Published in final edited form as:

Stat Med. 2015 February 28; 34(5): 844–858. doi:10.1002/sim.6371.

Assessing Temporal Agreement between Central and Local Progression-Free Survival Times

Donglin Zeng^a, Emil Cornea^a, Jun Dong^b, Jean Pan^b, and Joseph G. Ibrahim^a

^aDepartment of Biostatistics, CB7420, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27516, U.S.A

^bAmgen Inc., One Amgen Center Drive, Thousand Oaks, California 91320, U.S.A.

Abstract

In oncology clinical trials, progression-free survival (PFS), generally defined as the time from randomization until disease progression (PD) or death, has been a key endpoint to support licensing approval (e.g. [1]). When PFS is the primary or co-primary endpoint, it is recommended to have tumor assessments verified by an independent review committee (IRC) blinded to study treatments, especially in open-label studies (see [1]). It is considered reassuring about the lack of reader-evaluation bias if treatment effect estimates from the investigators' and IRCs' evaluations agree. The agreement between these evaluations may vary for subjects with short or long PFS, while there exist no such statistical quantities that can completely account for this temporal pattern of agreements. Therefore, in this paper, we propose a new method to assess temporal agreement between two time-to-event endpoints, while the two event times are assumed to have a positive probability of being identical. This method measures agreement in terms of the two event times being identical at a given time or both being greater than a given time. Overall scores of agreement over a period of time are also proposed. We propose a maximum likelihood estimation to infer the proposed agreement measures using empirical data, accounting for different censoring mechanisms, including reader's censoring (event from one reader dependently censored by event from the other reader). The proposed method is demonstrated to perform well in small-sample via extensive simulation studies and is illustrated through a head and neck cancer trial.

Keywords

Agreement measure; Copula distribution; EM algorithm; Kendall's τ ; Progression-free survival; Reader's censoring; Weibull distribution

1. Introduction

Progression-free survival (PFS) has been a key endpoint to support licensing approval in some cancer settings (e.g. [1]). PFS is generally defined as the time from randomization until disease progression (PD) or death. In oncology clinical trials, radiologic tumor assessments are usually performed at pre-defined intervals until PD as determined by an

investigator. When PFS is the primary or co-primary endpoint, it is recommended to have tumor assessments verified by an independent review committee (IRC) blinded to study treatments, especially in open-label studies (see [1]). The primary analysis of PFS for these studies may be based on the IRC's evaluation and a sensitivity analysis may be performed using the investigators' evaluation, or vice versa. It is generally considered reassuring about the lack of reader-evaluation bias if treatment effect estimates from the investigators' and IRC's evaluations agree.

Historically, less attention has been paid to the agreement (or disagreement) between the tumor assessments by the investigators and the IRC itself, compared to its impact on the treatment effect estimate. The agreement is of interest for several reasons. First, an agreement in treatment effect estimates does not necessarily mean the individual time-to-event times are the same or close. As pointed out in [2], hazard ratio estimates based on the investigators' and IRC's evaluations were highly concordant in their meta-analysis, but there existed a marked degree of individual patient discordance in the timing or occurrence of PD. In an extreme case, when the PD time by the IRC's evaluation and the investigators' evaluation differ by a constant for all subjects, the treatment effects measured by the hazard ratio are identical from both data sources, even though the evaluations of PD times never agree. Second, the median PFS time by treatment arm is often of great interest to clinicians. In the aforementioned extreme case, the median PFS times differ by the same constant between two data sources, in either treatment arm. Last, it is widely recognized that the investigators' evaluation of PD is potentially biased when the patients' treatment assignments are known to the investigators. Although the IRC is blinded to study treatments, and hence expected to reduce observation bias in the investigators' evaluation, results of the IRC's review may be subject to a new bias introduced by a special type of informative censoring: in many clinical trials, tumor assessments are discontinued when PD is determined by investigators, which precludes an IRC evaluation of PD in subsequent assessments. The IRC's evaluation of PD can be dependently censored by the investigator's evaluation of PD, which results in the IRC's evaluation of PFS be dependently censored if death is not further observed for the subject. Hence, informative censoring occurs for the IRC's evaluation of PFS endpoint. In this paper, we refer to this type of informative censoring as Reader Censoring (RC). Therefore, it is desirable to have a measure that can directly quantify the agreement between the IRC's and the investigators' evaluations, and also develop a robust estimator of an agreement measure in the presence of RC.

There are several agreement measures for assessing dependence or agreement for censored bivariate time-to-event data in the literature. The rank-based Kendall's coefficient of concordance τ (Section 4.2, [3]) is a measure of overall dependence. It is simple and can be estimated non-parametrically for censored data, but does not measure the degree of agreement at a single time point. In the aforementioned extreme case, the data achieves the maximum value of Kendall's τ , even though there is no agreement of PD time at all. Liu et al. [4] provided an estimation method of the concordance correlation coefficient for time-to-event data. It is a correlation type of measure and has the same issue as Kendall's τ . Guo and Manatunga [5] proposed a modified weighted kappa coefficient to measure agreement between discrete bivariate survival times, but it requires discretization of continuous

outcomes. Amit et al. [2] also proposed two discrepancy rates defined as the simple proportion of subjects whose PD time is strictly greater by a reader than the other. These rates are naïve approaches by definition and do not fully utilize the temporal nature of time-to-event data. Guo et al. [6] proposed a new agreement measure, which is formulated as the chance-corrected concordance between survival processes on the absolute distance scale.

In this paper, we propose a new method to assess temporal agreement between two time-to-event endpoints, where the two event times are assumed to have a positive probability of being identical. This method measures agreement in terms of the two event times being identical at a given time or both being greater than a given time. Overall scores of agreement over a period of time are also proposed. While the agreement measures are defined based on the underlying distributions of the two time-to-event endpoints, the proposed estimation method provides unbiased estimates of the proposed agreement measures (at a given time and overall) in the presence of RC in the observed data, for example, the IRC's evaluation of PFS. Although the focus of this paper is not to assess the impact of this agreement to the estimated treatment effect, the proposed measures can be applied to combined treatment arms as well as within each treatment arm to help assess the potential bias in estimating the treatment effect.

The development of our method is motivated by a small phase 2 head and neck cancer trial. A random subset of all randomized subjects is used here for demonstration purpose. Among 92 subjects followed-up in the trial, the local assessment yields 82 local PFS events while the central assessment gives 72 events and the number of agreed events is 35. In this paper, we will apply our development method to assess temporal agreement of the PFS between the local and central assessments in this trial dataset.

The paper is organized as follows. In Section 2, we present the methodology and notation for the underlying two time-to-event distributions. Furthermore, we propose time-varying agreement measures to assess the temporal agreement between two time-to-event endpoints. In Section 3, we propose a mixture bivariate survival model and, in Section 4, we describe an EM algorithm to estimate the agreement measures induced by this model. Section 5 studies the small-sample performance of the proposed methods through extensive simulation studies. A real data example from the aforementioned head and neck cancer study is analyzed in Section 6. Finally, we conclude with a discussion in Section 7.

