_Y^H(y)$ are the marginal survival functions of $S^H(x, y)$, and

$$
c_\rho^H = \left[\operatorname{Var}\left\{1 - S_X^H(T_X) - S_X^H(T_X^-)\right\} \\
\operatorname{Var}\left\{1 - S_Y^H(T_Y) - S_Y^H(T_Y^-)\right\}\right]^{-1/2}.
\tag{10}
$$

In other words, ρ_S^H is Spearman's correlation computed by setting $T_X = \min(T_X, \tau_X)$ and $T_Y = \min(T_Y, \tau_Y)$. Note that the population parameter, ρ_S^H, depends only on τ_X and τ_Y and is invariant to the censoring distribution within Ω.

For practical applications and intuitively, ρ_S^H can be viewed as the rank correlation computed for data with an upper detection limit, where all values above the detection limit are set to a common largest value. Note that although in general, $\rho_S^H \neq \rho_S$ (see examples in Section 2.5), unlike $\rho_{S \mid \Omega_R}$, parameter ρ_S^H is designed to take into account all observations including those outside of Ω. When the majority of the probability mass is within Ω, ρ_S^H can be viewed as an approximation of ρ_S. When $\tau_X = \infty$ and $\tau_Y = \infty$, then $\rho_S^H = \rho_S$.

2.5 | A few examples

Here, we illustrate how the restricted region can affect ρ_S^H and $\rho_{S \mid \Omega_R}$. Having a restricted region implies type I censoring. Although the figures in this section are generated with strict type I censoring, ρ_S^H remains the same for generalized type I censoring, as does $\rho_{S \mid \Omega_R}$.

In some settings, ρ_S, $\rho_{S \mid \Omega_R}$, and ρ_S^H will be quite similar. For example, the values of these parameters for the distribution shown in Figure 1 are 0.635, 0.549, and 0.634, respectively. However, these parameters may be very different in some settings.

Figure 2 shows $\rho_{S \mid \Omega_R}$ (left panel) and ρ_S^H (right panel) as a function of ρ_S for different τ_X and τ_Y for Frank's copula family. In this example, ρ_S and ρ_S^H are very similar. In contrast, ρ_S and $\rho_{S \mid \Omega_R}$ are very different, especially when ρ_S is negative and Ω_R is small (e.g., region defined by 0 to the 0.5 quantiles). This is because in these settings, only a small fraction of the underlying distribution is inside Ω_R, and therefore $\rho_{S \mid \Omega_R}$ shows a weak negative correlation. Figure 3 contains three additional examples; some of these distributions may not be realistic in practice, but are useful for illustrative purposes. The left panel shows an X-like distribution, for which $\rho_S = 0$, $\rho_{S \mid \Omega_R} = 0.69$, and $\rho_S^H = 0.06$. Here, ρ_S^H is similar to ρ_S because ρ_S^H incorporates the probability mass in the upper left, upper right, and lower right regions. However, $\rho_{S \mid \Omega_R}$ is very different from ρ_S because Ω_R only contains a positively correlated subset of the distribution. The middle panel of Figure 3 shows a distribution with highly correlated values in Ω_R, zero correlation in the upper right region, and no mass in the upper left and lower right regions. For this distribution, $\rho_S = 0.66$, $\rho_{S \mid \Omega_R} = 0.64$, and $\rho_S^H = 0.99$. Here, $\rho_{S \mid \Omega_R}$ and ρ_S are similar because only the points in the restricted region are correlated. In contrast, ρ_S^H is quite high because the large probability mass of the upper right region is concentrated on a single highest-rank point when computing ρ_S^H, which pulls its value upwards considerably. The right panel of Figure 3 shows a distribution with a highly negative overall correlation, for which $\rho_S = -0.90$,

$\rho_{S \mid \Omega_R} = -0.39$, and $\rho_S^H = -0.74$. Here, ρ_S^H and ρ_S are fairly different because although the probability mass of censored observations outside of Ω_R are taken into account, there is some loss of information with the highest rank assignment. The parameter $\rho_{S \mid \Omega_R}$ is very different from ρ_S because only a small fraction of the negatively correlated mass is included in Ω_R. Note that parametric or semiparametric approaches also struggle with many of these settings because they effectively use the information from the restricted region to impute what is occurring outside the restricted region. For example, $\hat{\rho}_{IMI}$ (see Schemper et al., 2013) for the middle panel is approximately 0.95.

3 | NONPARAMETRIC ESTIMATION

3.1 | Estimation of $\rho_{S \mid \Omega_R}$ under generalized type I censoring

We estimate $\rho_{S \mid \Omega_R}$ using a plug-in estimator for Equation (7):

$$\hat{\rho}_{S \mid \Omega_R}/\hat{c}_{\rho \mid \Omega_R} = \sum_{i:x_i < \tau_X} \sum_{j:y_j < \tau_Y} \left[\left\{ 1 - \hat{S}_X(x_i \mid \Omega_R) - \hat{S}_X(x_i^- \mid \Omega_R) \right\} \right.$$
$$\left. \times \left\{ 1 - \hat{S}_Y(y_j \mid \Omega_R) - \hat{S}_Y(y_j^- \mid \Omega_R) \right\} \hat{S}(dx_i, dy_j \mid \Omega_R) \right],$$

where

$$\hat{c}_{\rho \mid \Omega_R} = \left(\widehat{\mathrm{Var}}_X \cdot \widehat{\mathrm{Var}}_Y \right)^{-1/2},$$
$$\widehat{\mathrm{Var}}_X = \sum_{i:x_i < \tau_X} \left(\left[\left\{ 1 - \hat{S}_X(x_i \mid \Omega_R) - \hat{S}_X(x_i^- \mid \Omega_R) \right\} \right. \right.$$
$$\left. \left. - \sum_{i:x_i < \tau_X} \left\{ 1 - \hat{S}_X(x_i \mid \Omega_R) - \hat{S}_X(x_i^- \mid \Omega_R) \right\} \hat{S}_X(dx_i \mid \Omega_R) \right]^2 \times \hat{S}_X(dx_i \mid \Omega_R) \right),$$

$\qquad$ (11)

and $\widehat{\mathrm{Var}}_Y$ is computed similarly. The conditional survival curves, $\hat{S}_X(x \mid \Omega_R)$, $\hat{S}_Y(y \mid \Omega_R)$, and $\hat{S}(x, y \mid \Omega_R)$ are estimated using plug-in estimators for (4), (5), and (6).

For $\hat{S}(dx, dy)$ we use the estimator of Dabrowska (1988). The marginal distributions of Dabrowska's estimator, $\hat{S}(x)$ and $\hat{S}(y)$, are Kaplan–Meier estimators. There are other choices for nonparametrically estimating $S(dx, dy)$, including the estimators proposed by Prentice and Cai (1992), van der Laan (1997), Campbell (1981), and Lin and Ying (1993), to name a few. We considered the estimators of Campbell (1981) and Lin and Ying (1993) because of their computational simplicity, but ultimately chose Dabrowska's estimator because it is consistent for $S(dx, dy)$, straightforward to compute, and tended to result in estimates of Spearman's correlation with better performance (see Section 4). The confidence interval (CI) of $\hat{\rho}_{S \mid \Omega_R}$ is estimated using the bootstrap. Because of the consistency and asymptotic normality of Dabrowska's estimator, $\hat{\rho}_{S \mid \Omega_R}$ is consistent for $\rho_{S \mid \Omega_R}$ and asymptotically normal when $\rho_{S \mid \Omega_R} \in (-1, 1)$ (see Theorem A1 in the Appendix).

