Published in final edited form as:

Stat Med. 2017 April 30; 36(9): 1363–1382. doi:10.1002/sim.7225.

Sizing clinical trials when comparing bivariate time-to-event outcomes

Tomoyuki Sugimoto^a, Toshimitsu Hamasaki^{b,*†}, Scott Evans^c, and Takashi Sozu^d

^aDepartment of Mathematics and Computer Science, Kagoshima University Graduate School of Science and Technology, Kagoshima, Japan

^bDepartment of Data Science, National Cerebral and Cardiovascular Center, Saita, Japan

^cCenter for Biostatistics in AIDS Research, Department of Biostatistics, Harvard School of Public Health, Boston, MA, USA

^dDepartment of Information and Computer Technology, Faculty of Engineering, Tokyo University of Science, Tokyo, Japan

Abstract

Clinical trials with multiple primary time-to-event outcomes are common. Use of multiple endpoints creates challenges in the evaluation of power and the calculation of sample size during trial design particularly for time-to-event outcomes. We present methods for calculating the power and sample size for randomized superiority clinical trials with two correlated time-to-event outcomes. We do this for independent and dependent censoring for three censoring scenarios: (i) the two events are non-fatal; (ii) when one event is fatal (semi-competing risk); and (iii) when both are fatal (competing risk). We derive the bivariate log-rank test in all three censoring scenarios and investigate the behavior of power and the required sample sizes. Separate evaluations are conducted for two inferential goals, evaluation of whether the test intervention is superior to the control on (1) *all* of the endpoints (multiple co-primary) or (2) *at least one* endpoint (multiple primary).

Keywords

dependent censoring; log-rank test; multiple endpoints; semi-competing risk; time-dependent association

1. Introduction

The use of two time-to-event outcomes as primary endpoints has become common in clinical trials evaluating interventions in many disease areas. For example, co-infection/comorbidity trials may utilize primary endpoints to evaluate multiple comorbidities, for example, a trial

*Correspondence to: Tomoyuki Sugimoto, Department of Mathematics and Computer Science, Kagoshima University Graduate School of Science and Technology, Japan.

†sugimoto@sci.kagoshima-u.ac.jp

Supporting information

Additional supporting information may be found in the online version of this article at the publisher's web site.

evaluating therapies to treat Kaposi's sarcoma in HIV-infected individuals may have the time to Kaposi's sarcoma progression and the time to HIV virologic failure, as primary endpoints. In new anti-cancer drug trials, the most commonly used primary endpoint is overall survival (OS) defined as the time from randomization until death from any cause. However, OS in general requires long follow-up periods after disease progression leading to long and expensive studies. Therefore, in addition to OS, many clinical trials include progression-free survival (PFS) as a primary endpoint, defined as the time from randomization to the first of tumor progression or death. Meanwhile in trials aimed at evaluating treatments to reduce a specific type of mortality, the two endpoints of time to disease-specific mortality and all-cause mortality are primary endpoints. In the first example of co-infection/comorbidity trials, both events are non-fatal (the 'both non-fatal case' where neither event-time is censored by the other event). However, in the oncology example, one event is fatal (death) potentially censoring the other non-fatal event (the 'one fatal case'). Here, we use the term 'fatal' to describe an event that censors future events of interest. This is referred as to 'semi-competing risk problem', first introduced by [1]. In the last example, both events are fatal as each event-time may be censored by the other event (the 'both fatal case').

We developed methods for sizing clinical trials with two primary time-to-event outcomes under a time-dependent correlation structure of three bivariate exponential distributions, where (i) both events are non-fatal [2,3]. In this paper, we discuss the log-rank test-based method using the normal approximation for power and sample size calculations in clinical trials with two time-to-event outcomes, to accommodate two additional situations: (ii) when one event is fatal and (iii) when both are fatal. We also evaluate composite endpoints as a strategy. Hence, six scenarios are evaluated as classified by the three censoring schemes (both non-fatal, one fatal, and both fatal) for the two time-to-event outcomes and whether a composite endpoint is used. Table I illustrates examples of the six scenarios.

We consider two inferential goals for clinical trials with multiple endpoints: (1) 'multiple co-primary endpoints' (where the trial is designed to evaluate if the intervention is superior to the control on *all* of the endpoints) and (2) 'multiple primary endpoints' (or 'alternative primary endpoints', where the trial is designed to evaluate if the intervention is superior to the control on *at least one* endpoint [4–6]). When considering two or more endpoints as co-primary, no adjustment is needed to control the Type I error rate if the hypothesis associated with each endpoint is evaluated at the same significance level as that required for all of the objectives. However, the Type II error rate increases as the number of endpoints being evaluated increases. Thus, design adjustments are needed to maintain the overall power. In contrast, when designing the trial to evaluate an effect on at least one of the endpoints, an adjustment is needed to control the Type I error rate because the Type I error rate increases as the number of endpoints being evaluated increases.

The paper is structured as follows: in Section 2, we describe the dependence measure and censoring schemes, and then discuss the correlation structure of the bivariate log-rank statistic. In Sections 3 and 4, we provide the power for comparing two groups with respect to two time-to-event outcomes as co-primary or multiple primary and describe methods for calculating the sample size. We also investigate the behaviors of power and the required sample sizes with a real example. In Section 5, we summarize our findings.

2. Censoring schemes, dependency, and correlation

2.1. Notation and framework

Consider a randomized clinical trial designed to compare two interventions with total N participants being recruited and randomized. Suppose that $r^{(2)}N$ participants are assigned to the test intervention group and $r^{(1)}N$ participants to the control intervention group ($r^{(1)} + r^{(2)} = 1$). Patients are then followed to evaluate bivariate survival times for two endpoints. Let T_{ik}^* and C_{ik} be the underlying continuous survival time and potential censoring time of the k th primary endpoint for the i th participant ($k = 1, 2; i = 1, \dots, N$). Assume $C_i = C_{i1} = C_{i2}$, because C_{i1} and C_{i2} are usually the same time. Hence, we observe the bivariate time-to-event data of $\{(T_{i1}, T_{i2}, \Delta_{i1}, \Delta_{i2}, g_i)\}_{i=1}^N$, where T_{ik} and Δ_{ik} are the i th observable survival time and right-censoring indicator for the k th primary endpoint, respectively, and g_i is the group index ($j = 2$ if the i th participant belongs to the test, and $j = 1$ otherwise). For example, typically we see $T_{ik} = \min(T_{ik}^*, C_{ik})$ and $\Delta_{ik} = \mathbb{1}(T_{ik}^* \leq C_{ik})$ ($\mathbb{1}(\cdot)$ is the index function). The information of (T_{ik}, Δ_{ik}) is represented by the counting process $\mathcal{N}_{ik}(t) = \mathbb{1}(T_{ik} \leq t, \Delta_{ik} = 1)$ and the at-risk process $\mathcal{Y}_{ik}(t) = \mathbb{1}(T_{ik} > t)$.

Denote the marginal hazard function and its cumulative function for T_{ij}^* in the group j by

$$\lambda_k^{(j)}(t) = \lim_{dt \rightarrow 0} \frac{1}{dt} \Pr(t \leq T_{ik}^* < t + dt | t \leq T_{ik}^*, g_i = j) \quad \text{and} \quad \Lambda_k^{(j)} = \int_0^t \lambda_k^{(j)}(s) ds.$$

Let $\psi_k(t)$ be the hazard ratio (HR) $\lambda_k^{(2)}(t)/\lambda_k^{(1)}(t)$ between the two intervention groups. To test the single hypothesis $H_{0k}: \psi_k(t) = 1$ for all t restricted to the k th endpoint ($k = 1, 2$), the standardized log-rank statistic

$$Z_k^{(0)} = U_k(\tau) / \sqrt{\hat{V}_{kk}^{(0)}(\tau)} \quad (1)$$

can be applied to the univariate data set of $\{(T_{ik}, \Delta_{ik}, g_i)\}_{i=1}^N$, where τ is the maximum observed follow-up time, $U_k(t)$ is the log-rank process

$$U_k(t) = \sqrt{N} \int_0^t \hat{H}_k(s) \left\{ d\hat{\Lambda}_k^{(2)}(s) - d\hat{\Lambda}_k^{(1)}(s) \right\}, \quad k=1, 2,$$