2. Time-Varying Agreement Measures

Let T_c and T_l denote the two time-to-event endpoints under consideration (e.g. PFS per central and local assessment). The idea of defining an agreement measure is such that the analysis results based on comparing the hazard rates of the two event times should indicate similarity when the agreement between the two event times is high. In censored data analysis, the estimates of the hazard ratios, for example, fitting the Cox proportional hazards model or performing log-rank tests, are fully dependent on the behavior of the subjects at risk at each time t . In other words, the analysis relies on sufficient information regarding the number of events and the number of subjects at risk at any time t . Therefore, if T_l and T_c agree perfectly, we expect (pretending T_l and T_c to be discrete)

$$P(T_l=t)=P(T_c=t), \quad P(T_l \geq t)=P(T_c \geq t),$$

or equivalently,

$$P(T_c=t|T_l=t)=1, \quad P(T_c>t|T_l>t)=1.$$

This motivates us to define the following two time-varying agreement measures:

$$pA(t)=P(T_c=t|T_l=t), \quad nA(t)=P(T_c>t|T_l>t).$$

The above two quantities can be interpreted in terms of sensitivity and specificity as well. Given that the local event has not occurred prior to time t , at time t , T_l either occurs (“disease”) or not (“non-disease”). Treating T_c as a diagnostic measure, we then observe that pA is the sensitivity measure while nA is the specificity measure in the usual medical diagnostic context.

We vary time t from 0 to the maximum follow-up time T_E to obtain the curves $pA(t)$ and $nA(t)$. For practical purpose, it will be convenient to use a summary quantity of the curves to assess an overall agreement between the two types of events. To this end, we define the weighted area under these curves, denoted $wAUC(pA)$ and $wAUC(nA)$, respectively, by

$$\begin{aligned} wAUC(pA) &= \int_0^{T_E} pA(t) f_{T_l}(t) dt / \int_0^{T_E} f_{T_l}(t) dt, \\ wAUC(nA) &= \int_0^{T_E} nA(t) f_{T_l}(t) dt / \int_0^{T_E} f_{T_l}(t) dt, \end{aligned}$$

where $f_{T_l}(t)$ is the probability density function of T_l . The areas $wAUC(pA)$ and $wAUC(nA)$ can be used to measure the overall agreement. The measure $wAUC(pA)$ over the entire follow-up period of time is actually the agreement probability $P(T_c = T_l)$.

3. Mixture Models for Estimating Time-Varying Agreements

To estimate the time-varying agreement measures $pA(t)$ and $nA(t)$ using censored data, we introduce a mixture model for the bivariate distribution of (T_c, T_l) . Specifically, we assume that there is some positive chance for T_c and T_l to match. We define the agreement probability as $p = \Pr(T_c = T_l)$ and assume $p > 0$. Thus, we model the joint distribution by considering two different situations: given $T_c = T_l$, $\Pr(T_c > t | T_c = T_l) = S(t)$; while given $T_c \neq T_l$, we assume (T_c, T_l) follows a continuous bivariate distribution with cumulative distribution function (CDF) denoted by $Q(t, s)$. Therefore, for any t and s with $t \leq s$, we can write down the joint distribution for (T_c, T_l) as

$$\begin{aligned} \Pr(T_c \leq t, T_l \leq s) &= \Pr(T_c \leq t, T_l \leq t, T_c=T_l) + \Pr(T_c \leq t, T_l \leq s, T_c \neq T_l) \\ &= p[1 - S(t)] + (1 - p) \Pr(T_c \leq t, T_l \leq s | T_c \neq T_l) \\ &= p[1 - S(t)] + (1 - p) Q(t, s). \end{aligned}$$

By symmetry, for any t and s ,

$$Pr(T_c \leq t, T_l \leq s) = p[1 - S(\min(t, s))] + (1 - p)Q(t, s).$$

Another way of understanding this distribution is to introduce a latent Bernoulli variable $B = I(T_c = T_l)$. Then, given $B = 1$, $T_c = T_l$ follows the distribution $1 - S(t)$ while given $B = 0$, (T_c, T_l) follows the distribution $Q(t, s)$. This latent variable will be useful for developing the EM algorithm for empirical estimation later.

With practical sample sizes, estimating $S(t)$ non-parametrically may not be numerically stable. On the other hand, estimating $Q(t, s)$ nonparametrically is not feasible if there exists RC in the observed data. Therefore, in the following development, we will adopt a general class of parametric distributions for the estimation. Specifically, we will use the Weibull distribution to model $S(t)$ while assume $Q(t, s)$ to be from a copula distribution (e.g. see [7, 8]). We take $S(t) = \exp[-(\lambda_0 t)^{\alpha_0}]$, which is the survival function of the Weibull distribution $Weibull(\alpha_0, \lambda_0)$, with shape parameter λ_0 and inverse scale parameter λ_0 . We assume $Q(t, s)$ is from the following copula distribution: the CDFs of the marginal distributions of T_l and T_c given $T_l \leq T_c$ are $1 - \exp[-(\lambda_l t)^{\alpha_l}]$ and $1 - \exp[-(\lambda_c t)^{\alpha_c}]$, respectively; the joint distribution of T_l and T_c is given as

$$\left\{ \Phi^{-1} \left(1 - e^{-(\lambda_c T_c)^{\alpha_c}} \right), \Phi^{-1} \left(1 - e^{-(\lambda_l T_l)^{\alpha_l}} \right) \right\} \sim N \left(0, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \right),$$

where $\Phi(\cdot)$ is the CDF of the univariate standard normal distribution and ρ describes their correlation. For this setting, we obtain

$$Q(t, s) = P \left(Z_1 \leq \Phi^{-1} \left(1 - e^{-(\lambda_c t)^{\alpha_c}} \right), Z_2 \leq \Phi^{-1} \left(1 - e^{-(\lambda_l s)^{\alpha_l}} \right) \right),$$

where (Z_1, Z_2) follows the above bivariate normal distribution.

With the above distributions, we can easily calculate the proposed time-varying agreement measures as

$$\begin{aligned} pA(t) &= \frac{pf_{\alpha_0, \lambda_0}(t)}{pf_{\alpha_0, \lambda_0}(t) + (1-p)f_{\alpha_l, \lambda_l}(t)} \\ &= \frac{p\alpha_0\lambda_0(\lambda_0 t)^{\alpha_0-1}e^{-(\lambda_0 t)^{\alpha_0}}}{p\alpha_0\lambda_0(\lambda_0 t)^{\alpha_0-1}e^{-(\lambda_0 t)^{\alpha_0}} + (1-p)\alpha_l\lambda_l(\lambda_l t)^{\alpha_l-1}e^{-(\lambda_l t)^{\alpha_l}}}, \end{aligned}$$

and

$$\begin{aligned} nA(t) &= \frac{pS_{\alpha_0, \lambda_0}(t) + (1-p)[S_{\alpha_l, \lambda_l}(t) + S_{\alpha_c, \lambda_c}(t) - 1 + Q(t, t)]}{pS_{\alpha_0, \lambda_0}(t) + (1-p)S_{\alpha_l, \lambda_l}(t)} \\ &= \frac{pe^{-(\lambda_0 t)^{\alpha_0}} + (1-p)[e^{-(\lambda_l t)^{\alpha_l}} + e^{-(\lambda_c t)^{\alpha_c}} - 1 + Q(t, t)]}{pe^{-(\lambda_0 t)^{\alpha_0}} + (1-p)e^{-(\lambda_l t)^{\alpha_l}}}, \end{aligned}$$

where $f_{\alpha,\lambda}(t)$ and $S_{\alpha,\lambda}(t)$ are the density and, respectively, the survival functions of the Weibull distribution $Weibull(\alpha, \lambda)$.