A few problems can arise in practice when computing $\hat{\rho}_{S \mid \Omega_R}$, all due to negative mass at some points in $\hat{S}(dx, dy)$. First, in some extreme cases (e.g., small sample sizes, strong positive or negative correlation, and heavy censoring), $|\hat{\rho}_{S \mid \Omega_R}|$ may exceed 1; if this happens, our software sets it to $\text{Sign}(\hat{\rho}_{S \mid \Omega_R})$ and generates a warning. Second, negative mass can also lead to problems computing $\hat{c}_{\rho \mid \Omega_R}$ in (11) because $\widehat{\text{Var}}_X$ or $\widehat{\text{Var}}_Y$ may be negative because $\hat{S}_X(dx \mid \Omega_R)$ or $\hat{S}_Y(dy \mid \Omega_R)$ is negative at some points. If this happens, we correct the guilty conditional marginal probability mass estimator by assigning negative values to 0, and by normalizing the rest of the probability mass values; specifically, the corrected probability mass is $\hat{S}_X^*(dx \mid \Omega_R) = \max\{0, \hat{S}_X(dx \mid \Omega_R)\} / \sum_x \max\{0, \hat{S}_X(dx \mid \Omega_R)\}$. Third, negative mass may lead to a negative estimate of P_R, the probability of both events occurring in Ω_R (see Equation (4)); when this occurs, $\hat{\rho}_{S \mid \Omega_R}$ is not defined.

Although the number of points with negative mass does not decrease as the sample size increases (Pruitt, 1991), the amount of negative mass at each point does go to zero, which therefore reduces the possibility of these problems occurring. Also, the tendency of having negative mass is lower when a lower proportion of observations are singly or doubly censored. To give a sense of the magnitude of these problems, for a sample size of 50 with 70% bivariate censoring, $\widehat{\text{Var}}_X$ or $\widehat{\text{Var}}_Y$ was negative for 1.3% of 1000 simulations, and $\hat{P}_R$ was less than zero in 2.4% of simulations. With a sample size of 100 and 70% bivariate censoring, these problems occurred in 0% and 0.8% of simulations, respectively.

3.2 | Estimation of ρ_S^H under generalized type I censoring

Equation (9) provides a straightforward way of estimating ρ_S^H, given a nonparametric estimate of the bivariate survival surface, $\hat{S}(dx, dy)$:

$$\hat{\rho}_S^H / \hat{c}_\rho^H = \sum_{i*} \sum_{j*} \left\{ 1 - \hat{S}_X^H(x_{i*}) - \hat{S}_X^H(x_{\bar{i}*}) \right\} \\ \left\{ 1 - \hat{S}_Y^H(y_{j*}) - \hat{S}_Y^H(y_{\bar{j}*}) \right\} \hat{S}^H(dx_{i*}, dy_{j*}),$$

where $i*$ enumerates all the events for X plus τ_X, $j*$ enumerates all the event for Y plus τ_Y, and $\hat{S}^H(dx_i, dy_j)$ and $\hat{c}_\rho^H$ are the plug-in estimators for (8) and (10), respectively. As before, we compute $\hat{S}(x, y)$ using Dabrowska's estimator. The CI of $\hat{\rho}_S^H$ is estimated using the bootstrap. Again, $\hat{\rho}_S^H$ is consistent for ρ_S^H and asymptotically normal when $\rho_S^H \in (-1, 1)$ (see Theorem A2 in the Appendix). In practice, for some extreme cases similar to those mentioned in Section 3.1 for $\hat{\rho}_{S \mid \Omega_R}$, $\left| \hat{\rho}_S^H \right|$ may exceed 1; when this occurs, we set it to $\text{Sign}(\hat{\rho}_S^H)$, and our software generates a warning.

3.3 | Estimation of ρ_S under unbounded censoring

The estimator $\hat{\rho}_S^H$ may also be used to estimate ρ_S with unbounded censoring. Under unbounded censoring, $\tau_X = \tau_Y = \infty$ by definition, and one would naturally estimate $\hat{\rho}_S$ by plugging $\hat{S}(x, y)$, $\hat{S}_X(x)$, and $\hat{S}_Y(y)$ into (3) in a manner similar to that described above. However, if the maximum value of X, for example, is a censored event (i.e., $\Delta_X = 0$), then $\hat{S}(dx, dy)$ will not sum to 1 resulting in improper integration when using plug-in estimators in (3). A workaround is to assign the remaining mass (which is typically very little) to a point just beyond the largest observed event time of X and set τ_X to this point. We then estimate ρ_S with $\hat{\rho}_S^H$. Although in this setting, τ_X is no longer fixed but unbounded, estimation of ρ_S in this manner performs well (see Section 4). Under unbounded censoring, $\hat{\rho}_S^H$ is consistent for ρ_S and asymptotically normal when $\rho_S \in (-1, 1)$ (see Theorem A3 in the Appendix).

4 | SIMULATIONS

4.1 | Simulation set-up

We performed several simulations to investigate the finite sample performance of our estimators. The random variables T_X and T_Y were simulated using various choices of copula families and parameters. Specifically, following Fan et al. (2000), we simulated dependent random uniform variables, U and V, from Clayton's and Frank's copula families; both copulas are defined by a single parameter, θ. The dependence between U and V was specified by choosing the parameter θ in such a way that the true Spearman's correlation varied among no correlation ($\rho_S = 0$), moderate correlation ($\rho_S = -0.2$ and 0.2 for Frank's family and 0.2 for Clayton's family), and strong correlation ($\rho_S = -0.6$ and 0.6 for Frank's family and 0.6 for Clayton's family). (Clayton's family does not permit negative correlation.) We then set $T_X = -\log(1 - U)$ and $T_Y = -\log(1 - V)$ such that T_X and T_Y were exponentially distributed with mean 1.

Four types of censoring scenarios were implemented: (1) unbounded univariate, (2) unbounded bivariate, (3) generalized type I univariate, and (4) generalized type I bivariate. Each censoring scenario was implemented for censoring proportions $P_C = \{0.3, 0.7\}$. Bivariate unbounded censoring times C_X and C_Y for each observation were simulated independently from an exponential distribution, $Pr(C_X \leq t) = Pr(C_Y \leq t) = 1 - e^{-\lambda t}$, with $\lambda = P_C/(1 - P_C)$. The event times T_X and T_Y were censored if $T_X > C_X$ and $T_Y > C_Y$, respectively. Univariate unbounded censoring was implemented in a similar manner except only one censoring event was generated per (T_X, T_Y) pair. For generalized type I censoring, C_X^* and C_Y^*, were first simulated as described above with probability P_C and then C_X and C_Y were defined as (C_X^*, τ_X) and (C_Y^*, τ_Y), respectively, with $\tau_X = \tau_Y$ set at the median survival time, $S_X^{-1}(0.5) = S_Y^{-1}(0.5)$. The resulting censoring proportions for generalized type I censoring were therefore higher than P_C: for example, for generalized type I bivariate censoring with $P_C = 0.3$ and 0.7, the outcomes were censored for approximately 56% and 73% of observations, respectively.