$\hat{\Lambda}_k^{(j)}(t) = \int_0^t d\mathcal{N}_k^{(j)}(s) / \mathcal{Y}_k^{(j)}(s)$ is the Nelson–Aalen estimator of $\Lambda_k^{(j)}(t)$, $\hat{V}_{kk}^{(0)}(t)$ is the conditional variance of $U_k(t)$ under the null hypothesis H_{0k} ,

$$\hat{V}_{kk}^{(0)}(t) = N \int_0^t \hat{H}_k(s)^2 \left\{ 1 - \frac{d\mathcal{N}_k(s) - 1}{\mathcal{Y}_k(s) - 1} \right\} \left\{ \frac{d\hat{\Lambda}_k^{(1)}(s)}{\mathcal{Y}_k^{(2)}(s)} + \frac{d\hat{\Lambda}_k^{(2)}(s)}{\mathcal{Y}_k^{(1)}(s)} \right\},$$

where

$\hat{H}_k(t) = N^{-1} \mathcal{Y}_k^{(1)}(t) \mathcal{Y}_k^{(2)}(t) / \mathcal{Y}_k(t)$, $\mathcal{Y}_k(t) = \mathcal{Y}_k^{(1)}(t) + \mathcal{Y}_k^{(2)}(t)$, $\mathcal{N}_k(t) = \mathcal{N}_k^{(1)}(t) + \mathcal{N}_k^{(2)}(t)$, $\mathcal{Y}_k^{(j)}(t) = \sum_{i=1}^N (g_i = j) \mathcal{I}_{ik}(t)$ and $\mathcal{N}_k^{(j)}(t) = \sum_{i=1}^N (g_i = j) \mathcal{N}_{ik}(t)$ are the at-risk and counting processes for the k th endpoint of individuals belonging to the group j .

2.2. Censoring schemes and dependence measures for two time-to-event outcomes

In the trial where the bivariate time-to-event data of $\{(T_{i1}, T_{i2}, \Delta_{i1}, \Delta_{i2}, g_i)\}_{i=1}^N$ is observed, various situations are considered. First, consider the censoring scheme. [3] discuss the simplest case where both events are non-fatal. We consider the two other situations, that is, where one event is fatal but another is non-fatal (one fatal), and when both events are fatal (both fatal). We also briefly describe the measure of dependence between the two time-to-event outcomes.

Examples of the three censoring scenarios

- i. **Both non-fatal outcomes:** In HIV clinical trials, if T_{i1}^* is the time to infant HIV infection and T_{i2}^* is the time to infant Hepatitis B infection, neither event-time is censored by the other event. If the subjects do not experience both events, then the events are censored at the same time (e.g., by the end of the study or patient drop-out) in the end of follow-up period. So that we have $T_{i1} = \min(T_{i1}^*, C_i)$ and $T_{i2} = \min(T_{i2}^*, C_i)$.
- ii. **One fatal outcome:** In oncology trials, if T_{i1}^* is the time-to-progression (TTP: defined as the time from randomization until objective tumor progression and does not include deaths) and T_{i2}^* is the OS (time to all cause death), one endpoint (TTP) is censored by the other endpoint (OS) being completely observed. So that, we have $T_{i1} = \min(T_{i1}^*, T_{i2}^*, C_i)$ and $T_{i2} = \min(T_{i2}^*, C_i)$. As T_{i2} is a competing risk for T_{i1} but T_{i1} is not for T_{i2} ($T_{i1} < T_{i2}$), this situation describes a semi-competing risk discussed by [1] (see [7] for further discussions). If there is a non-zero correlation between the endpoints (i.e., dependent censoring), the situation requires modifying the standard log-rank test to account for the dependent censoring. We may be able to avoid this problems by creating a composite endpoint.
- iii. **Both fatal outcomes:** In trials aimed at reducing a specific type of mortality, if T_{i1}^* is the time to disease-specific mortality and T_{i2}^* is the time to other-cause mortalities, each event may be censored by the other event. This situation is called a competing risk. Here, we have $T_{i1} = \min(T_{i1}^*, T_{i2}^*, C_i)$ and $T_{i2} = \min(T_{i1}^*, T_{i2}^*, C_i)$.

Definition for the composite endpoint and handling censoring—We provide the definition for the composite endpoint in the scenarios with two time-to-event outcomes. See Table I for examples of the composite endpoints, such as PFS composing (combining) the TTP and OS or major adverse cardiac events, where the difference between composing and not composing the endpoints is based on handling rule for censoring in each endpoint. That is, we can consider handling censoring in two ways in our time-to-event context. The first way is as usual to use $\Delta_{ik}=(T_{ik}=T_{ik}^*)$, $k=1, 2$, which is the non-composite setting in the context of this paper. The second one is to handle the censoring indicators by

$\Delta_{i1}=(T_{i1}=\min(T_{i1}^*, T_{i2}^*))$ and $\Delta_{i2}=(T_{i2}=T_{i2}^*)$, which is the definition of the composite setting in this paper. For an example, consider the situation where T_{i1}^* and T_{i2}^* are the TTP and OS endpoints, respectively. Then, (T_{i1}, δ_{i1}) is the i th observation for the TTP endpoint in the former handling censoring (the TTP outcome), while (T_{i1}, δ_{i1}) is that for the PFS endpoint in the latter (the PFS outcome). Note that the difference between the TTP and PFS outcomes is brought by handling of censoring in this paper, but T_{i1} is the observable time with the same length and notation in both TTP and PFS endpoints. Alternatively, because one usually define the PFS endpoint as $\tilde{T}_{i1}^*=\min(T_{i1}^*, T_{i2}^*)$, the former handling of censoring for the composite endpoint $\tilde{T}_{i1}^*$ leads to the PFS outcome. However, to avoid confusion in derivations, we discuss the case of composite setting based on the handling rule for censoring without notations to define the composite endpoint such as $\tilde{T}_{i1}^*$. Also, in this paper, the observations for the 1st endpoint, (T_{i1}, δ_{i1}) , are used in either non-composite or composite setting. On the other hand, those for the second endpoint, (T_{i2}, δ_{i2}) , are consistently considered as the non-composite setting.

Dependence measure—Two times T_{i1}^* and T_{i2}^* may be correlated. We consider a correlation structure between T_{i1}^* and T_{i2}^* . Let $S^{(j)}(t, s)=\Pr(t < T_{i1}^*, s < T_{i2}^* | g_i=j)$ and $S_k^{(j)}(t)=\Pr(t < T_{ik}^* | g_i=j)$ ($k=1, 2$) ($k=1, 2$) be the joint survival and marginal survival functions for the bivariate survival data $\{(T_{i1}^*, T_{i2}^*)\}_{i=1}^n$ in the group j , respectively. We consider the correlation between the two cumulative hazard variates [8] defined by

$$\rho^{(j)}=\text{corr} \left[\Lambda_1^{(j)}(T_{i1}^*), \Lambda_2^{(j)}(T_{i2}^*) \right] = \int_0^\infty \int_0^\infty S^{(j)}(t, s) d\Lambda_1^{(j)}(t) d\Lambda_2^{(j)}(s) - 1.$$

If the marginal of the bivariate survival data are exponential, $\rho^{(j)}$ is the same as the correlation coefficient of the raw data $\{(T_{i1}^*, T_{i2}^*)\}_{i=1}^n$ [3]. In order to generate the joint survival function $S^{(j)}(t, s)$ from the $S_1^{(j)}(t)$ and $S_2^{(j)}(s)$, we prepare a copula function $\mathcal{C}(\cdot)$, which gives

$$S^{(j)}(t, s)=\mathcal{C}(S_1^{(j)}(t), S_2^{(j)}(s); \theta^{(j)}),$$

where the association parameter $\theta^{(j)}$ included in $\mathcal{C}(\cdot)$ is a one-to-one function of $\rho^{(j)}$.