As an illustrative example, we use some distribution settings from the simulation section and plot the curves of these agreement measures in Figure 1, where we fix $\rho = 0.5$ while varying $p = 0, 0.2, 0.4, 0.6, 0.8, 1$, or we fix $p = 0.5$ while varying $\rho, \rho = 0, 0.2, 0.4, 0.6, 0.8$, while all the other parameters are set to the values in Tables 1 & 2. Values for pA and nA at selected times are also given in Tables 1 & 2. Furthermore, we plot the weighted area under the curves from 0 to T_E , where T_E varies from 0.5 to 3.0, $p = 0.2, 0.4, 0.6, 0.8$, and $\rho = 0.5$, in Figure 2. For comparison, we also plot the Kendall's τ values over $[0, T_E]$. Values for $wAUC(pA)$ and $wAUC(nA)$ for selected T_E 's are also given in Table 3. Tables 1 and 2 and Figure 1 show that for given T_E , $pA(t)$ and $nA(t)$ vary over time. Table 3 and Figure 2 show that by varying T_E , $wAUC(pA)$ and $wAUC(nA)$ vary as well and become flat after most events have occurred.

4. Observed Likelihood and Inference

In general, the observed data from n i.i.d subjects are

$$(Y_{ci} = \min(T_{ci}, C_{ci}), Y_{li} = \min(T_{li}, C_{li}), \Delta_{ci}, \Delta_{li}), \quad i=1, \dots, n, \quad (1)$$

where C_{ci} and C_{li} are the respective censoring times for T_{ci} and T_{li} , $I_{li} = I(T_{li} \leq C_{li})$, and $I_{ci} = I(T_{ci} \leq C_{ci})$. We can allow the situation when the RC is present, in which case the observed data are

$$(Y_{ci} = \min(T_{ci}, \min(C_{ci}, \{(1 - \Delta_{li}) \infty + \Delta_{li} T_{li}\})), Y_{li} = \min(T_{li}, C_{li}), \Delta_{ci}, \Delta_{li}), \quad i=1, \dots, n, \quad (2)$$

where C_{ci} and C_{li} are the respective censoring times for T_{ci} and T_{li} , $I_{li} = I(T_{li} \leq C_{li})$, and $I_{ci} = I(T_{ci} \leq \min(C_{ci}, \{(1 - I_{li}) \infty + I_{li} T_{li}\}))$. Here, we use the convention $0 \cdot \infty = 0$.

Assuming that (C_{ci}, C_{li}) is noninformative for the joint distribution of (T_{ci}, T_{li}) , we can write down the observed likelihood function as follows. In the case of data without RC, for a given observation $(y_{ci}, y_{li}, \delta_{ci}, \delta_{li})$, the likelihood contributions from each subject are from one of four groups:

$$\begin{aligned} (\delta_{ci}=1, \delta_{li}=1) & \quad Pr(T_c=y_{ci}, T_l=y_{li}) Pr(C_c \geq y_{ci}, C_l \geq y_{li}), \\ (\delta_{ci}=1, \delta_{li}=0) & \quad Pr(T_c=y_{ci}, T_l > y_{li}) Pr(C_c \geq y_{ci}, C_l = y_{li}), \\ (\delta_{ci}=0, \delta_{li}=1) & \quad Pr(T_c > y_{ci}, T_l=y_{li}) Pr(C_c = y_{ci}, C_l \geq y_{li}), \\ (\delta_{ci}=0, \delta_{li}=0) & \quad Pr(T_c > y_{ci}, T_l > y_{li}) Pr(C_c = y_{ci}, C_l = y_{li}). \end{aligned} \quad (3)$$

In the case of data with RC, it comes from one of the following four groups:

$$\begin{aligned} (\delta_{ci}=1, \delta_{li}=1) & \quad Pr(T_c=y_{ci}, T_l=y_{li}) Pr(\min(C_c, y_{li}) \geq y_{ci}, C_l \geq y_{li}), \\ (\delta_{ci}=1, \delta_{li}=0) & \quad Pr(T_c=y_{ci}, T_l > y_{li}) Pr(C_c \geq y_{ci}, C_l = y_{li}), \\ (\delta_{ci}=0, \delta_{li}=1) & \quad Pr(T_c > y_{ci}, T_l=y_{li}) Pr(\min(C_c, y_{li}) = y_{ci}, C_l \geq y_{li}), \\ (\delta_{ci}=0, \delta_{li}=0) & \quad Pr(T_c > y_{ci}, T_l > y_{li}) Pr(C_c = y_{ci}, C_l = y_{li}). \end{aligned} \quad (4)$$

It is clear from the above likelihood formulation that no matter if the censoring is RC or even partially RC (i.e., only some central events may be censored by local events), we only

need the likelihood from the observed local/central events to make inference about the agreement measures.

Specifically, we maximize the observed likelihood function where the part concerning the joint distribution of (C_c, C_l) can be dropped. The maximization can be carried out via the EM algorithm by treating the latent status, $B_i = I(T_{ci} = T_{li})$ as missing data. Then the complete-data likelihood function for the above expression corresponds to:

$$\begin{aligned} (\delta_{ci}=1, \delta_{li}=1) & \{p \ f(y_{li})\}^{B_i} \{(1-p) q(y_{ci}, y_{li})\}^{1-B_i}, \\ (\delta_{ci}=1, \delta_{li}=0) & \{p \ f(y_{ci})\}^{B_i} \left\{ (1-p) \int_{v>y_{li}} q(y_{ci}, v) dv \right\}^{1-B_i}, \\ (\delta_{ci}=0, \delta_{li}=1) & \{p \ f(y_{li})\}^{B_i} \left\{ (1-p) \int_{u>y_{ci}} q(u, y_{li}) du \right\}^{1-B_i}, \\ (\delta_{ci}=0, \delta_{li}=0) & \{p \ S(\max(y_{ci}, y_{li}))\}^{B_i} \{(1-p) Q^c(y_{ci}, y_{li})\}^{1-B_i}, \end{aligned} \quad (5)$$

where $f(t)$ is the derivative of $-S(t)$, $q(u, v)$ is $^2Q(u, v)/u \ v$, and $Q^c(t, s)$ is the probability $P(T_c > t, T_l > s)$, which is equal to $1 - Q(t, \infty) - Q(\infty, s) + Q(t, s)$. The complete-data log-likelihood function becomes

$$\begin{aligned} & \sum_{i=1}^n \delta_{ci} \delta_{li} \{B_i \log f(y_{li}) + (1-B_i) \log q(y_{ci}, y_{li})\} \\ & + \sum_{i=1}^n \delta_{ci} \delta_{li} \{B_i \log p + (1-B_i) \log(1-p)\} \\ & + \sum_{i=1}^n \delta_{ci} (1-\delta_{li}) \left\{ B_i \log f(y_{ci}) + (1-B_i) \log \int_{v>y_{li}} q(y_{ci}, v) dv \right\} \\ & + \sum_{i=1}^n \delta_{ci} (1-\delta_{li}) \{B_i \log p + (1-B_i) \log(1-p)\} \\ & + \sum_{i=1}^n (1-\delta_{ci}) \delta_{li} \left\{ B_i \log f(y_{li}) + (1-B_i) \log \int_{\mu>y_{ci}} q(\mu, y_{li}) d\mu \right\} \\ & + \sum_{i=1}^n (1-\delta_{ci}) \delta_{li} \{B_i \log p + (1-B_i) \log(1-p)\} \\ & + \sum_{i=1}^n (1-\delta_{ci}) (1-\delta_{li}) \{-B_i \log S(\max(y_{ci}, y_{li})) + (1-B_i) \log Q^c(y_{ci}, y_{li})\} \\ & + \sum_{i=1}^n (1-\delta_{ci}) (1-\delta_{li}) \{B_i \log p + (1-B_i) \log(1-p)\}. \end{aligned} \quad (6)$$