We evaluated the performance of $\hat{\rho}_S^H$ and $\hat{\rho}_{S \mid \Omega_R}$ in the presence of unbounded and generalized type I censoring using a sample size of 200 for strong, moderate, and no correlation. For unbounded censoring, the population parameter of ρ_S^H is the same as ρ_S. For generalized type I censoring, the population parameters of ρ_S^H and $\rho_{S \mid \Omega_R}$ were the same as ρ_S for Clayton's family. For Frank's family with the overall Spearman's correlation of −0.6, −0.2, 0.2, and 0.6, the population parameters of $\rho_{S \mid \Omega_R}$ were −0.098, −0.042, 0.058, and 0.261; and the population parameters of ρ_S^H were −0.512, −0.173, 0.180, and 0.545. These population parameters were empirically estimated with a sample size of 10^6.

The bias, root mean squared error (RMSE), type I error rate, and power, computed as the proportion of times that bootstrap CIs (based on 1000 bootstrap samples) did not include zero, were also evaluated for sample sizes of 100 and 200 under unbounded censoring for moderate and no correlation. The performance of $\hat{\rho}_S^H$ was compared to estimator $\hat{\rho}_{IMI}$ proposed by Schemper et al. (2013) for these simulation scenarios. We also evaluated the performance of $\hat{\rho}_S^H$ as an estimator of ρ_S using survival surfaces proposed by Lin and Ying (1993) and Campbell (1981).

Lastly, we did some additional comparisons of $\hat{\rho}_S^H$ to semiparametric estimators $\hat{\rho}_{IMI}$ and $\hat{\rho}_S^{MLE}$ (maximum likelihood estimator assuming Frank's copula dependency structure). These comparisons were made in the context of a well-behaved dependency structure induced by Frank's copula and in the context of a complex dependency structure, a mixture of 60% highly negatively correlated data ($\rho_S = -0.8$, Frank's copula with $\theta = -8$) and 40% perfectly correlated data ($\rho_S = 1$) with the overall Spearman's correlation being about −0.0813. This simulation scenario was loosely motivated by data from site B in our real data analysis presented in Section 5; Supplementary Figure S5 shows the scatter plot of the uncensored data. Sample sizes of 200, 500, and 1000 were used and unbounded univariate censoring with $P_C = 0.5$ was applied as described above. The goal of these comparisons was to show better efficiency of $\hat{\rho}_S^{MLE}$ compared to $\hat{\rho}_S^H$ under a correctly specified model and to demonstrate greater accuracy of $\hat{\rho}_S^H$ compared to $\hat{\rho}_{IMI}$ and $\hat{\rho}_S^{MLE}$ under model misspecification.

All simulations used 1000 replications.

4.2 | Simulation results

Figure 4 shows the mean point estimates and the 0.025th and 0.975th quantiles of estimators $\hat{\rho}_S^H$ under unbounded censoring (row 1), the semiparametric estimator $\hat{\rho}_{IMI}$ proposed by Schemper et al. (2013) under unbounded censoring (row 2), $\hat{\rho}_S^H$ under generalized type I censoring (row 3), and $\hat{\rho}_{S \mid \Omega_R}$ under generalized type I censoring (row 4). The sample size was 200, the censoring was bivariate with varying censoring proportions, and $\hat{\rho}_S^H$ and $\hat{\rho}_{S \mid \Omega_R}$ were computed using Dabrowska's survival surface estimator.

For Clayton's and Frank's families under unbounded censoring (row 1), the mean of $\hat{\rho}_S^H$ was very close to the true population parameter verifying the consistency of $\hat{\rho}_S^H$ for ρ_S. When data were generated under Frank's copula, the semiparametric estimator, $\hat{\rho}_{IMI}$ (row 2), was similarly unbiased for unbounded censoring, and it tended to be less variable than $\hat{\rho}_S^H$. However, when data were generated using Clayton's copula, $\hat{\rho}_{IMI}$ was biased for ρ_S, also noted by Schemper et al. (2013).

Supplementary Tables S1 and S2 provide more details and additional comparisons between $\hat{\rho}_{IMI}$ and $\hat{\rho}_S^H$ under unbounded censoring for different sample sizes and censoring proportions in terms of bias, RMSE, type I error rate and power. In short, the bias of $\hat{\rho}_S^H$ for ρ_S was low, even with fairly small numbers of events; both the bias and RMSE decreased as the number of events increased. In general, $\hat{\rho}_S^H$ compared favorably to $\hat{\rho}_{IMI}$.

Under generalized type I censoring (the third and fourth rows of Figure 4), the means of $\hat{\rho}_S^H$ and $\hat{\rho}_{S \mid \Omega_R}$ are very close to their true population parameters, suggesting with $n = 200$, these estimators are essentially unbiased for ρ_S^H and $\rho_{S \mid \Omega_R}$ respectively. The variance of our estimators naturally increased as the probability of censoring increased. The variance of $\hat{\rho}_{S \mid \Omega_R}$ was greater than that of $\hat{\rho}_S^H$ under generalized type I censoring (rows 3 and 4), presumably because $\hat{\rho}_{S \mid \Omega_R}$ only uses the events inside $\Omega = \Omega_R$, whereas $\hat{\rho}_S^H$ assigns probability mass to values outside of Ω. Note that the variance of $\hat{\rho}_S^H$ under unbounded censoring was slightly larger than that under generalized type I censoring in spite of the lighter censoring. This is probably because under generalized type I censoring, the probability mass, $\hat{S}^H(dx, dy)$, calculated outside of Ω is concentrated on the same points, making its variance smaller compared to the case of unbounded censoring. Results for univariate censoring were very similar to those for bivariate censoring, except the estimators were slightly less variable (see Supplementary Figure S1).

With unbounded censoring of 50% and $n = 200$, the correctly specified semiparametric $\hat{\rho}_S^{MLE}$ was more efficient than $\hat{\rho}_S^H$ (relative efficiency in terms of variance ranging from 1.19 (for $\rho_S = 0$) to 1.60 (for $\rho_S = 0.6$), Supplementary Figure S4); both approaches yielded unbiased estimates of ρ_S. In contrast, when data were generated using a mixture of positively and negatively correlated bivariate distributions, the misspecified semiparametric estimators $\hat{\rho}_S^{MLE}$ and $\hat{\rho}_{IMI}$ were substantially biased and this bias did not decrease with increasing sample size. On the other hand, the nonparametric $\hat{\rho}_S^H$ was unbiased for ρ_S (see Supplementary Figure S6). In this more complicated setting, estimates of $\rho_{S \mid \Omega_R}$ with Ω_R being defined using the median survival times were also unbiased (Supplementary Figure 7).

The simulations reported above incorporated 1000 bootstrap replications; in general, CI coverage and width were stable and adequate with as few as 200 bootstrap replications (see Supplementary Figure S8).

We also evaluated the performance of $\hat{\rho}_S^H$ with survival surfaces of Campbell (1981) for univariate and bivariate censoring and of Lin and Ying (1993) for univariate censoring only (see Supplementary Figures S2 and S3). With univariate censoring, using the estimator of Lin and Ying (1993) resulted in unbiased estimation although with larger variance than that using Dabrowska's estimator. Estimator $\hat{\rho}_S^H$ computed using the survival surface estimator of Campbell (1981) was visibly biased for the sample size of 200 under heavy censoring.