2.3. Bivariate structure of two log-rank test statistics

Consider applying the log-rank statistic (1) to the data $\{(T_{ik}, \Delta_{ik}, g_i)\}_{i=1}^N$ for each endpoint ($k = 1, 2$) extracted from the bivariate time-to-event data $\{(T_{i1}, T_{i2}, \Delta_{i1}, \Delta_{i2}, g_i)\}_{i=1}^N$

obtained under the aforementioned situation. Then, the pair $(Z_1^{(0)}, Z_2^{(0)})^T$ of the standardized log-rank statistics is approximately bivariate normally distributed with mean vector $\sqrt{N}\delta$ and variance–covariance matrix Σ when N is sufficiently large (see [3]), where

$$\delta = \begin{pmatrix} \delta_1 \\ \delta_2 \end{pmatrix} = \begin{pmatrix} \frac{\mu_1(\tau)}{\sqrt{V_{11}^{(0)}(\tau)}} \\ \frac{\mu_2(\tau)}{\sqrt{V_{22}^{(0)}(\tau)}} \end{pmatrix} \text{ and } \Sigma = \begin{pmatrix} \frac{V_{11}(\tau)}{V_{11}^{(0)}(\tau)} & \frac{V_{12}(\tau)}{\sqrt{V_{11}^{(0)}(\tau)V_{22}^{(0)}(\tau)}} \\ \frac{V_{12}(\tau)}{\sqrt{V_{11}^{(0)}(\tau)V_{22}^{(0)}(\tau)}} & \frac{V_{22}(\tau)}{V_{22}^{(0)}(\tau)} \end{pmatrix},$$

$\mu_k(t)$ and $V_{kk}^{(0)}(t)$ are asymptotic forms of $N^{-1/2}U_k(t)$ and $\hat{V}_{kk}^{(0)}(t)$, respectively, and $V_{kk}(t)$, $k = 1, 2$ and $V_{12}(t)$ are the asymptotic variances and covariance of $U_1(t)$ and $U_2(t)$, respectively. These elements can be written as, in all censoring scenarios with/without the composite setting,

$$\begin{aligned} \mu_k(\tau) &= \int_0^\tau H_k(t) \left\{ d\tilde{\Lambda}_k^{(2)}(t) - d\tilde{\Lambda}_k^{(1)}(t) \right\}, \quad k=1, 2 \\ V_{kk}^{(0)}(\tau) &= \int_0^\tau H_k(t)^2 \left\{ d\tilde{\Lambda}_k^{(1)}(t)/r^{(2)}y_k^{(2)}(t) + d\tilde{\Lambda}_k^{(2)}(t)/r^{(1)}y_k^{(1)}(t) \right\}, \quad k=1, 2 \\ V_{kk}(\tau) &= \int_0^\tau H_k(t)^2 \left\{ d\tilde{\Lambda}_k^{(1)}(t)/r^{(1)}y_k^{(1)}(t) + d\tilde{\Lambda}_k^{(2)}(t)/r^{(2)}y_k^{(2)}(t) \right\}, \quad k=1, 2 \\ V_{12}(\tau) &= \int_0^\tau \int_0^\tau H_1(t)H_2(s)G(t \vee s) \left\{ \frac{dA^{(1)}(t,s)}{r^{(1)}y_1^{(1)}(t)y_2^{(1)}(s)} + \frac{dA^{(2)}(t,s)}{r^{(2)}y_1^{(2)}(t)y_2^{(2)}(s)} \right\}, \end{aligned} \quad (2)$$

where $H_k(t)$ is the asymptotic form of $\hat{H}_k(t)$,

$$y_k^{(j)}(t) = E[\mathcal{Y}_{ik}(t)|g_i=j], \quad d\tilde{\Lambda}_k^{(j)}(t) = E[d\mathcal{N}_{ik}(t)|\mathcal{Y}_{ik}(t)=1, g_i=j],$$

$$G(t \vee s) dA^{(j)}(t, s) = E[dM_{i1}(t)dM_{i2}(s)|g_i=j], \quad j=1, 2$$

is a covariance function for a martingale variation of $dM_{ik}(t) = d\mathcal{N}_{ik}(t) - \mathcal{Y}_{ik}(t)d\tilde{\Lambda}_k^{(j)}(t)$, $G(t)$ is the survival function of censoring time C_i and $t \vee s$ represents $\max(t, s)$. However, note that the forms of $H_k(t)$, $y_k^{(j)}(t)$, $d\tilde{\Lambda}_k^{(j)}(t)$, and $dA^{(j)}(t, s)$ are different among the censoring scenarios as provided thereafter. Some details of the derivations are in Appendix A. In advance, let $\Lambda^{(j)}(t, s) = -\log S^{(j)}(t, s)$ be the joint cumulative hazard function in group j . Using this notation, note that we can write $d\Lambda_1^{(j)}(t) = \Lambda^{(j)}(dt, 0)$ and $d\Lambda_2^{(j)}(s) = \Lambda^{(j)}(0, ds)$.

- (i) **Both non-fatal outcomes:** In this situation, the setting of the censoring indicators is consistently $\Delta_{ik} = (T_{ik}^* = T_{ik}) = (T_{ik}^* \leq C_i)$. The forms of $H_k(t)$,

$y_k^{(j)}(t)$, $d\tilde{\Lambda}_k^{(j)}(t)$ and $dA^{(j)}(t, s)$ are given by [3]. See Table II for the details. In particular, $\tilde{\Lambda}_k^{(j)}(t) = \Lambda_k^{(j)}(t)$ and

$$dA^{(j)}(t, s) = S^{(j)}(dt, ds) + S^{(j)}(t, ds)d\Lambda_1^{(j)}(t) + S^{(j)}(dt, s)d\Lambda_2^{(j)}(s) + S^{(j)}(t, s)d\Lambda_1^{(j)}(t)d\Lambda_2^{(j)}(s).$$

- (ii) **One fatal outcome:** Assume that T_{i1}^* is the time to non-fatal event while T_{i2}^* is the time to fatal event, so that $T_{i1} = \min(T_{i1}^*, T_{i2}^*, C_i)$ and $T_{i2} = \min(T_{i2}^*, C_i)$. For example, in oncology trials, the TTP and OS endpoints are defined by non-fatal event time T_{i1}^* and fatal one T_{i2}^* , respectively, while the PFS endpoint is defined by $\min(T_{i1}^*, T_{i2}^*)$ as the composite. The log-rank test is applied to survival outcomes obtained using their endpoints. The applications to two sets of bivariate data defined by the TTP and OS endpoints and by the PFS and OS endpoints are discussed in the following non-composite and composite settings, respectively. See Appendix A for the details provided below.

Non-composite case—Consider the non-composite setting, that is, set $\Delta_{ik} = (T_{ik}^* = T_{ik})$, $k = 1, 2$ for the censoring indicators, where non-fatal T_{i1}^* is censored by observing fatal T_{i2}^* earlier than T_{i1}^* . The forms of $H_k(t)$, $y_k^{(j)}(t)$, and $d\tilde{\Lambda}_k^{(j)}(t)$, $j = 1, 2$ are that, for the non-fatal endpoint,

$$H_1(t) = G(t) \frac{r^{(1)}r^{(2)}S^{(1)}(t, t)S^{(2)}(t, t)}{r^{(1)}S^{(1)}(t, t) + r^{(2)}S^{(2)}(t, t)}, \quad y_1^{(j)}(t) = G(t)S^{(j)}(t, t) \quad \text{and} \quad d\tilde{\Lambda}_1^{(j)}(t) = \Lambda^{(j)}(dt, t),$$

depending on the correlation between T_{i1}^* and T_{i2}^* while, for the fatal endpoint,

$$H_2(t) = G(t) \frac{r^{(1)}r^{(2)}S_2^{(1)}(t)S_2^{(2)}(t)}{r^{(1)}S_2^{(1)}(t) + r^{(2)}S_2^{(2)}(t)}, \quad y_2^{(j)}(t) = G(t)S_2^{(j)}(t) \quad \text{and} \quad d\tilde{\Lambda}_2^{(j)}(t) = d\Lambda_2^{(j)}(t)$$

being the same as those of situation (i). Also, $dA^{(j)}(t, s)$ related to the covariance of the martingale variation is

$$dA^{(j)}(t, s) = (t \leq s)S^{(j)}(dt, ds) + (t \leq s)S^{(j)}(t, ds)\Lambda^{(j)}(dt, t) + S^{(j)}(dt, t \vee s)d\Lambda_2^{(j)}(s) + S^{(j)}(t, t \vee s)\Lambda^{(j)}(dt, t)d\Lambda_2^{(j)}(s). \quad (3a)$$

Composite case—In the composite setting, set $\Delta_{i1} = (T_{i1} = \min(T_{i1}^*, T_{i2}^*))$ and $\Delta_{i2} = (T_{i2} = T_{i2}^*)$ for the censoring indicators. The forms of $H_k(t)$, $y_k^{(j)}(t)$ and $d\tilde{\Lambda}_2^{(j)}(t)$ are the same as those in the non-composite case, while $d\tilde{\Lambda}_1^{(j)}(t)$ and $dA^{(j)}(t, s)$ are changed from the

non-composite case, where the intensity information on the second endpoint (fatal event) is added into them ($k = 1, 2, j = 1, 2$). That is, in this case, we have

$$d\tilde{\Lambda}_1^{(j)}(t) = \Lambda^{(j)}(dt, t) + \Lambda^{(j)}(t, dt)$$

and

$$\begin{aligned} dA^{(j)}(t, s) = & (t \leq s)S^{(j)}(dt, ds) - (t=s)S^{(j)}(t, ds) \\ & + (t \leq s)S^{(j)}(t, ds)\{\Lambda^{(j)}(dt, t) + \Lambda^{(j)}(t, dt)\} \\ & + \{S^{(j)}(dt, t \vee s) + (t \leq s)S^{(j)}(dt, t)\}d\Lambda_2^{(j)}(s) \\ & + S^{(j)}(t, t \vee s)\{\Lambda^{(j)}(dt, t) + \Lambda^{(j)}(t, dt)\}d\Lambda_2^{(j)}(s). \end{aligned} \quad (3b)$$

- (iii) **Both fatal outcomes:** Both T_{i1}^* and T_{i2}^* are the times to fatal events, so that we observe $T_{i1} = \min(T_{i1}^*, T_{i2}^*, C_i)$ and $T_{i2} = \min(T_{i1}^*, T_{i2}^*, C_i)$. See Appendix A for the details provided in the subsequent section.