For the E-step, we calculate the posterior probability of $B_i = 1$, denoted by p_i , given the observed data as follows:

$$\begin{aligned} \delta_{ci}=1, \delta_{li}=1 & \quad p_i = I(y_{ci}=y_{li}); \\ \delta_{ci}=1, \delta_{li}=0 & \quad p_i = \frac{I(y_{li}<y_{ci})f(y_{ci})p}{I(y_{li}<y_{ci})f(y_{ci})p + \int_{v>y_{li}} q(y_{ci}, v)dv(1-p)}; \\ \delta_{ci}=0, \delta_{li}=1 & \quad p_i = \frac{I(y_{ci}<y_{li})f(y_{li})p}{I(y_{ci}<y_{li})f(y_{li})p + \int_{\mu>y_{ci}} q(\mu, y_{li})d\mu(1-p)}; \\ \delta_{ci}=0, \delta_{li}=0 & \quad p_i = \frac{S(\max(y_{ci}, y_{li}))p}{S(\max(y_{ci}, y_{li}))p + Q^c(y_{ci}, y_{li})(1-p)}. \end{aligned} \quad (7)$$

For the M-step, we estimate the parameters in $S(t)$ by maximizing

$$\sum_{i=1}^n p_i \{ \delta_{ci} \delta_{li} \log f(y_{li}) + \delta_{ci} (1-\delta_{li}) \log f(y_{ci}) + (1-\delta_{ci}) \delta_{li} \log f(y_{li}) + (1-\delta_{ci}) (1-\delta_{li}) (-\log S(\max(y_{ci}, y_{li}))) \}, \quad (8)$$

and the parameters in $Q(u, v)$ by maximizing

$$\sum_{i=1}^n (1-p_i) \left\{ \delta_{ci} \delta_{li} \log q(y_{ci}, y_{li}) + \delta_{ci} (1-\delta_{li}) \int_{v>y_{li}} q(y_{ci}, v) dv + (1-\delta_{ci}) \delta_{li} \int_{\mu>y_{ci}} q(\mu, y_{li}) d\mu + (1-\delta_{ci}) (1-\delta_{li}) \log Q^c(y_{ci}, y_{li}) \right\},$$

We estimate p by $\sum_{i=1}^n p_i/n$. We iterate between the E- and M-steps until convergence. For inference, the information is calculated based on the Louis formula [9] and its inverse is used as the sample estimate of the asymptotic covariance of the maximum likelihood estimators.

5. Simulation Study

We conducted two sets of simulations, one for data with RC and the other for data without RC, to evaluate the accuracy of the parameter estimates, their associated variance estimates, and the estimates of the agreement measures $pA(t)$ and $nA(t)$ for our proposed method. We also evaluated the performance of the true pA , nA , $wAUC(pA)$, and $wAUC(nA)$. In addition, we also evaluated the performance of Kendall's τ and its estimate. All computations were done in R.

For these sets of simulations, we assume that along the diagonal (i.e., $T_c = T_l$) (T_c, T_l) is parametrically modeled by the Weibull distribution with shape parameter α_0 and inverse scale parameter λ_0 , while outside of the diagonal (i.e., $T_c \neq T_l$), T_c and T_l marginally follow Weibull distributions with parameters α_c, λ_c and, respectively, α_l, λ_l , and their joint distribution is modeled by the copula distribution based on the bivariate normal distribution with zero mean, unit variances, and correlation parameter ρ , as in the example of Section 3. Recall that we denoted the agreement probability, $P(T_c = T_l)$, by p . For a sample size n , we generated the simulated data as follows. We generated $B_i \in \{0, 1\}$, $i = 1, \dots, n$, independently from $Bernoulli(p)$. The subjects with $B_i = 1$ have the event times along the diagonal, i.e. $T_{ci} = T_{li}$, while those with $B_i = 0$ have the event times outside the diagonal, i.e. $T_{ci} \neq T_{li}$. Given $B_i = 1$, the common event time $T_{ci} = T_{li}$ is generated as a random sample from a $Weibull(\alpha_0, \lambda_0)$ distribution, while given $B_i = 0$, (T_{ci}, T_{li}) is generated so that $(\Phi^{-1}(1 - e^{-(\lambda_c T_{ci})^{\alpha_c}}), \Phi^{-1}(1 - e^{-(\lambda_l T_{li})^{\alpha_l}}))$ is a random sample from the bivariate normal distribution $N_2(\mathbf{0}, \Sigma)$, with

$$\Sigma = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}.$$

For the time-to-event without RC scenario, the simulated data (Y_{ci}, Y_{li}, C_i) are obtained by censoring T_{ci} and T_{li} with a single censoring time C_i , i.e. $C_{ci} = C_{li}$, and an end time T_E (maximum follow-up time). That is, $Y_{ci} = \min(T_{ci}, C_i, T_E)$, $Y_{li} = \min(T_{li}, C_i, T_E)$, $C_i = I(T_{ci} \leq \min(C_i, T_E))$, and $_{li} = I(T_{li} \leq \min(C_i, T_E))$, $i = 1, \dots, n$. The censoring times (C_i 's) are independent of the T 's and independently distributed from $Weibull(\alpha_{cens}, \lambda_{cens})$, $i = 1, \dots, n$. For the time-to-event data with RC scenario, in addition to the time-to-event data without RC case, the central time T_{ci} is further censored by the local time T_{li} , $i = 1, \dots, n$, as presented in Section 4.

For each simulation setting, we considered sample sizes of $n = 100$ and 200 and varied the agreement proportion p from 0.2 to 0.8 . Furthermore, we let $\lambda_0 = 0.90$, $\alpha_0 = 1.50$, $\lambda_c = 1.35$, $\alpha_c = 1.20$, $\lambda_l = 1.40$, $\alpha_l = 1.30$, and $\pi = 0.50$. For censoring, we set $\alpha_{cens} = 1.10$, $\lambda_{cens} = 0.75$, and $T_E = 3.0$ resulting in 36.4% to 43.8% censoring rates. The parametric estimation is based on the Weibull distribution. In each case, we generated 500 datasets. For each simulated dataset, we applied the EM algorithm to obtain the maximum likelihood estimators for the model parameters and their standard errors using the Louis formula. Additionally, we calculated the agreement measures $pA(t)$ and $nA(t)$, at time points $t = 0.5$,

1.5, and 2.5. The standard errors for the latter quantities were obtained using the delta method.

Tables 1 and 2 give the results from the first simulation setting with RC and from the second simulation setting without RC, respectively, where column “SD” denotes the empirical standard deviations of the obtained estimates, column “SE” is the estimated standard errors by the Louis formula (in [9]), and column “CP” is the 95% coverage probability. Both tables indicate that the proposed estimation and inference work well in small samples: the biases are small, the estimated standard errors agree well with the empirical standard deviations, and the coverage probabilities are close to the nominal level. The coverage probabilities of the estimated $pA(t)$ or $nA(t)$ tend to be larger for the time point at the tail but are improved when increasing the sample size. Additionally, for each agreement probability, $p = 0.2, 0.6, 0.8$, we calculated the estimates and biases of the weighted areas under the curves, $wAUC(pA)(T_E)$ and $wAUC(nA)(T_E)$, of pA and nA respectively, and Kendall's coefficient of concordance $\tau(T_E)$ as functions of the maximum follow-up time T_E , in the time-to-event data with or without RC. The results for maximum follow-up times $T_E = 0.5, 1.5, 2.5$, and 3.0 are presented in Table 3. These results show that the estimates are quite accurate and that $wAUC(pA)$ approximates the true p for large enough T_E when most events have occurred.