5 | APPLICATION

We apply our methods to a study of 6691 HIV-positive adults starting ART in Latin America. We are interested in estimating the correlation between two right-censored variables: (1) the time from ART initiation to viral failure and (2) the time from ART initiation to major regimen change. Patients experience viral failure when the amount of virus circulating in their blood (their viral load) is above a certain threshold, which may make them infectious and vulnerable to HIV-related diseases. Viral failure may be caused by many factors including poor adherence or drug resistance; it often triggers changing a patient's ART regimen. However, patients may also change their ART regimens for reasons other than viral failure (e.g., poor tolerability, discovery of a simpler regimen, or patient/provider choice).

Our study uses data from the Caribbean, Central, and South America network for HIV Epidemiology (CCASAnet). The definitions of viral failure and regimen change were the same as those used in a prior CCASAnet study (Cesar et al., 2015; CCASAnet, 2020). In short, viral failure was defined as a single viral load > 1000 copies/mL or two viral loads > 400 copies/mL after a person's virus had been suppressed or they had been on ART long enough that it should have been suppressed (i.e., 6 months). Regimen change was limited to major changes such that the patient switched drug classes or changed multiple drugs. Each study subject may have had one, both, or neither of these events. Follow-up ended at the last clinic visit; censoring was univariate. Our analysis data set includes patients from Brazil, Chile, Honduras, Mexico, and Peru; sites have been anonymized for presentation. After a median follow-up of 4.1 years (ranging from 1 day to 18.2 years), 1916 persons (28.6%) had a viral failure and 1895 persons (28.3%) changed regimens. Approximately 16.1% of patients had both events over the follow-up period, 12.2% changed regimens but did not have viral failure, 12.5% had viral failure but did not change regimens, and 59.1% were not observed to have either event. The upper left panel of Figure 5 shows Kaplan–Meier estimates for the marginal probabilities of viral failure and regimen change as a function of time since ART initiation. Ten years after ART initiation, the estimated probability of not having viral failure was 0.58 and the estimated probability of remaining on the initial regimen was 0.51. The upper right panel shows the estimated joint bivariate probability mass function, $\hat{S}^H(dx, dy)$, based on Dabrowska's estimator; estimated marginal probability mass functions, $\hat{S}_X^H(dx)$ and $\hat{S}_Y^H(dy)$, are also included. Note that because a large proportion of patients experienced only one or neither event, a large amount of mass has been assigned to $\tau_X = 18$ and $\tau_Y = 17$ years.

The estimated highest rank correlation, $\hat{\rho}_S^H$, between time to viral failure and time to regimen change was 0.35 with a 95% CI based on 1000 bootstrap replications of 0.27 to 0.44. This result is fairly similar, albeit with a wider CI, to the estimator of Schemper et al., $\hat{\rho}_{IMI}$, that imputes censored values: 0.37, 95% CI = [0.33, 0.41].

The estimated rank correlation over the restricted region, $\hat{\rho}_{S \mid \Omega_R}$, with $\Omega_R = [0, 15) \times [0, 15)$ was substantially higher, 0.65, with a wide 95% CI = [0.30, 0.97]. This can be explained by a careful look at the conditional probability mass over Ω_R (lower left plot of Figure 5). Notice that there are several points in the upper right corner of this surface with large probability mass. Large amounts of mass are assigned to these points because a few events occurred at later follow-up times when there were fairly small numbers of patients remaining in the risk set. This is also seen in the Kaplan–Meier estimates, where relatively large drops in the marginal survival curves are noted between 10 and 15 years. These points pushed the probability mass function closer to the diagonal, leading to a larger estimated rank correlation. In addition, since there were only a few points with substantial probability mass, their inclusion/exclusion in various bootstrap samples led to wide variation in CIs. This example serves as a nice illustration of the potential perils of estimating rank correlations over restricted regions that include tail areas with small numbers of events. Perhaps a more reliable rank correlation would be over the restricted region, $\Omega_R = [0, 10) \times [0, 10)$, in which case $\hat{\rho}_{S \mid \Omega_R}$ was 0.26 (95% CI = [0.17, 0.36]); the lower right plot of Figure 5 shows the conditional probability mass over this smaller region.

In addition to showing estimates of overall correlation, Table 1 shows estimates based on sex and study site. For the most part, $\hat{\rho}_S^H$ is fairly close to $\hat{\rho}_{IMI}$, except for those sites with small sample sizes (i.e., A, B, and D). Rank correlations over the restricted region, $\hat{\rho}_{S \mid \Omega_R}$ with $\Omega_R = [0, 15) \times [0, 15)$ were generally more variable and typically higher than those over $\Omega_R = [0, 10) \times [0, 10)$.

6 | DISCUSSION

We have proposed two nonparametric methods of quantifying correlation with bivariate right-censored data. One estimator, $\hat{\rho}_{S \mid \Omega_R}$ computes Spearman's correlation within a restricted region. The other estimator, $\hat{\rho}_S^H$, computes Spearman's correlation for an estimable bivariate distribution, which is analogous to assigning data censored beyond the estimable region to the highest rank values. Under unbounded censoring, $\hat{\rho}_S^H$ is consistent for the overall Spearman's correlation. Under generalized type I censoring, with the majority of events happening in the restricted region, $\hat{\rho}_S^H$ can be viewed as an approximation of the overall Spearman's correlation. Because our methods assume neither marginal nor joint parametric distributions, they have potential advantages over parametric and semiparametric methods.

The main limitations of our estimators stem from challenges with nonparametrically estimating the bivariate survival surface. The first challenge is that with generalized type

I censoring, the bivariate survival surface, and hence Spearman's rank correlation ρ_S, cannot be identified beyond the region of data support without parametric assumptions. Hence, our inference targets were the estimable parameters, $\rho_{S \mid \Omega_R}$ and ρ_S^H. Although under generalized type I censoring they do not equal the overall Spearman's correlation, both parameters have sensible interpretations. The alternative, using parametric and semiparametric models to estimate Spearman's correlation—an inherently nonparametric statistic— has other limitations. Parametric and semiparametric approaches assume a dependency structure; and as seen in our simulations, semiparametric estimators such as $\hat{\rho}_{IMI}$ proposed by Schemper et al. (2013), may be biased when the dependency structure is misspecified. In addition, they implicitly assume that the dependency structure outside the region of observation is the same as that seen inside the observation region. Although in the real data example, $\hat{\rho}_{IMI}$ appeared more stable than our nonparametric estimators, particularly for small sample sizes, with complex real data there is a real possibility of model misspecification. As with most statistics, there may be settings where one might prefer the nonparametric estimators over the semiparametric estimators, or vice versa.

The second challenge is that even in regions where the bivariate survival curve is identifiable, nonparametric estimators of the survival surface may have negative mass, which leads to potential downstream problems with estimation of $\rho_{S \mid \Omega_R}$ and ρ_S^H. Researchers have grappled with nonparametric approaches to avoid negative mass, for example, van der Laan (1996). Thankfully, in our simulations, only a minor proportion of our estimates encountered problems due to negative mass, the problems go away as the numbers of events increase, and there are typically workaround solutions that appear to behave reasonably.

In our approach we considered a rectangular restricted region, $[0, \tau_X] \times [0, \tau_Y]$. An anonymous associate editor correctly pointed out that the top right corner of this rectangle is not always identifiable nonparametrically. Alternatively, one could consider defining estimators using the identifiable region $\Omega_R = \{(t_X, t_Y) : T_{X,i} \quad t_X, T_{Y,i} \quad t_Y$ for some subject $i\}$ (see Prentice and Zhao, 2019). Although this idea is intriguing, we decided to keep the definition of Ω_R as $[0, \tau_X] \times [0, \tau_Y]$ due to easier interpretation. Note that Dabrowska's estimator can be computed and is consistent for all points inside $[0, \tau_X] \times [0, \tau_Y]$.