Non-composite case—Assume the non-composite setting, that is, $\Delta_{ik} = (T_{ik}^* = T_{ik})$, $k = 1, 2$. The forms of $H_k(t)$, $y_k^{(j)}(t)$, and $d\tilde{\Lambda}_k^{(j)}(t)$, $k = 1, 2, j = 1, 2$ are similar to those for the 2nd endpoint in situation (ii), because, if either of T_{i1} or T_{i2} is completely observed, the other is always censored. In this case, we can derive

$$\begin{aligned} H_k(t) &= G(t) \frac{r^{(1)}r^{(2)}S^{(1)}(t, t)S^{(2)}(t, t)}{r^{(1)}S^{(1)}(t, t) + r^{(2)}S^{(2)}(t, t)}, \quad y_k^{(j)}(t) = G(t)S^{(j)}(t, t), \quad k=1, 2 \\ d\tilde{\Lambda}_1^{(j)}(t) &= \Lambda^{(j)}(dt, t), \quad d\tilde{\Lambda}_2^{(j)}(t) = \Lambda^{(j)}(t, dt) \end{aligned}$$

and

$$dA^{(j)}(t, s) = 0. \quad (4a)$$

Composite case—Here, set $\Delta_{i1} = (T_{i1} = \min(T_{i1}^*, T_{i2}^*))$ and $\Delta_{i2} = (T_{i2} = T_{i2}^*)$ for the composite setting. In this case, only $d\tilde{\Lambda}_1^{(j)}(t)$ and $dA^{(j)}(t, s)$, $j = 1, 2$ are changed from the non-composite version. Hence, we have

$$d\tilde{\Lambda}_1^{(j)}(t) = \Lambda^{(j)}(dt, t) + \Lambda^{(j)}(t, dt)$$

and

$$\begin{aligned} dA^{(j)}(t, s) = & -(t=s)S^{(j)}(t \vee s, ds) + (t \leq s)S^{(j)}(s, ds)\Lambda^{(j)}(t, dt) \\ & + (t \geq s)S^{(j)}(t, dt)\Lambda^{(j)}(s, ds) + S^{(j)}(t \vee s, t \vee s)\Lambda^{(j)}(t, dt)\Lambda^{(j)}(s, ds). \end{aligned} \quad (4b)$$

The forms of $H_k(t)$, $y_k^{(j)}(t)$, and $d\tilde{\Lambda}_2^{(j)}(t)$, $k, j = 1, 2$ are the same as those in the non-composite setting.

Table II summarizes the statistics regarding $y_k^{(j)}(t)$, $d\tilde{\Lambda}_k^{(j)}(t)$, $H_k(t)$, and $dA^{(j)}(t, s)$ among the three censoring scenarios under the non-composite setting. The table clearly shows how these statistics change among the three scenarios. For all scenarios, the power is given by

$$1 - \beta = \Pr \left[(Z_1^{(0)}, Z_2^{(0)}) \in \mathcal{Z}^{(0)} \right] \approx \iint_{(z_1, z_2) \in \mathcal{Z}^{(0)}} \phi(z_1, z_2; \sqrt{N}\boldsymbol{\delta}, \boldsymbol{\Sigma}) dz_1 dz_2 \quad (5)$$

because the hypothesis is rejected if the bivariate statistic $(Z_1^{(0)}, Z_2^{(0)})^T$ takes the realized values on the area $\mathcal{Z}^{(0)}$, where $\phi(z_1, z_2; \sqrt{N}\boldsymbol{\delta}, \boldsymbol{\Sigma})$ is the bivariate normal density with mean vector $\sqrt{N}\boldsymbol{\delta}$ and variance–covariance matrix $\boldsymbol{\Sigma}$.

3. Co-primary endpoints

3.1. Hypothesis testing, power, and sample sizes

We are interested in testing the hypotheses on the HRs to evaluate a joint reduction of occurrence of events over time on both outcomes, that is,

$H_0^{\text{cp}} = \{H_{01} : \psi_1(t) = 1 \text{ or } H_{02} : \psi_2(t) = 1 \text{ for all } t\}$ versus

$H_1^{\text{cp}} = \{H_{11} : \psi_1(t) < 1 \text{ and } H_{12} : \psi_2(t) < 1 \text{ at some } t\}$. In all of the censoring scenarios, using the

two log-rank statistics, $Z_k^{(0)} = U_k(\tau) / \sqrt{\hat{V}_{kk}^{(0)}(\tau)}$, $k = 1, 2$ for this hypothesis testing, we are able to

$$\text{reject the null hypothesis } H_0^{\text{cp}} \text{ if and only if } Z_1^{(0)} < -z_\alpha \text{ and } Z_2^{(0)} < -z_\alpha \quad (6)$$

at the prespecified significance level of α , where z_α is the $100(1 - \alpha)$ percentile of $N(0, 1)$.

The overall Type I error associated with the null hypothesis H_0^{cp} is controlled by the maximum of the marginal Type I errors [5]. This means that, to investigate whether the overall Type I error is larger than the nominal level, it is enough to investigate whether the marginal ones are larger than the nominal level [2]. The behaviors of the Type I error for the univariate log-rank test are well known [9–11]. The one-sided test procedure (6) based on the asymptotic normality may inflate under some situations such as small sample size and/or unbalanced design (in particular, with $r^{(1)} > 0.5$). We can improve the precision by correcting the critical value based on the sample size N and the allocation rate $r^{(1)}$ to control the

marginal Type I errors. However, the overall Type I error on H_0^{cp} is, for example, the product of marginal ones for independent endpoints, and usually quite smaller than the maximum of marginal ones as long as $\rho^{(j)}$ is not so high, so that the inflation problem seen in one-sided log-rank test may be moderately reduced in multiple co-primary problem.

In the procedure (6), the rejection region of H_0^{cp} is $\{(Z_1^{(0)}, Z_2^{(0)}) : Z_1^{(0)} < -z_\alpha \text{ and } Z_2^{(0)} < -z_\alpha\}$. In all of the scenarios, therefore, we have the power function for the joint reduction in both time-to-event outcomes given as

$$1-\beta = \Pr \left[\bigcap_{k=1}^2 \{Z_k^{(0)} < -z_\alpha\} | H_1^{\text{cp}} \right]$$

This overall power is referred to as ‘complete power’ [12] or ‘conjunctive power’ [13], which is simply calculated using the cumulative distribution function of the bivariate normal distribution. The power can be approximately calculated (under large samples) by

$$1-\beta \approx \int_{-\infty}^{z_\alpha} \int_{-\infty}^{-z_\alpha} \phi \left(z_1, z_2; \sqrt{N} \boldsymbol{\delta}, \boldsymbol{\Sigma} \right) dz_1 dz_2 \quad (7)$$

Let N_{cp} be the minimum of total sample size N required for testing H_0^{cp} against H_1^{cp} . Thus, N_{cp} is the smallest integer not less than satisfying the power (7) and given by

$$N_{\text{cp}} = \lceil \left(K_\beta + z_\alpha \sigma_2^{(0)} / \sigma_2 \right)^2 / \delta_2^2 \rceil \quad (8)$$

where $\lceil x \rceil$ is the smallest integer not less than x , $\sigma_k^{(0)}$ and $\sigma_k (k=1, 2)$ are the standard deviations $\sqrt{V_{kk}^{(0)}(\tau)}$ and $\sqrt{V_{kk}(\tau)}$, respectively, K_β is the solution of the integral equation

$$1-\beta = \int_{-\infty}^{\frac{\delta_1}{\delta_2} K_\beta + z_\alpha \left\{ \frac{\delta_1}{\delta_2} \frac{\sigma_2^{(0)}}{\sigma_2} - \frac{\sigma_1^{(0)}}{\sigma_1} \right\}} \int_{-\infty}^{K_\beta} \phi(z_1, z_2; \mathbf{0}, \mathbf{R}) dz_1 dz_2$$

and $\mathbf{R}$ is the following correlation matrix

$$\mathbf{R} = \begin{pmatrix} 1 & V_{12}(\tau) / \sqrt{V_{11}(\tau)V_{22}(\tau)} \\ V_{12}(\tau) / \sqrt{V_{11}(\tau)V_{22}(\tau)} & 1 \end{pmatrix}.$$

A grid search to find a value of N_{cp} often requires considerable computing time. The search proceeds by gradually increasing N_{cp} until the power (7) exceeds the desired power.