Finally, we conducted some simulations to examine the robustness of the estimated pA or nA to the model misspecification. In the simulations, instead of using the copula model for Q , we used the following distribution. Assume that T_c is marginally $Weibull(a_c, \lambda_c)$ distributed and $Pr(T_l = T_c | T_c = t) = p(t)$, where $p(t), 0 \leq p(t) \leq 1$, for $t \geq 0$, is a given probability function. Given $T_l \leq T_c$, we assume T_l given $T_c = t$ satisfies $T_l = \exp(X)$ where X is normally distributed with mean $\log(t)$ and variance σ^2 . We set $\lambda_c = 1.35$, $a_c = 1.2$, $p(t) = ae^{-rt}$, with $a = 0.8$ and $r = 0.4$, and $\sigma^2 = 0.5$. For this setting, the true value of the agreement probability is $p = 0.620$. For this simulation study, we considered the sample size $n = 200$, and we generated 500 datasets. The results are shown in Table 4. The biases of $wAUC(pA)$ and $wAUC(nA)$ are still relatively small, even though the actual joint distribution of (T_c, T_l) given $T_c \leq T_l$ is not of copula type. The agreement measure nA seems to be more sensitive to this model assumption than pA when $t \rightarrow 1$.

6. Head and Neck Cancer Trial

We consider data from a relatively small phase 2 head and neck cancer clinical trial with $n = 92$ subjects (a random subset of all randomized subjects for demonstration purpose). The PFS times (in days) were recorded from two different sources, an IRC's evaluation (central) and the investigators' evaluation (local). The range of the central PFS times is from 0 to 861 days and that of the local PFS times is from 0 to 832 days. There are 72 central and 82 local events, respectively. For the purpose of analyzing the data, we exclude the subjects with zero central time or zero local time or zero both times from the analysis. There are 89 subjects with nonzero observed times. For the purpose of this analysis, we rescaled the times-to-event to years, i.e., we divided the event times (in days) by 365.25. The event times are now in the range $[0, 3]$, similar to the time range in our simulation studies. The maximum

follow-up time is $T_E = 2.36$. A descriptive summary of the data is shown in Table 5. The probability of exact agreement of T_c and T_l is at least 30%.

We fit the model described in Section 3. The estimates, standard deviations, and 95% confidence intervals for the parameters p , λ_0 , a_0 , λ_c , a_c , λ_l , a_l , ρ are shown in Table 6. The estimate of the agreement probability p is $\hat{p} = 0.413$. The estimates of the agreement measures $pA(t)$ and $nA(t)$, together with their 95% pointwise confidence bands, are shown in Table 7 and displayed in Figure 3. The weighted area under the estimated curves $\widehat{pA}(t)$ and $\widehat{nA}(t)$ are 0.412 and 0.843. The $wAUC$ s are calculated over the study duration, that is, over the time period from 0 to the last event/censoring time, $T_E = 2.36$. Table 7 shows the estimated values of $wAUC(pA)$, $wAUC(nA)$, and Kendall's coefficient of concordance τ . The statistic τ only reflects the rank correlation between the two outcomes, T_c and T_l , at any time, while our measures $pA(t)$ and $nA(t)$ give the exact agreement probabilities at any time t . Our measures are more informative than the Kendall's coefficient τ .

The results indicate that both the local and central times agreed well at the beginning and end of this study. The relatively higher agreement at the end of the study is mostly due to death for subjects without either local or central PD. The local and central diagnosis for the non-event cases were similar. The plots in Figure 3 indicate a large variability of the agreements over time; thus, it is necessary to account for this temporally variable agreement in any further analysis.

7. Discussion

We have proposed a new method to assess agreement between two time-to-event endpoints, where the two event times are assumed to have a positive probability of being identical. Such an assumption is motivated by type of data we targeted. For example, the head and neck cancer trial data show that the probability of exact agreement of the two event times is at least 30%. We can view our model as a two-pattern mixture model, with $T_c = T_l$ being one pattern. Although, it is not uncommon that the two evaluations are based on measurements which are usually performed at predefined time intervals, there might be a large amount of variation in the two observed event times, so $T_c \neq T_l$, with T_c and T_l continuous random variables, is another pattern (see head and neck cancer trial data, Table 5). Other approaches may be envisioned for modeling same type of data. For example, one can set up the two event times in a discrete failure time setting, by considering a model arising from a single copula model, which is then discretized according to the observation windows and estimate accordingly. However, such an approach is not the scope of this paper. Nonetheless, it is an interesting alternative approach and can be addressed in a separate research.

Our method measures agreement in terms of two event times of being identical at a given time point or both being greater than a given time, assuming that the event had not occurred prior to the given time. The proposed agreement measures are based on the underlying joint distribution of the time-to-event endpoints. We estimate the proposed agreement measures using a mixture parametric distribution based on Weibull distributions, which are general enough to cover a wide range of applications. However, it is of interest to further relax this parametric assumption while consider even more flexible or nonparametric distributions for

both $S(t)$ and $Q(t, s)$ in future work. Sieve estimation based on splines may be useful for carrying out inference, but will increase the computation complexity. Although our agreement measures are defined as the conditional probability of the central event given the local event, this definition can be similarly defined as the reverse conditional probability or the summation of both conditional probabilities, especially when there is no preference to either event.

Our approach is fully likelihood-based and the informative censoring RC is naturally accounted for in the likelihood formulation. So, it has no impact on the validity of our method. As shown in the simulation study, the relative biases of the proposed agreement measures for data without RC and data with RC are very similar. Our method yields robust estimators of the agreement measures in the presence of RC. However, if the censoring is truly informative (depends on potential outcome value), modelling censoring mechanism is necessary to obtain a valid inference.

The proposed method can be directly applied to other commonly used endpoints for oncology clinical trials, for example, time-to-progression and disease free survival ([1]), to assess their agreement between IRC's and investigators' evaluation. In addition, by jointly modeling the IRC's and investigators' assessments, the proposed estimation method can provide less-biased estimates of the parameters in the assumed underlying distribution than those based on IRC's assessment alone. One application is to apply the method within each treatment arm and then make inference about treatment effect with the estimated underlying distribution. High values (e.g., $> 80\%$) of the measure $wAUC(pA)$ in both treatment and control arms provide a high assurance of assessing treatment effects using the investigator's evaluations only, which implies investigator's reading data are robust and reliable to estimate treatment effects; on the contrast, lower values require the need for the IRC's readings in clinical trials. Another application is for event projection. The primary analysis in most oncology trials are event driven (i.e., based on planned number of events). When the primary analysis is based on IRC's evaluation, event modeling based on the estimated underlying distribution of the IRC's data from joint modeling of IRC's and investigator's data provide more accurate projection of future events than that from modeling of IRC's data alone.