Another limitation of our method is its computational time complexity, up to $\mathcal{O}(mn^2)$, where n is the sample size and m is the number of bootstrap samples. For example, on a MacBook Air with 1.7 GHz Intel Core i7, the computational time for $n = 100, 200, 500, 1000$ and $m = 200$ is approximately 0.2, 0.7, 5.1, and 30.8 min, respectively.

Although not studied here, our approaches can be directly applied to settings where only one of the variables is right censored. This special case allows nonparametric estimation of the bivariate survival surface without negative mass (Stute, 1993). In future work, we plan to study semiparametric methods for estimating covariate-adjusted partial and conditional Spearman's correlation for bivariate survival data.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

DATA AVAILABILITY STATEMENT

The deidentified data used in this article are openly available at https://biostat.app.vumc.org/ArchivedAnalyses.

APPENDIX

Theorem A1.

Under uninformative and generalized type I censoring and for a nonempty rectangular region Ω_R such that $\Omega_R \subseteq \Omega$, $\hat{\rho}_{S \mid \Omega_R}$ is a consistent estimator of $\rho_{S \mid \Omega_R}$. When $\rho_{S \mid \Omega_R} \in (-1, 1)$, estimator $\hat{\rho}_{S \mid \Omega_R}$ is also asymptotically normal, and the bootstrap estimation of its variance is asymptotically valid.

Proof.

The consistency of $\hat{\rho}_{S \mid \Omega_R}$ follows from the consistency of Dabrowska's estimator $\hat{S}(x, y)$ in Ω under uninformative and generalized type I censoring and the fact that $\hat{S}(x, y \mid \Omega_R)$ is a continuous function of $\hat{S}(x, y)$. To prove asymptotic normality, we note that under uninformative and generalized type I censoring, $\hat{S}(x, y)$, $\hat{S}_X(x)$, and $\hat{S}_Y(y)$ are Hadamard differentiable estimators that converge to Gaussian processes; see Dabrowska (1989) and van der Vaart and Wellner (1996). Because estimator $\hat{\rho}_{S \mid \Omega_R}$ is a continuous function of $\hat{S}(x, y)$, $\hat{S}_X(x)$, and $\hat{S}_Y(y)$, it follows from the chain rule that it is also Hadamard differentiable and, therefore, it is asymptotically normal by the functional delta method (van der Vaart and Wellner, 1996). In addition, its bootstrap estimator converges to the same Gaussian process as $\hat{\rho}_{S \mid \Omega_R}$, justifying the use of the bootstrap to construct CIs (van der Vaart and Wellner, 1996). □

Theorem A2.

Under uninformative and generalized type I censoring, estimator $\hat{\rho}_S^H$ is consistent. When, $\rho_S^H \in (-1, 1)$ it is also asymptotically normal, and the bootstrap estimation of its variance is asymptotically valid.

Proof.

Similar to Theorem A1. □

Theorem A3.

Under uninformative and unbounded censoring, $\hat{\rho}_S^H$ is a consistent estimator of the overall Spearman's correlation, ρ_S. Estimator $\hat{\rho}_S^H$ is also asymptotically normal when $\rho_S \in (-1, 1)$, and the bootstrap estimation of its variance is asymptotically valid.

Proof.

When the maximum of X and Y are observed events, then $\hat{S}^H(x, y) = \hat{S}(x, y)$, so the consistency of $\hat{\rho}_S^H$ for ρ_S follows from the consistency of Dabrowska's estimator and the continuous mapping theorem. When the maximum of either X or Y is censored, the plug-in estimator of (9), $\hat{\rho}_S^H$, can be defined as a sum of four terms A, B, C, and D that are functions of the joint and conditional survival surface estimators (Theorem A.1, Supplementary Appendix). Because of Lemma A.1 (Supplementary Appendix), B, C, $D \xrightarrow{p} 0$ and $A \xrightarrow{p} \rho_S$. The proofs of the asymptotic normality and bootstrap validity are similar to Theorem A1. $\square$

REFERENCES

Campbell G (1981) Nonparametric bivariate estimation with randomly censored data. Biometrika, 68, 417–422.

Carriere JF (2000) Bivariate survival models for coupled lives. Scandinavian Actuarial Journal, 2000, 17–32.

CCASAnet (2020). Times to viral failure and regimen change data. Anonymized for presentation, https://biostat.app.vumc.org/ArchivedAnalyses accessed March 17, 2021.

Cesar C, Jenkins CA, Shepherd BE, Padgett D, Mejía F, Ribeiro SR, et al. (2015) Incidence of virological failure and major regimen change of initial combination antiretroviral therapy in the Latin America and the Caribbean: an observational cohort study. The Lancet HIV, 2, e492–e500. [PubMed: 26520929]

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

Cuzick J (1982) Rank tests for association with right censored data. Biometrika, 69, 351–364.

Dabrowska DM (1986) Rank tests for independence for bivariate censored data. The Annals of Statistics, 14, 250–264.

Dabrowska DM (1988) Kaplan–Meier estimate on the plane. The Annals of Statistics, 16, 1475–1489.

Dabrowska DM (1989) Kaplan–Meier estimate on the plane: weak convergence, LIL, and the bootstrap. Journal of Multivariate Analysis, 29, 308–325.

Eden SK, Li C and Shepherd BE (2021). survSpearman: Nonparametric Spearman's Correlation for Survival Data. R package version 1.0.0, https://CRAN.R-project.org/package=survSpearman accessed March 17, 2021.

Fan J, Hsu L and Prentice RL (2000) Dependence estimation over a finite bivariate failure time region. Lifetime Data Analysis, 6, 343–355. [PubMed: 11190604]

Klein JP and Moeschberger ML (1997). Survival Analysis: Techniques for Censored and Truncated Data. New York: Springer Science & Business Media.

Kruskal WH (1958) Ordinal measures of association. Journal of the American Statistical Association, 53, 814–861.

Li C and Shepherd BE (2012) A new residual for ordinal outcomes. Biometrika, 99, 473–480. [PubMed: 23843667]

Lin D and Ying Z (1993) A simple nonparametric estimator of the bivariate survival function under univariate censoring. Biometrika, 80, 573–581.

Liu Q, Li C, Wanga V and Shepherd BE (2018) Covariate-adjusted Spearman's rank correlation with probability-scale residuals. Biometrics, 74, 595–605. [PubMed: 29131931]

Nelsen RB (2007). An Introduction to Copulas. New York: Springer Science & Business Media.

Oakes D (1982) A model for association in bivariate survival data. Journal of the Royal Statistical Society: Series B (Methodological), 44, 414–422.

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

Prentice RL and Cai J (1992) Covariance and survivor function estimation using censored multivariate failure time data. Biometrika, 79, 495–512.

Prentice RL and Zhao S (2019). The Statistical Analysis of Multivariate Failure Time Data: A Marginal Modeling Approach. Boca Raton, FL: Chapman and Hall/CRC.

Pruitt RC (1991) On negative mass assigned by the bivariate Kaplan–Meier estimator. The Annals of Statistics, 19, 443–453.