Alternative methods to reduce the computational time are the Newton–Raphson algorithm for finding K_{β} in [14] or the basic linear interpolation algorithm in [2].

One may expect to calculate the required number of participants from that of events such as the Reference [15] procedure in univariate data. However, we encounter difficulty in finding such a procedure in the bivariate case [3]. One of the causes is that the relationship between numbers of events and participants is more complicated than the univariate case, because the patterns of observations (censored or not) increase twofold. Another is that it is difficult to let the difference between the correlation models (such as early or late dependency) on a restricted time-interval reflect using both numbers of events without censoring, because both numbers of events are usually considered under the uncensored model (or after long enough follow-up). Even so, the required numbers of events are useful to monitor a trial and they can be obtained using $D_{ij} = N_{cp} \times P_{ij}$, where P_{ij} is the proportion of each observation pattern, $P_{\delta I \check{Z} d \delta I \check{Z} \check{e}} = \Pr(\Delta_i^{(1)} = \delta I \check{Z} d, \Delta_i^{(2)} = \delta I \check{Z} \check{e})$, D_{ij} is the expected number of events in each case where the first endpoint is observed ($i = 1$) or censored ($i = 0$) and the second is observed ($j = 1$) or not ($j = 0$). See Appendix B for details for calculation of P_{ij} under the three censoring schemes.

3.2. Behavior of the sample size

We investigate the behavior of the sample size and power for detecting the joint reduction H_1^{cp} in bivariate time-to-event data under the three censoring schemes with the two time-dependent association structures, where the time-dependent association structures are asymmetric late (tail) dependency generated by the Clayton copula [16] and early (tail) dependency by the Gumbel copula [17], which have been widely used in practice. We generate bivariate time-to-event data by supposing that the marginals of T_{i1}^* and T_{i2}^* are exponential, and that $C_i = U(0, \tau_a) + \tau_f$ (hence, $\tau = \tau_a + \tau_f$), where τ_a and τ_f are the lengths of the entry period to the trial and follow-up period, respectively, and $U(0, \tau_a)$ denotes a uniform random number on $(0, \tau_a)$. The target power $1 - \beta = 0.8$, the significance level $\alpha = 0.025$, $\tau_a = 2$, and $\tau_f = 3$ are used, and all empirical powers are computed by Monte Carlo trials with 100,000 replications throughout.

Let N_{sim} be the simulation-based sample size required for testing H_0^{cp} , and let N_k denote the minimum of total sample size required to test the single hypothesis H_{0k} for the k th endpoint. To avoid confusion in N_{cp} and N_1 , we write them as $N_{cp}^{(c)}$ and $N_1^{(c)}$ if the first endpoint is replaced by the composite of the first and the second ones (composite setting), and as $N_{cp}^{(nc)}$ and $N_1^{(nc)}$ otherwise (non-composite setting). Let $\tilde{P}_{cp}$ denote the empirical power (%) for detecting H_1^{cp} when the total number of participants is N_{cp} designed using the formula (8). See Section B.1 of Supporting Information for results other than ones provided later.

Formula performance and sample size behavior: one fatal case—Focusing on the one fatal case scenario, suppose that T_{i1}^* and T_{i2}^* are the TTP and OS endpoints, respectively. We evaluate the performance of the sample size formula (8) for practicality by comparing it with alternative solutions sizing based on univariate versions (N_{TTP} , N_{PFS} , and

N_{OS}) and simulation (N_{sim}) under the common α and β , where

$N_{TTP} (=N_1^{(nc)})$, $N_{PFS} (=N_1^{(c)})$, and $N_{OS} (=N_2)$ are N_k when the k th endpoint is the TTP, PFS, and OS. In the one fatal case, the bivariate survival data $(T_{i1}, \dots)$ and $(T_{i2}, \dots)$ are of the TTP and OS outcomes under the non-composite setting, but they are of the PFS and OS outcomes under the composite setting.

Table III displays the required total sample sizes N_{cp} , N_{sim} , N_1 , and N_2 with the empirical power $\tilde{P}_{cp}$ under $\psi_{TTP}^{-1}=1.7$, $S_{TTP}^{(1)}(\tau)=0.3$ and $S_{OS}^{(1)}(\tau)=0.4$, when ψ_{OS}^{-1} and $\rho^{(k)}$ vary for the combinations of $\psi_{OS}^{-1}=1.3, 1.5, 1.7$ and $\rho^{(1)} = \rho^{(2)} = 0, 0.3, 0.5$ and 0.8 . Note that $\psi_{TTP} (= \psi_1)$ and $\psi_{OS} (= \psi_2)$ are the HRs for the TTP and OS, respectively. Similarly,

$S_{TTP}^{(1)}(\tau) (=S_1^{(1)}(\tau))$ and $S_{OS}^{(1)}(\tau) (=S_2^{(1)}(\tau))$ are the τ -time survival rates for the TTP and OS.

The sample sizes N_{cp} calculated from the formula (8) usually provide slightly conservative results compared with N_{sim} , and the corresponding empirical powers $\tilde{P}_{cp}$ are preferable, that is, $\tilde{P}_{cp}$ are slightly larger than the target powers (although $\tilde{P}_{cp}$ tends to be further away from the target as the HR ψ_2^{-1} increases larger than 1). Times to compute N_{sim} are usually much longer than ones for N_{cp} , which becomes greater as the effect size is smaller. Hence, the formula (8) reduces the cost greatly, regardless of the effect size, and it is also useful as an initial value to search N_{sim} . Also, $N_{cp}^{(c)}$ (N_{cp} under the composite setting) is smaller than $N_{cp}^{(nc)}$ (N_{cp} under the non-composite setting) in all cases. As the absolute value of $\log \psi_{OS}^{-1}$ decreases below $|\log \psi_{TTP}^{-1}|$, the value of N_{cp} approaches N_{OS} and diverges from N_{PFS} and N_{TTP} . If both HRs are approximately equal, $\psi_{TTP}^{-1} \approx \psi_{OS}^{-1}$, then N_{cp} is slightly larger than the $\max(N_1, N_2)$ (the ratios of $N_{cp}/\max(N_1, N_2)$ are at most about 1.29 in Table III). Further, when comparing late dependency (Clayton copula) and early one (Gumbel copula), if ψ_{TTP}^{-1} is relatively close to ψ_{OS}^{-1} , $N_{cp}^{(nc)}$ increases proportionally to $\rho^{(j)}$ in the late dependency, while $N_{cp}^{(c)}$ decreases or does not change as $\rho^{(j)}$ varies in the early dependency. Similar tendencies but more moderate variation is observed for $N_{cp}^{(c)}$. That is, a high value of $\rho^{(j)}$ makes the two log-rank statistics correlated, but higher $\rho^{(j)}$ increases the censoring rate, so that the higher $\rho^{(j)}$ does not contribute very much to the reduction of N_{cp} compared with the sample size when $\rho^{(j)} = 0$. Hence, N_{cp} increases in the late dependency, because many observations are censored before the dependent effect works.

Sample size behavior under correlated both fatal outcomes—Consider the both fatal outcomes scenario such that T_{i1}^* is the time to disease-specific mortality and T_{i2}^* is the time to other-cause mortalities. If there are such fatal endpoints, it is often assumed that their endpoints are uncorrelated for purposes of ease. But the assumption of no correlation is often unjustified scientifically. We investigate how the required sample sizes

$N_{cp}^{(c)}$, $N_{cp}^{(nc)}$, $N_1^{(c)}$, $N_1^{(nc)}$, and N_2 behave when the fatal endpoints are correlated. We compute the sample sizes using formula (8) and provide results in Table IV, displaying N_{cp} , N_1 , and

N_2 with the empirical power $\tilde{P}_{cp}$ under $\psi_2^{-1}=1.7$, $S_1^{(1)}(\tau)=S_2^{(1)}(\tau)=0.5$, when ψ_2^{-1} and $\rho^{(k)}$ vary for the combinations of $\psi_2^{-1}=1.5, 1.7$ and $\rho^{(1)}=\rho^{(2)}=0, 0.3, 0.5$, and 0.8 .