The proposed temporal agreement measurements are potentially useful for future trials when central assessment is needed for auditing purposes, since time-to-event outcomes such as PFS, are usually subject to measurement error. For example, it was reported that a discrepancy in PFS times between local and central assessment could range from 23% to 36% [10]. Recently, there has been some development to address this discrepancy [10, 11]. With the development of our temporal agreement measurements, we will be able to not only capture the overall agreement, but also assess the temporal discrepancy between these two assessments; therefore, the differences in the analyses based on local and central assessments will be quantified more precisely. Furthermore, the proposed temporal agreement measurements will also be useful to help design more effective and efficient audit strategies. For instance, if the disagreement occurs most often for long survivals, an audit should then be applied to those subjects. In future work, we will examine how to incorporate

this temporal agreement measure into designing an audit strategy to correct bias due to imprecise local assessments.

References

1. FDA.. Guidance for industry: Clinical trial endpoints for the approval of cancer drugs and biologics. May. 2007 URL <http://www.fda.gov/downloads/drugs/guidancecomplianceregulatoryinformation/guidances/ucm071590.pdf>
2. Amit O, Mannino F, Stone A, Bushnell W, Denne J, Helterbrand J, Burger H. Blinded independent central review of progression in cancer clinical trials: Results from a meta-analysis. *European Journal of Cancer*. 2011; 47(12):1772–1778. doi:<http://dx.doi.org/10.1016/j.ejca.2011.02.013>. URL <http://www.sciencedirect.com/science/article/pii/S0959804911001122>. [PubMed: 21429737]
3. Hougaard, P. *Statistics for Biology and Health*. Springer-Verlag; New York: 2000. Analysis of multivariate survival data.. doi:10.1007/978-1-4612-1304-8. URL <http://dx.doi.org/10.1007/978-1-4612-1304-8>.
4. Liu X, Du Y, Teresi J, Hasin DS. Concordance correlation in the measurements of time to event. *Statistics in Medicine*. 2005; 24(9):1409–1420. doi:10.1002/sim.2004. URL <http://dx.doi.org/10.1002/sim.2004>. [PubMed: 15568183]
5. Guo Y, Manatunga AK. Measuring agreement of multivariate discrete survival times using a modified weighted kappa coefficient. *Biometrics*. 2009; 65(1):125–134. doi:10.1111/j.1541-0420.2008.01054.x. URL <http://dx.doi.org/10.1111/j.1541-0420.2008.01054.x>. [PubMed: 18505422]
6. Guo, Y.; Li, R.; Peng, L.; Manatunga, AK. New agreement measures based on survival processes.; *Biometrics*. 2013. p. 1-9.doi:10.1111/biom.12063. URL <http://dx.doi.org/10.1111/biom.12063>.
7. Joe, H. *Monographs on Statistics and Applied Probability*. Vol. 73. Chapman & Hall; London: 1997. Multivariate models and dependence concepts.
8. Durante, F.; Sempi, C. Copula theory: An introduction. *Copula Theory and Its Applications*. In: Jaworski, P.; Durante, F.; Hrdle, WK.; Rychlik, T., editors. *Lecture Notes in Statistics*. Springer; Berlin Heidelberg: 2010. p. 3-31.doi:10.1007/978-3-642-12465-5_1. URL http://dx.doi.org/10.1007/978-3-642-12465-5_1.
9. Louis TA. Finding the observed information matrix when using the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*. 1982; 44(2):226–233. URL <http://www.jstor.org/stable/2345828>.
10. Dodd L, Korn E, Freidlin B, Robert G, Bhattacharya S. An audit strategy for progression-free survival. *Biometrics*. 2011; 67(3):1092–1099. doi: 101111/j.1541-0420.2010.01539.x. [PubMed: 21210772]
11. Dodd L, Korn E, Freidlin B, Gu W, Abrams J, Bushnell W, Canetta R, Doroshow J, Gray R, Sridhara R. An audit strategy for time-to-event outcomes measured with error: Application to five randomized controlled trials in oncology. *Clinical Trials*. 2013; 10(5):754–760. doi:10.1177/1740774513493973. [PubMed: 23935162]

Figure 1.

Plots of the curves of the agreement measures pA and nA . *Top:* $\rho = 0.5$ fixed and p varying, $p = 0, 0.2, 0.4, 0.6, 0.8, 1$; *bottom:* $p = 0.5$ fixed and ρ varying, $\rho = 0, 0.2, 0.4, 0.6, 0.8$; all other parameters being set to the values as in Tables 1 & 2, and $T_E = 3.0$.

Figure 2.

Plots of the areas, $wAUC(pA)$ and $wAUC(nA)$, under the agreement measure curves pA and nA , respectively, and the Kendall's coefficient of concordance τ .

Estimated curves and pointwise 95% confidence bands of the agreement measures pA and nA for PFS Data

Figure 3.

Plots of the estimated curves of the agreement measures pA and nA , together with the pointwise 95% confidence bands, over the study duration for PFS data. The weighted AUCs are $wAUC(pA) = 0.412$ and $wAUC(nA) = 0.843$.

Table 1

Summary of simulation results from data with RC

p	Par	True	n = 100				n = 200			
			Mean	SD	SE	CP	Mean	SD	SE	CP
0.2	p	0.20	0.201	0.053	0.055	0.96	0.203	0.039	0.039	0.94
	λ_0	0.90	0.978	0.264	0.307	0.91	0.921	0.165	0.187	0.96
	α_0	1.50	1.741	0.520	0.526	0.92	1.596	0.298	0.299	0.94
	λ_c	1.35	1.348	0.220	0.225	0.95	1.359	0.156	0.159	0.95
	α_c	1.20	1.219	0.167	0.164	0.94	1.213	0.112	0.115	0.95
	λ_L	1.40	1.400	0.160	0.166	0.95	1.410	0.124	0.117	0.93
	α_L	1.30	1.332	0.148	0.146	0.94	1.316	0.101	0.101	0.95
	ρ	0.50	0.507	0.130	0.138	0.99	0.499	0.097	0.098	0.95
	$pA(0.5)$	0.161	0.168	0.054	0.058	0.96	0.162	0.038	0.039	0.95
	$pA(1.5)$	0.331	0.312	0.151	0.149	0.99	0.335	0.103	0.101	0.97
0.6	$pA(2.5)$	0.517	0.418	0.314	0.308	0.998	0.479	0.251	0.257	0.99
	$nA(0.5)$	0.771	0.775	0.046	0.044	0.95	0.773	0.032	0.032	0.95
	$nA(1.5)$	0.620	0.611	0.153	0.158	0.96	0.622	0.113	0.114	0.96
	$nA(2.5)$	0.659	0.588	0.284	0.296	0.91	0.619	0.242	0.254	0.89
	p	0.60	0.594	0.063	0.062	0.96	0.597	0.042	0.044	0.97
	λ_0	0.90	0.925	0.114	0.124	0.95	0.911	0.079	0.082	0.95
	α_0	1.50	1.544	0.226	0.216	0.94	1.532	0.151	0.148	0.93
	λ_c	1.35	1.385	0.344	0.340	0.94	1.355	0.228	0.234	0.95
	α_c	1.20	1.277	0.283	0.256	0.92	1.229	0.171	0.169	0.95
	λ_L	1.40	1.438	0.269	0.251	0.92	1.408	0.164	0.175	0.96
	α_L	1.30	1.365	0.222	0.216	0.92	1.332	0.145	0.147	0.95
	ρ	0.50	0.494	0.167	0.211	0.996	0.496	0.131	0.143	0.99
	$pA(0.5)$	0.535	0.529	0.075	0.076	0.96	0.532	0.052	0.054	0.97
	$pA(1.5)$	0.748	0.753	0.118	0.106	0.97	0.749	0.073	0.075	0.96
	$pA(2.5)$	0.865	0.802	0.221	0.182	0.98	0.832	0.148	0.133	0.97
	$nA(0.5)$	0.900	0.900	0.030	0.030	0.96	0.900	0.022	0.022	0.94