R Core Team (2017). R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. [Version 3.4.3., https://www.R-project.org/ (accessed July 9, 2019)]

Romeo JS, Tanaka NI and Pedroso-de Lima AC (2006) Bivariate survival modeling: a Bayesian approach based on copulas. Lifetime Data Analysis, 12, 205–222. [PubMed: 16868839]

Royston P and Parmar MK (2013) Restricted mean survival time: an alternative to the hazard ratio for the design and analysis of randomized trials with a time-to-event outcome. BMC Medical Research Methodology, 13, 152. [PubMed: 24314264]

Schemper M, Kaider A, Wakounig S and Heinze G (2013) Estimating the correlation of bivariate failure times under censoring. Statistics in Medicine, 32, 4781–4790. [PubMed: 23775542]

Shepherd BE, Li C and Liu Q (2016) Probability-scale residuals for continuous, discrete, and censored data. Canadian Journal of Statistics, 44, 463–479.

Shih JH and Louis TA (1995) Inferences on the association parameter in copula models for bivariate survival data. Biometrics, 51, 1384–1399. [PubMed: 8589230]

Shih JH and Louis TA (1996) Tests of independence for bivariate survival data. Biometrics, 52, 1440–1449. [PubMed: 8962462]

Stute W (1993) Consistent estimation under random censorship when covariables are present. Journal of Multivariate Analysis, 45, 89–103.

van der Laan M (1997) Nonparametric estimators of the bivariate survival function under random censoring. Statistica Neerlandica, 51, 178–200.

van der Laan MJ (1996) Efficient estimation in the bivariate censoring model and repairing NPMLE. The Annals of Statistics, 24, 596–627.

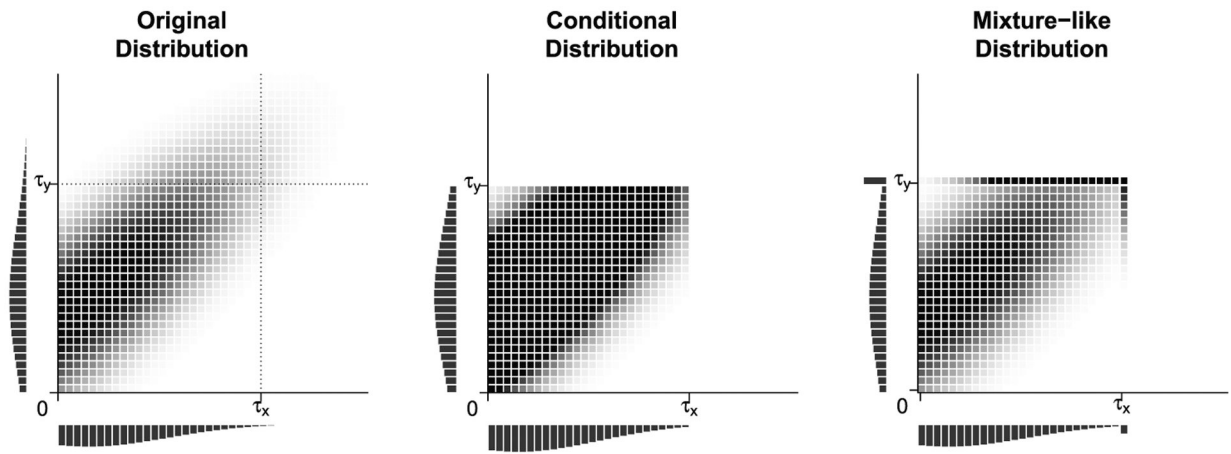van der Vaart AW and Wellner JA (1996). Weak Convergence and Empirical Processes. Springer.

FIGURE 1.

Illustration of bivariate distributions underlying the three population parameters. Left: Original distribution over $[0, \infty) \times [0, \infty)$, which has Spearman's correlation ρ_S. Middle: Conditional distribution over $\Omega_R = \Omega$, which has Spearman's correlation $\rho_{S \mid \Omega_R}$. Right: Mixture-like distribution S^H over region $\Omega \cup [0, \tau_X] \times \tau_Y \cup \tau_X \times [0, \tau_Y]$, which has Spearman's correlation ρ_S^H
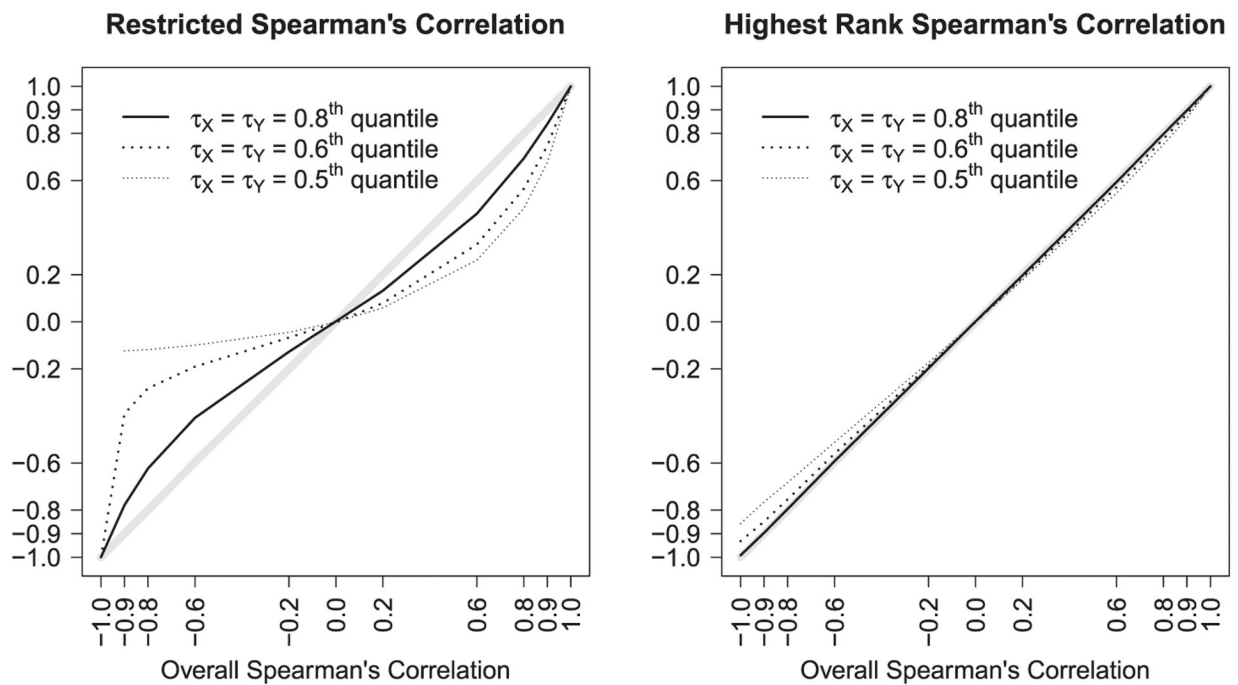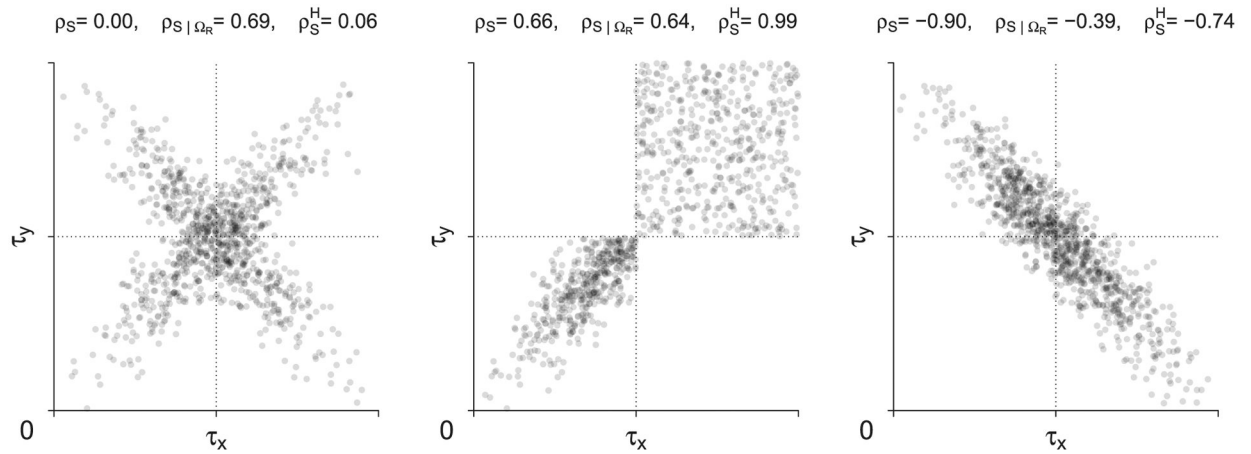
FIGURE 2.