The sample sizes N_{cp} from (8) usually provide slightly more conservative results from the perspective that the corresponding empirical powers $\tilde{P}_{cp}$ are slightly larger than the target powers, similarly to the one fatal case. The value of $N_{cp}^{(c)}$ is smaller than $N_{cp}^{(nc)}$ in any case, and $N_{cp}^{(c)}, N_{cp}^{(nc)}, N_1^{(c)}, N_1^{(nc)}$ and N_2 usually increase proportionally to $\rho^{(j)}$ in both of late and early dependencies, but $N_{cp}^{(c)}$ and N_2 decrease inversely only when $(\psi_1^{-1}, \psi_2^{-1})=(1.5, 1.7)$, $\rho^{(j)}=0.8$ and the Gumbel copula. That is, the stronger dependency makes the censoring rate increase more than the one fatal case, and hence, the reduction effect of N_{cp} based on correlated log-rank statistics are not obtained. On the other hand, the composite endpoint strategy using $N_1^{(c)}$ is the most reasonable in terms of smaller sample sizing. In particular, if $\psi_1^{-1} \approx \psi_2^{-1}$, we have $N_{cp}^{(c)} \approx N_2 \approx 2N_1^{(c)}$.

3.3. Illustration: the ICON7 study

We illustrate the sample size methods with an example. Consider ‘A Randomized, Two-Arm, Multi-Centre Gynaecologic Cancer Inter Group Trial of Adding Bevacizumab to Standard Chemotherapy (Carboplatin and Paclitaxel) in Patients With Epithelial Ovarian Cancer’ (ICON7) [18]. The study was designed to investigate the addition of bevacizumab to standard chemotherapy for first-line treatment of woman with ovarian cancer. The primary endpoints of interest were the PFS and OS. The protocol stated that the number of PFS events, 684 was needed to detect a 28% change in the PFS from a median value of 18 months in the control group to 23 months in the bevacizumab group (i.e., $\psi_{PFS}^{-1} \doteq 1.28$), with the power of 90% at the significance level of 5% (a two-sided log-rank test), while the number of OS events, 715 was required to detect a 23% improvement in the OS from a median value of 43 months in the control group to 53 months in the bevacizumab group (i.e., $\psi_{OS} = 1.23$), with the power of 80% at a significance level of 5% (two-sided test). The protocol sample size of 1520 patients was determined for the reasons that the required numbers of PFS and OS events were expected to occur until 36 and 60 months after the first randomized, respectively, assuming constant recruitment over 24 months, and some elements of uncertainty are considered.

We set 60 months PFS and OS rates in the control group as $S_{PFS}^{(1)}(60)=0.1$ and $S_{OS}^{(1)}(60)=0.4$ based on the aforementioned information, which is taking account of hazard rates

$\lambda_{PFS}^{(1)} = -\frac{1}{18} \log(0.5)$ and $\lambda_{OS}^{(1)} = -\frac{1}{43} \log(0.5)$ from the protocol and the exponential assumptions, and its uncertainty. Also, we derive the HR for TTP, because our calculation is based on the TTP and OS rather than the PFS and OS. That is, the HR and survival rate for

TTP can be computed as $\psi_{TTP}^{-1} = \lambda_{TTP}^{(1)} / \lambda_{TTP}^{(2)} \doteq 1.31$ and $S_{TTP}^{(1)}(60) \doteq 0.25$ from the information on medians based on

$$e^{-\lambda_{\text{TTP}}^{(j)} t} = S_{\text{TTP}}^{(j)}(t) = S_{\text{PFS}}^{(j)}(t) / S_{\text{OS}}^{(j)}(t) = e^{-\lambda_{\text{PFS}}^{(j)} t} / e^{-\lambda_{\text{OS}}^{(j)} t}, j=1, 2$$

from the independence and exponential assumptions. These values of the HR and $S_{\text{TTP}}^{(1)}$ are used for an illustration although the independence assumption is suspect.

Table V shows the total sample size N_{cp} , and the empirical powers $\tilde{P}_{\text{cp}}$, $\tilde{P}_{\text{TTP}}$, $\tilde{P}_{\text{PFS}}$, and $\tilde{P}_{\text{OS}}$ (%) to evaluate the joint and single hypotheses (H_1^{cp} and $H_{1,j}$) of TTP, PFS, and OS given N_{cp} , respectively, under $\psi_{\text{TTP}}^{-1} = 1.31$ and $\psi_{\text{OS}}^{-1} = 1.23$. Alternative sizing of MTTP, MPFS, and MOS gives the sample sizes required to test the single hypotheses of the TTP, PFS, and OS, respectively. The sample sizes were calculated to evaluate the joint reduction in both time-to-event outcomes of the TTP and OS or the PFS and OS with the target power of 80% at the significance level of 2.5%, based on the assumptions of the ICON7 study, assuming $\tau_a = 24$ months accrual duration and additional $\tau_f = 36$ months follow-up. The empirical power $\tilde{P}_{\text{PFS}}$ is larger than the target power of 90% to test the PFS written in the protocol, because the analysis times in the protocol design ($\tau_f = 12$) and ours ($\tau_f = 36$) are different. Similarly, note that N_{PFS} is calculated under the target power of 80% smaller than the power 90% in the protocol. Also, the expected numbers of bivariate events are listed using the notation corresponding to $D_{ij} = N_{\text{cp}} \times P_{ij}$ in which N_{cp} is replaced by $N_{\text{cp}}^{(\text{c})}$ or $N_{\text{cp}}^{(\text{nc})}$, and P_{ij} is the probability ($P_{\text{did\ddot{e}d\ddot{e}}^{\text{TTP}}}$; see Appendix B) calculated under the non-composite setting. Note that, even in the composite setting with the PFS and OS, D_{ij} represents the number that either or both of the TTP and OS are observed. In addition, the correlation is assumed to be common between the two groups, that is, $\rho^{(1)} = \rho^{(2)}$.

In calculating the sample size, we assume that one event is fatal because the TTP may be censored by the OS of fatal event. If the association between the TTP and OS is late-time dependent, then the total sample sizes $N_{\text{cp}}^{(\text{c})}$ required to test the PFS and OS jointly with common correlation between the two groups, $\rho^{(k)} = 0.0, 0.3, 0.5$, and 0.8 are 1510, 1528, 1544, and 1564, respectively. The sample size increases monotonically to $\rho^{(k)} = 0.8$ from 0: the difference between the smallest and largest sample sizes may seem relatively large, although the ratio is only about 1.04. If the association is early-time dependent, then the total sample sizes $N_{\text{cp}}^{(\text{c})}$ required under $\rho^{(k)} = 0.0, 0.3, 0.5$ and 0.8 are 1510, 1508, 1506, and 1498, respectively. The sample size decreases with increasing correlation, but the reduction rate from the largest sample size given by $\rho^{(k)} = 0.0$ is quite small. Also, comparing the composite and non-composite endpoints, $N_{\text{cp}}^{(\text{c})}$ is smaller than the $N_{\text{cp}}^{(\text{nc})}$ required to test the TTP and OS jointly, but the difference is slight. The expected numbers of bivariate events, D_{ij} provide useful information in monitoring process of the trial.

4. Multiple primary endpoints

4.1. Hypothesis testing, power, and sample sizes

We discuss calculating the required sample size for trials with multiple primary endpoints. For simplicity, we will consider the problem on the power and sample size calculation using the simplest procedures, the (weighted) Bonferroni procedure as well known and widely used. The other procedures for controlling the Type I error rate are available [6, 19, 20].

The weighted Bonferroni procedure allocates the Type I error rate α between the endpoints with weight ω , that is, $\alpha_1 = \omega\alpha$ for the first endpoint and $\alpha_2 = (1-\omega)\alpha$ for the second endpoint. We are then interested in testing $H_0^{\text{mp}} = \{H_{01}:\psi_1(t)=1 \text{ and } H_{02}:\psi_2(t)=1\} \text{ for all } t\}$ versus $H_1^{\text{mp}} = \{H_{11}:\psi_1(t)<1 \text{ or } H_{12}:\psi_2(t)<1 \text{ at some } t\}$ at the (overall) significance level of .3, based on the log-rank test statistics $Z_1^{(0)}$ and $Z_2^{(0)}$ given in Section 2. The testing procedure is to

$$\text{reject the null hypothesis } H_0^{\text{mp}} \text{ if and only if } Z_1^{(0)} < -z_{\alpha_1} \text{ or } Z_2^{(0)} < -z_{\alpha_2}, \quad (9)$$

where z_{α_1} and z_{α_2} are the $100(1 - \alpha_1)$ and $100(1 - \alpha_2)$ percentiles of $N(0, 1)$, respectively.