p	Par	True	n = 100				n = 200			
			Mean	SD	SE	CP	Mean	SD	SE	CP
0.8	$nA(1.5)$	0.877	0.870	0.094	0.085	0.96	0.874	0.061	0.061	0.96
	$nA(2.5)$	0.913	0.848	0.194	0.164	0.94	0.875	0.138	0.119	0.95
	ρ	0.80	0.794	0.050	0.051	0.96	0.800	0.037	0.037	0.94
	λ_0	0.90	0.914	0.095	0.098	0.94	0.904	0.071	0.070	0.95
	α_0	1.50	1.547	0.189	0.184	0.94	1.516	0.122	0.126	0.94
	λ_C	1.35	1.436	0.478	0.537	0.95	1.377	0.312	0.365	0.96
	α_C	1.20	1.308	0.373	0.421	0.96	1.251	0.242	0.256	0.95
	λ_L	1.40	1.450	0.363	0.386	0.93	1.423	0.245	0.273	0.95
	α_L	1.30	1.424	0.338	0.340	0.93	1.365	0.215	0.222	0.95
	ρ	0.50	0.512	0.179	0.373	1.00	0.493	0.161	0.219	1.00
	$pA(0.50)$	0.754	0.745	0.067	0.069	0.95	0.751	0.051	0.049	0.95
	$pA(1.5)$	0.888	0.891	0.071	0.071	0.98	0.892	0.049	0.050	0.98
	$pA(2.5)$	0.945	0.907	0.151	0.113	0.98	0.926	0.095	0.078	0.97
	$nA(0.5)$	0.953	0.951	0.021	0.022	0.97	0.953	0.014	0.015	0.97
	$nA(1.5)$	0.948	0.941	0.059	0.054	0.97	0.947	0.039	0.037	0.96
	$nA(2.5)$	0.965	0.922	0.136	0.114	0.96	0.941	0.090	0.072	0.96

Table 2

Summary of simulation results from data without RC

p	Par	True	n = 100					n = 200				
			Mean	SD	SE	CP	Mean	SD	SE	CP	SE	CP
0.2	p	0.20	0.202	0.053	0.055	0.95	0.202	0.039	0.038	0.94	0.038	0.94
	λ_0	0.90	0.966	0.300	0.308	0.91	0.920	0.165	0.175	0.95	0.175	0.95
	α_0	1.50	1.715	0.489	0.517	0.91	1.596	0.303	0.304	0.93	0.304	0.93
	λ_C	1.35	1.354	0.169	0.174	0.95	1.362	0.125	0.122	0.93	0.122	0.93
	α_C	1.20	1.215	0.133	0.133	0.94	1.212	0.096	0.093	0.95	0.093	0.95
	λ_L	1.40	1.405	0.154	0.165	0.96	1.409	0.125	0.116	0.93	0.116	0.93
	α_L	1.30	1.329	0.142	0.145	0.95	1.315	0.102	0.101	0.95	0.101	0.95
	ρ	0.50	0.505	0.103	0.103	0.93	0.502	0.071	0.073	0.95	0.073	0.95
	$pA(0.5)$	0.161	0.166	0.053	0.056	0.97	0.162	0.038	0.038	0.94	0.038	0.94
	$pA(1.5)$	0.331	0.320	0.147	0.143	0.99	0.334	0.104	0.099	0.97	0.099	0.97
	$pA(2.5)$	0.517	0.436	0.308	0.298	0.99	0.479	0.252	0.242	0.98	0.242	0.98
	$nA(0.5)$	0.771	0.775	0.045	0.043	0.94	0.773	0.031	0.031	0.95	0.031	0.95
	$nA(1.5)$	0.620	0.616	0.148	0.150	0.94	0.621	0.112	0.109	0.94	0.109	0.94
	$nA(2.5)$	0.659	0.594	0.281	0.287	0.87	0.616	0.242	0.243	0.87	0.243	0.87
0.6	p	0.60	0.594	0.061	0.062	0.96	0.598	0.042	0.045	0.97	0.045	0.97
	λ_0	0.90	0.924	0.113	0.121	0.96	0.910	0.077	0.086	0.95	0.086	0.95
	α_0	1.50	1.539	0.221	0.216	0.95	1.533	0.151	0.155	0.93	0.155	0.93
	λ_C	1.35	1.366	0.254	0.260	0.94	1.363	0.175	0.187	0.96	0.187	0.96
	α_C	1.20	1.262	0.211	0.200	0.93	1.230	0.141	0.140	0.94	0.140	0.94
	λ_L	1.40	1.436	0.253	0.251	0.93	1.410	0.160	0.181	0.97	0.181	0.97
	α_L	1.30	1.365	0.221	0.216	0.93	1.334	0.144	0.148	0.95	0.148	0.95
	ρ	0.50	0.490	0.145	0.152	0.99	0.500	0.102	0.108	0.96	0.108	0.96
	$pA(0.5)$	0.535	0.528	0.074	0.076	0.97	0.532	0.052	0.054	0.97	0.054	0.97
	$pA(1.5)$	0.748	0.752	0.114	0.105	0.97	0.751	0.072	0.074	0.97	0.074	0.97
	$pA(2.5)$	0.865	0.807	0.207	0.180	0.98	0.835	0.140	0.133	0.97	0.133	0.97
	$nA(0.5)$	0.900	0.900	0.029	0.030	0.96	0.900	0.022	0.021	0.95	0.021	0.95

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

p	Par	True	n = 100				n = 200			
			Mean	SD	SE	CP	Mean	SD	SE	CP
0.8	$nA(1.5)$	0.877	0.869	0.092	0.085	0.96	0.874	0.061	0.061	0.96
	$nA(2.5)$	0.913	0.846	0.194	0.165	0.94	0.875	0.137	0.122	0.95
	ρ	0.80	0.791	0.046	0.051	0.96	0.801	0.036	0.035	0.95
	λ_0	0.90	0.915	0.090	0.098	0.95	0.903	0.070	0.067	0.95
	α_0	1.50	1.549	0.194	0.185	0.94	1.516	0.124	0.125	0.94
	λ_C	1.35	1.387	0.387	0.404	0.95	1.372	0.262	0.273	0.95
	α_C	1.20	1.290	0.325	0.303	0.95	1.254	0.210	0.202	0.94
	λ_L	1.40	1.468	0.367	0.395	0.94	1.422	0.242	0.258	0.94
	α_L	1.30	1.398	0.334	0.323	0.93	1.361	0.215	0.218	0.95
	ρ	0.50	0.506	0.163	0.219	1.00	0.493	0.139	0.157	0.99
	$pA(0.5)$	0.754	0.744	0.067	0.069	0.95	0.752	0.051	0.048	0.95
	$pA(1.5)$	0.888	0.891	0.065	0.070	0.98	0.893	0.048	0.049	0.98
	$pA(2.5)$	0.945	0.904	0.151	0.113	0.97	0.927	0.091	0.074	0.96
	$nA(0.5)$	0.953	0.954	0.020	0.020	0.97	0.953	0.014	0.014	0.96
	$nA(1.5)$	0.948	0.941	0.060	0.053	0.97	0.947	0.038	0.036	0.97
	$nA(2.5)$	0.965	0.914	0.151	0.113	0.95	0.941	0.091	0.071	0.96