Restricted Spearman's correlation, $\rho_{S \mid \Omega_R}$ (left panel), and highest rank Spearman's correlation, ρ_S^H (right panel), for Frank's copula family for different restricted regions defined by τ_X and $\tau_Y (\tau_X = \tau_Y)$: 0.5th (50% censored), 0.6th (40% censored), 0.8th (20% censored) quantiles. A diagonal gray line is added for reference. Although the plots are generated based on data under strict type I censoring, the population parameters are the same for generalized type I censoring and are invariant to the rate of censoring within the restricted region

FIGURE 3.

Example of bivariate distributions and their population parameters ρ_S with no censoring, and ρ_S^H and $\rho_{S \mid \Omega_R}$ with strict type I censoring with $\Omega_R = [0,\ \tau_X) \times [0,\ \tau_Y)$. The proportions of observed double events in the left, middle, and right panels are 25%, 43%, and 7%, respectively. Drawn are 1000 points randomly selected from the underlying distributions. Although the plots are based on strict type I censoring, the population parameters are the same for generalized type I censoring and are invariant to the rate of censoring within Ω_R
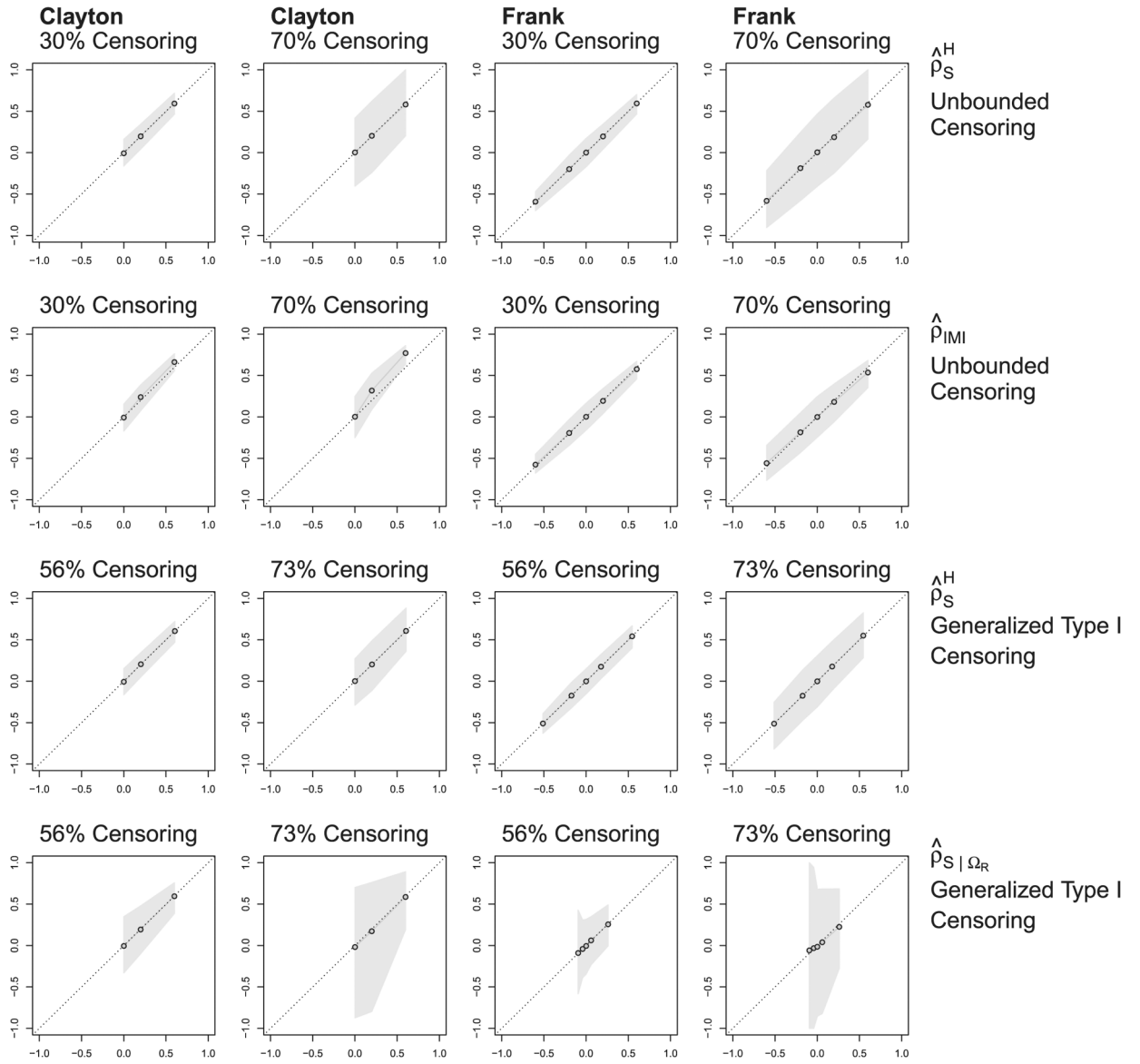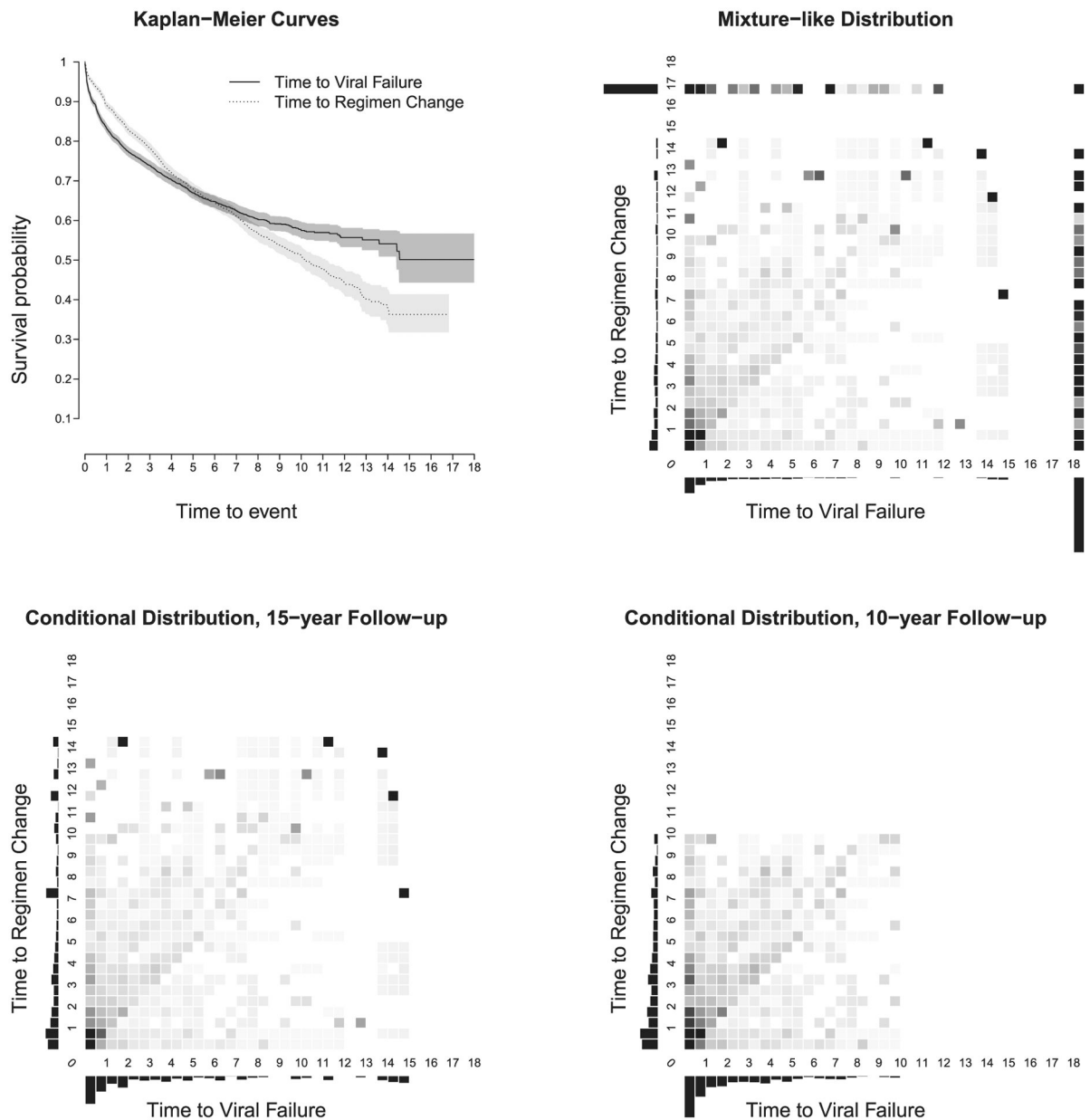
FIGURE 4.