Therefore, because the rejection region of H_0^{mp} is $\{(Z_1^{(0)}, Z_2^{(0)}) : Z_1^{(0)} < -z_{\alpha_1} \text{ or } Z_2^{(0)} < -z_{\alpha_2}\}$, we have the power function for a reduction in either of the time-to-event outcomes with the weighted Bonferroni procedure,

$$1 - \beta = \Pr \left[\bigcup_{k=1}^2 \{Z_k^{(0)} < -z_{\alpha_k}\} \mid H_1^{\text{mp}} \right] = 1 - \Pr \left[\bigcap_{k=1}^2 \{Z_k^{(0)} \geq -z_{\alpha_k}\} \mid H_1^{\text{mp}} \right]$$

This overall power is referred to as ‘minimal power’[12] or ‘disjunctive power’[13]. Similarly to Section 3.1, for all of the three censoring scenarios, the power based on the function (5) can be approximated by

$$1 - \beta \approx 1 - \int_{-z_{\alpha_1}}^{\infty} \int_{-z_{\alpha_2}}^{\infty} \phi \left(z_1, z_2; \sqrt{N}\delta, \sum \right) dz_1 dz_2. \quad (10)$$

Letting N_{mp} be the minimum of total sample size N required for testing H_0^{mp} against H_1^{mp} , the formula for N_{mp} is given by

$$N_{\text{mp}} = \lceil \left(L_{\beta} - z_{\alpha_2} \sigma_2^{(0)} / \sigma_2 \right)^2 / \delta_2^2 \rceil, \quad (11)$$

where L_{β} is the solution of the integral equation

$$\beta = \int_{-\infty}^{\frac{\delta_1}{\delta_2} L_\beta + \left\{ z_{\alpha_1} \frac{\sigma_1^{(0)}}{\sigma_1} - z_{\alpha_2} \frac{\delta_1}{\delta_2} \frac{\sigma_2^{(0)}}{\sigma_2} \right\}} \int_{-\infty}^{L_\beta} \phi(z_1, z_2; \mathbf{0}, \mathbf{R}) dz_1 dz_2$$

and $\sigma_k^{(0)}$, σ_k , and $\mathbf{R}$ are the same as the definitions given in Section 3.1. Similarly to the methods to obtain N_{cp} , we can use a search method of L_β such as the Newton–Raphson algorithm [14] or the basic linear interpolation algorithm [2] in order to compute N_{mp} , which generally take shorter computing time than direct search that let N_{mp} increase sequentially until (10) exceeds the desired power. Also, similarly to Section 3.1, the required numbers of events that are useful to monitor a trial can be obtained by $D_{ij} = N_{mp} \times P_{ij}$ (see Appendix B for details regarding calculation of P_{ij} under the three censoring scenarios).

The overall Type I error associated with the null H_0^{mp} is controlled by the sum of marginal Type I errors. Similarly to Section 3, the inflation problem of the overall Type I error in the procedure (6) based on the asymptotic normal approximation is parallel to that of the marginal ones. The overall significance level $\alpha_1 + \alpha_2$ on H_0^{mp} is quite close to the level $\alpha_1 + \alpha_2 - \alpha_1 \alpha_2$ considered when two endpoints are independent, so that the influence on marginal inflations may be practically larger in multiple primary problem than in multiple co-primary one. However, we can control the errors by correcting the critical values based on the sample size N and the allocation rate $r^{(1)}$, using simulation or theoretical result well known in the univariate log-rank statistic [9–11]. The practical need for sample size formula is a balanced weighting of the precision and cost, and then the use of the formula based on the asymptotic normal approximation is left in each stage of practice.

4.2. Illustration: one fatal case

We illustrate the behavior of the sample size and power for detecting at least one reduction H_1^{mp} in bivariate time-to-event data focusing on the one fatal outcome scenario with the two time-dependent association structures. Similarly to Section 3.2, we generate bivariate time-to-event data and perform Monte Carlo trials. The target power $1 - \beta = 0.8$, the significance level $\alpha = 0.025$, $\tau_a = 2$, and $\tau_f = 3$ are used. The notations of N_{EP} , ψ_{EP} , and $S_{EP}^{(j)}(t)$ for EP = TTP, PFS, and OS are the same as those in Section 3. Also, N_{mp} is written as $N_{mp}^{(nc)}$ when the first endpoint is the TTP, or as $N_{mp}^{(c)}$ when the first endpoint is the PFS. Let $\tilde{P}_{mp}$ denote the empirical power (%) for detecting H_1^{mp} under N_{mp} participants designed using the formula (11). Further, we select $\omega = \omega_0$ as an optimal value of ω to consider a variable weighting strategy for the weighted Bonferroni procedure, that is, ω_0 is ω that gives a minimum of N_{mp} over $\omega = 0, 0.05, \dots, 0.95, 1$ ($\omega_0 = \operatorname{argmin}_{\omega \in \{0, 0.05, \dots, 0.95, 1\}} N_{mp}$). One may consider the other testing procedures such as the fixed-sequence procedure when $\omega_0 = 0$ or $\omega_0 = 1$ is suggested.

Table VI displays the required total sample sizes N_{mp} (given $\omega = \omega_0$), the alternative sizing solutions N_{PFS} , N_{TTP} , and N_{OS} , the selected Bonferroni weight ω_0 and the empirical power

$\tilde{P}_{mp}$ under $\psi_{TTP}^{-1}=1.7$, $S_{TTP}^{(1)}(\tau)=0.3$, and $S_{OS}^{(1)}(\tau)=0.5$ when ψ_{OS}^{-1} and $\rho^{(k)}$ vary in the combinations of $\psi_{OS}^{-1}=1.3, 1.5, 1.7$ and $\rho^{(1)}=\rho^{(2)}=0, 0.3, 0.5, 0.8$.

The sample sizes N_{mp} obtained from the formula (11) consistently provide conservative results because the empirical powers $\tilde{P}_{mp}$ are slightly larger than the target powers. Although $N_{mp}^{(c)} < N_{mp}^{(nc)}$ in many cases, we observe $N_{mp}^{(c)} > N_{mp}^{(nc)}$ in some cases where N_{TTP} is relatively close to N_{PFS} or smaller than N_{PFS} , which occurs when $\rho^{(j)}$ is relatively large and the effect size of OS (absolute value of $\log \psi_{OS}^{-1}$) is smaller than that of TTP. Also, if the effect size of OS is smaller than that of TTP, then $N_{TTP} \ll N_{OS}$ or $N_{PFS} \ll N_{OS}$ occurs, so that we have $N_{mp}^{(nc)} \approx N_{TTP}$ or $N_{mp}^{(c)} \approx N_{PFS}$ accompanied with the value of ω_0 close to 1. When N_{TTP} (or N_{PFS}) is closer to N_{OS} , ω_0 takes a value between 0 and 1.

Further, we add a summary about the tendency how ω_0 -value ranges from 0 to 1 in Section B.2 of Supporting Information, based on the underlying results by varying $S_{TTP}(\tau)$ and $\rho^{(k)}$ leaving the other factors fixed, whose parts are provided in Figures B.4, B.5, B.6, and B.7. We provide a guideline based on the summary about ω_0 -value: generally, the correlation $\rho^{(k)}$ and the dependence structure are unknown, so that these should be examined using meta analysis and/or data from a pilot study, similarly to consideration for the HRs of effect sizes. Although it may be desirable to consider a maximum sample size considering such unknown factors, it is necessary to balance on a degree of uncertainty and the cost. When the TTP and OS are used as two primary endpoints, ω_0 ranges from 0 to 1, but ω_0 is approximately 0.5 when N_{TTP}/N_{OS} is close to 1. When the PFS and OS are used as two primary endpoints, the ω_0 -value is usually close to 1 under early dependency and under a late dependency with not high $\rho^{(k)}$, if $\psi_{PFS}^{-1} \geq \psi_{OS}^{-1}$. The value of ω_0 may be away from 0 and 1 otherwise. These considerations about ω_0 -value suggest that the use of PFS endpoint is one of the reasonable strategies in the multiple primary problem for the TTP and OS if $\psi_{TTP}^{-1} \geq \psi_{OS}^{-1}$, although the PFS endpoint may be used without consideration in practice. Hence, it is useful to incorporate ω_0 in the sample size calculation proposed for multiple primary endpoints.