Table 3

True values, estimates, and biases of the weighted area under the curve of the agreement measures pA and nA , and of Kendall's coefficient of concordance τ , in both data with RC and without RC cases, for maximum follow-up times $T_E = 0.5, 1.5, 2.5, 3.0$, agreement probabilities $p = 0.2, 0.6, 0.8$, sample size $n = 200$, and all the other parameters as in Tables 1 and 2.

p	Agreement Measure	Maximum Follow-up Time	True Value	Data with RC		Data without RC	
				Estimate	Bias	Estimate	Bias
p = 0.2	wAUC(pA)	0.5	0.1224	0.1202	-0.0023	0.1197	-0.0027
		1.5	0.1759	0.1814	0.0055	0.1809	0.0050
		2.5	0.1955	0.1997	0.0042	0.1992	0.0038
		3.0	0.1984	0.2017	0.0034	0.2013	0.0029
	wAUC(nA)	0.5	0.8696	0.8695	-0.0001	0.8695	-0.0001
		1.5	0.7734	0.7721	-0.0013	0.7720	-0.0013
		2.5	0.7602	0.7580	-0.0022	0.7579	-0.0023
		3.0	0.7595	0.7570	-0.0025	0.7569	-0.0026
	τ	0.5	0.2602	0.3108	0.0506	0.4237	0.1634
		1.5	0.3727	0.3457	-0.0270	0.4530	0.0803
		2.5	0.4244	0.3476	-0.0769	0.4536	0.0291
		3.0	0.4328	0.3477	-0.0851	0.4536	0.0208
p = 0.6	wAUC(pA)	0.5	0.4557	0.4520	-0.0037	0.4520	-0.0037
		1.5	0.5615	0.5599	-0.0016	0.5602	-0.0012
		2.5	0.5931	0.5911	-0.0020	0.5917	-0.0014
		3.0	0.5975	0.5952	-0.0024	0.5958	-0.0018
	wAUC(nA)	0.5	0.9379	0.9376	-0.0003	0.9375	-0.0004
		1.5	0.9027	0.9018	-0.0009	0.9021	-0.0006
		2.5	0.9009	0.9005	-0.0004	0.9010	9.5e-05
		3.0	0.9012	0.9009	-0.0004	0.9014	0.0002
	τ	0.5	0.5597	0.5876	0.0280	0.6532	0.0936
		1.5	0.6392	0.6370	-0.0022	0.6925	0.0532
		2.5	0.6797	0.6390	-0.0407	0.6934	0.0138
		3.0	0.6863	0.6391	-0.0472	0.6934	0.0072
p = 0.8	wAUC(pA)	0.5	0.6906	0.6896	-0.0010	0.6900	-0.0006
		1.5	0.7735	0.7719	-0.0015	0.7726	-0.0009
		2.5	0.7954	0.7954	-0.0000	0.7959	0.0006
		3.0	0.7983	0.7984	9.4e-05	0.7990	0.0007
	wAUC(nA)	0.5	0.9695	0.9692	-0.0002	0.9696	9.9e-05
		1.5	0.9543	0.9551	0.0008	0.9554	0.0011
		2.5	0.9543	0.9564	0.0021	0.9566	0.0023
		3.0	0.9546	0.9568	0.0023	0.9570	0.0024

p	Agreement Measure	Maximum Follow-up Time	True Value	Data with RC		Data without RC	
				Estimate	Bias	Estimate	Bias
τ		0.5	0.7596	0.7727	0.0131	0.8038	0.0443
		1.5	0.8068	0.8103	0.0035	0.8359	0.0291
		2.5	0.8315	0.8117	-0.0199	0.8369	0.0054
		3.0	0.8356	0.8118	-0.0238	0.8369	0.0014

Table 4

Summary of simulation results for data from the distribution described at the end of Section 5, for which the model is misspecified. Sample size is $n = 200$.

Maximum Follow-up Time	Agreement Measure							
	True Value	$wAUC(pA)$			True Value	$wAUC(nA)$		
		Estimate	Bias	Rel. Bias (%)		Estimate	Bias	Rel. Bias (%)
0.50	0.6963	0.7176	0.0213	3.06	0.9763	0.9709	-0.0055	-0.56
1.00	0.6824	0.6727	-0.0098	-1.43	0.9596	0.9338	-0.0258	-2.69
1.50	0.6677	0.6511	-0.0167	-2.50	0.9524	0.9063	-0.0461	-4.84
2.00	0.6550	0.6416	-0.0134	-2.04	0.9499	0.8910	-0.0589	-6.20
2.50	0.6452	0.6379	-0.0073	-1.13	0.9492	0.8842	-0.0650	-6.85
3.00	0.6383	0.6367	-0.0016	-0.25	0.9491	0.8817	-0.0674	-7.10

Table 5Head and neck cancer trial data information at times $t = 0.5, 1.5, 2.5$. ($n = 89$)

Average percentage of subjects (%)				
Time		$Y_C = Y_L$	$Y_C < Y_L$	$Y_C > Y_L$
0.5	Both Censored	28.09	0.00	1.12
	Central Event Only	1.12	11.24	1.12
	Local Event Only	6.74	0.00	6.74
	Both Events	31.46	12.36	0.00
1.5	Both Censored	4.49	0.00	1.12
	Central Event Only	1.12	3.37	1.12
	Local Event Only	11.24	0.00	7.87
	Both Events	37.08	24.72	7.86
2.5	Both Censored	2.25	0.00	1.12
	Central Event Only	1.12	2.25	1.12
	Local Event Only	11.24	0.00	4.49
	Both Events	39.33	25.84	11.24

Table 6

Parameter estimates, standard deviations, and 95% confidence intervals of the parameters for the head and neck cancer trial data. (sample size $n = 89$).

Parameter	Estimated Value	Standard Deviation (Louis)	95% CI	
			lower bound	upper bound
p	0.413	0.053	0.314	0.520
λ_0	2.209	0.419	1.522	3.205
α_0	0.998	0.129	0.774	1.286
λ_C	1.284	0.194	0.955	1.727
α_C	1.088	0.140	0.845	1.401
λ_L	1.463	0.139	1.215	1.762
α_L	1.664	0.192	1.327	2.087
ρ	0.612	0.118	0.330	0.794

Estimated values, standard deviations, and 95% confidence intervals of the agreement measures pA and nA . Efrom the second simulation setting without RCestimated values of the weighted area under the agreement measures pA and nA , and of the Kendall's coefficient of concordance τ for the head and neck cancer trial data.

Table 7

Time	Time-Varying Agreement Measure						Agreement Measure over Maximum Follow-Up Time		
	pA			nA					
	Estimated Value	Standard Deviation (Louis)	95% CI		Estimated Value	Standard Deviation (Louis)	95% CI		τ
			lower bound	upper bound			lower bound	upper bound	
0.50	0.320	0.059	0.217	0.445	0.808	0.044	0.708	0.880	0.617
1.00	0.264	0.072	0.147	0.426	0.770	0.078	0.585	0.888	0.587
1.50	0.359	0.193	0.097	0.744	0.826	0.124	0.465	0.963	0.600
2.00	0.599	0.391	0.058	0.973	0.924	0.128	0.254	0.998	0.609
2.36	0.800	0.377	0.038	0.998	0.972	0.074	0.136	0.999	0.613