Point estimates (X-axis) vs population parameters (Y-axis) under different bivariate censoring scenarios. The top and second rows are $\hat{\rho}_S^H$ and $\hat{\rho}_{IMI}$ as estimators of the overall Spearman's correlation, ρ_S. The third row is $\hat{\rho}_S^H$ as an estimator of ρ_S^H. The bottom row is $\hat{\rho}_{S \mid \Omega_R}$ as an estimator of $\rho_{S \mid \Omega_R}$. The columns represent Clayton's and Frank's copulas. The population parameters for Clayton's family are 0, 0.2, and 0.6 for all estimates. For Frank's family, the population parameters of ρ_S are −0.6, −0.2, 0.2, and 0.6; the population parameters of ρ_S^H are −0.512, −0.173, 0.180, and 0.545; the population parameters of $\rho_{S \mid \Omega_R}$ are −0.098, −0.042, 0.058, and 0.261. The dots are the mean point estimates based on 1000 simulations. The shaded areas represent the 0.025th and 0.975th quantiles. For generalized type I censoring, the restricted region, Ω_R, was defined by the median survival times

FIGURE 5.

Upper left: Kaplan–Meier curves for time to viral failure and time to regimen change, where time is measured in years. Upper right: bivariate probability mass function for the mixture-like distribution, $\hat{S}^H(dx, dy)$. Lower left: conditional bivariate probability mass function for 15-year follow-up. Lower right: conditional bivariate probability mass function for 10-year follow-up. For probability mass functions, the bars on the left and on the bottom represent histograms of the univariate survival mass for each event. The probability mass function was computed from the Dabrowska's survival surface and then aggregated over half-year bivariate time periods. After aggregation, any negative values were set to 0. Lighter shade represents smaller values

TABLE 1

Correlation of time to viral failure and time to regimen change in CCASAnet cohort measured using three estimators: $\hat{\rho}_S^H$, $\hat{\rho}_{S\,|\,\Omega_R}$, and $\hat{\rho}_{IMI}$. n is the number of subjects for each subgroup. Columns P_{VR}, $P_{\overline{V}R}$, $P_{V\overline{R}}$, and $P_{\overline{VR}}$ show the percent of subjects having both events (P_{VR}), having regimen change event but censored viral failure ($P_{\overline{V}R}$), having viral failure event but censored regimen change ($P_{V\overline{R}}$), and having both events censored ($P_{\overline{VR}}$). For $\hat{\rho}_S^H$ and $\hat{\rho}_{S\,|\,\Omega_R}$, the numbers inside brackets are 95% confidence intervals (CIs) estimated from 1000 bootstrap samples. Column $\hat{\rho}_{IMI}$ shows the estimates of Schemper et al. (2013) and their CIs

Subgroup	n	P_{VR}	$P_{\overline{V}R}$	$P_{V\overline{R}}$	$P_{\overline{VR}}$	$\hat{\rho}_S^H$	$\hat{\rho}_{S\,|\,\Omega_R}=[0,15]\times[0,15]$	$\hat{\rho}_{S\,|\,\Omega_R}=[0,10]\times[0,10]$	$\hat{\rho}_{IMI}$
All	6691	16.1	12.2	12.5	59.1	0.35 [0.27, 0.44]	0.65 [0.30, 0.97]	0.26 [0.17, 0.36]	0.37 [0.33, 0.41]
Male	5185	14.8	12.0	12.3	60.9	0.32 [0.22, 0.42]	0.57 [0.14, 1.00]	0.32 [0.18, 0.44]	0.36 [0.30, 0.41]
Female	1506	20.7	13.1	13.3	52.9	0.45 [0.30, 0.57]	0.70 [0.20, 1.00]	0.13 [−0.04, 0.32]	0.40 [0.32, 0.48]
A	1040	19.2	23.0	11.5	46.2	−0.06 [−0.21, 0.30]	0.73 [−0.24, 1.00]	0.19 [0.03, 0.34]	0.26 [0.00, 0.37]
B	975	13.7	16.3	11.9	58.1	0.52 [0.14, 0.74]	−0.39 [−0.54, 0.41]	−0.43 [−1.00, 0.08]	0.25 [0.10, 0.38]
C	2225	8.9	5.5	12.1	73.4	0.45 [0.37, 0.53]	0.69 [0.42, 0.93]	0.58 [0.34, 0.80]	0.55 [0.43, 0.66]
D	138	18.1	17.4	10.1	54.3	−0.18 [−0.53, 0.44]	0.21 [−1.00, 1.00]	0.50 [−0.09, 0.96]	0.29 [0.00, 0.55]
E	2313	22.4	11.8	13.8	51.9	0.45 [0.36, 0.53]	0.42 [0.10, 0.73]	0.17 [−0.01, 0.34]	0.36 [0.30, 0.42]