4.3. Behavior of the sample size as a function of the correlation

We extend the study to the both non-fatal and both fatal cases. We use the same setting as in Section 4.2 ($1 - \beta = 0.8$, $\alpha = 0.025$, $\tau_a = 2$, and $\tau_f = 3$), while the notations of N_k , ψ_k , and $S_k^{(j)}(t)$ for the k th endpoint are used because of the consistent expression of the results from the three situations. Similarly to Section 4.2, we select a minimum of N_{mp} over $\omega = 0, 0.05, \dots, 0.95, 1$ via a strategy for the weighted Bonferroni procedure, $N_{mp}^{(c)}$ is N_{mp} calculated under the composite setting, and $N_{mp}^{(nc)}$ is N_{mp} calculated under the non-composite setting.

Figure 1 shows the required total sample sizes N_{mp} as a function of $\rho^{(1)} = \rho^{(2)} = 0, 0.05, \dots, 0.95$ when ψ_2^{-1} varies from 1.3, 1.5 and 1.7 under $\psi_1^{-1}=1.7$, $S_1^{(1)}(\tau)=0.3$ and the early dependency (Gumbel copula). The 12 plots are arranged from the left for $S_2^{(1)}(\tau)=0.4, 0.5$ and 0.6, and from the top for the four scenarios: the both non-fatal case without the

composite setting ($N_{cp}^{(nc)}$), the one fatal case without and with the composite setting ($N_{mp}^{(nc)}$ and $N_{mp}^{(c)}$), and the both fatal case with the composite setting ($N_{mp}^{(c)}$). The plots for the late dependency case are provided in Figure B.3 in Section B.1 of Supporting Information (generated under the Clayton copula (late dependency) using the same conditions as Figure 1 except the copula model).

As $\rho^{(j)}$ increases, the behavior of $N_{mp}^{(nc)}$ is monotone increasing up to a constant value (univariate sample size) in the both non-fatal case but is more complicated in the one fatal case. In the one fatal case, there is a complicated interaction between the effect size and the correlation, where a smaller sample size is required with increasing effect size and decreasing correlation, while the censoring rate for non-fatal events increases with increasing correlation. Consider the case of $\rho^{(j)} = 1$ as a standard, because the sample size N_{mp} required in bivariate primary endpoints reduces to that of the single endpoint in the optimal weighting Bonferroni strategy when the correlations $\rho^{(j)}$ are 1. In the one fatal case with the composite setting, the complicated behavior is moderated, and $N_{mp}^{(c)}$ tends to decrease as the correlation decreases, but $N_{mp}^{(c)}$ is sometimes larger than $N_{mp}^{(nc)}$. The behavior of $N_{mp}^{(c)}$ in the both fatal case with the composite setting is similar to that of the one fatal case.

5. Summary

Utilizing multiple endpoints in clinical trials may provide the opportunity for characterizing intervention's multidimensional effects but also creates challenges in design and analysis of clinical trials. Specifically controlling the Types I and II error rates is non-trivial when the multiple primary endpoints are potentially correlated. When designing the trial to detect effects for *all* of the endpoints, no adjustment is needed to control the Type I error. However, the Type II error increases as the number of endpoints being evaluated increases. In contrast, when designing the trial to detect an effect for *at least one* of the endpoints, then an adjustment is needed to control the Type I error.

We describe an approach to the evaluation of power and sample size for comparing the effect of two interventions in superiority clinical trials with two time-to-event outcomes for multiple co-primary and multiple primary cases. Designing clinical trials with multiple time-to-event outcomes is more complex compared with endpoints with the other scales, requiring censoring scheme challenges and time-dependent associations among outcomes. We consider the three censoring scenarios based on the types of outcomes: (i) both outcomes are non-fatal; (ii) one outcome is fatal; and (iii) both outcomes are fatal, and we evaluate their composite and non-composite settings. We discuss the two time-dependent association structures: the asymmetric late time-dependency generated by the Clayton copula and the early time-dependency generated by the Gumbel copula. Our findings are summarized as follows.

- In the co-primary endpoint situation, if the two time-to-event outcomes are non-fatal, then the required sample size $N_{cp}^{(nc)}$ decreases with increasing correlations

$\rho^{(j)}$, in both of the late and early time-dependencies, except for the case where one HR is larger than the other. When the correlations are zero, $N_{cp}^{(nc)}$ is the largest. However, when one or both outcomes are non-fatal, the behaviors of required sample sizes $N_{cp}^{(nc)}$ and $N_{cp}^{(c)}$ have complicated shapes owing to an interaction between the correlations and effect sizes based on the HRs. Zero correlation does not provide the largest required sample size. Thus, careful consideration is required in practice.

- In the multiple primary endpoint situation based on the optimal weighting Bonferroni strategy, a standard situation occurs when the correlations $\rho^{(j)}$ are one, corresponding to the single endpoint case. Considering the two non-fatal case, the sample size $N_{mp}^{(nc)}$ does not vary or increases monotonously up to a constant (univariate sample size) as a function of the correlations, in both of the late and early time-dependencies. However, when one or both outcomes are non-fatal, the behaviors of the required sample sizes $N_{mp}^{(nc)}$ and $N_{mp}^{(c)}$ are complex with an interaction between the correlations and the effect size ratios. Thus, the proposed formula is useful for determining the sample size in practice, noting, when the sample size is small, for example, < 100 , one may have to modify a critical value based on the normal approximation.
- Unlike the both non-fatal case, larger correlations do not increase the statistical power in the one fatal or both fatal cases. The reason for this is that higher correlation leads to increased censoring. Hence, the standard log-rank statistic must be corrected by incorporating information from informative censoring. We focus on providing a foundation for designing clinical trials with two time-to-event outcomes. Informative censoring is a topic for future work.

When designing a clinical trial with the proposed methods, one needs parameter estimates using the available methods ([7, 21–25, 27, 28]). But specification for the joint distribution is challenging as data to base selection is often limited during trial design. One conservatively alternative is to select the largest sample size of all of the correlation and joint distribution combinations, and stop the clinical trial when the appropriate number of events required for each outcome is observed. Another option is to use group-sequential designs. This may lead to potentially fewer patients than the fixed-sample designs when evidence is overwhelming and thus offers efficiency but introduces the other challenges. Information for the endpoints may not accrue at the same rate and require different information times.

As discussed in [29] and [26], the Type I error rate of the log-rank test for each endpoint may be inflated in small sample sizes or with unequally sized intervention groups. When this occurs in the co-primary endpoint situation, our simulation studies suggest that the overall Type I error associated with the null hypothesis H_0^{cp} is not larger than the target significance level, except when the correlation is very high (i.e., close to one), even though the marginal Type I error rate is inflated. On the other hand with multiple primary endpoints, the overall Type I error associated with the null hypothesis H_0^{mp} is larger than the target significance level, particularly when the correlation is small. In these cases, we may consider more direct

ways of calculating sample size without using a normal approximation such as the methods in [29] and [26].

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank the two anonymous referees for their helpful suggestions and constructive comments that improved the content and presentation. We also thank Dr. Lu Tian, Dr. H.M. James Hung and Dr. Sue-Jane Wang in encouraging us with their valuable comments for this research. This work was supported by JSPS KAKENHI Grant Number 26330032.

References

1. Fine JP, Jiang H, Chappell R. On semi-competing risks data. *Biometrika*. 2001; 88:907–919. DOI: 10.1093/biomet/88.4.907
2. Hamasaki T, Sugimoto T, Evans SR, Sozu T. Sample size determination for clinical trials with co-primary outcomes: exponential event times. *Pharmaceutical Statistics*. 2013; 12:28–34. DOI: 10.1002/pst.1545 [PubMed: 23081932]
3. Sugimoto T, Sozu T, Hamasaki T, Evans SR. A logrank test-based method for sizing clinical trials with two co-primary time-to-event endpoints. *Biostatistics*. 2013; 14:409–421. DOI: 10.1093/biostatistics/kxs057 [PubMed: 23307913]
4. Offen W, Chuang-Stein C, Dmitrienko A, Littman G, Maca J, Meyerson L, Muirhead R, Stryszak P, Boddy A, Chen K, Copley-Merriman K, Dere W, Givens S, Hall D, Henry D, Jackson JD, Krishen A, Liu T, Ryder S, Sankoh AJ, Wang J, Yeh CH. Multiple co-primary endpoints: medical and statistical solutions. *Drug Information Journal*. 2007; 41:31–46. DOI: 10.1177/009286150704100105
5. Hung HMJ, Wang SJ. Some controversial multiple testing problems in regulatory applications. *Journal of Biopharmaceutical Statistics*. 2009; 19:1–11. DOI: 10.1080/10543400802541693 [PubMed: 19127460]
6. Dmitrienko, A., Tamhane, AC., Bretz, F. *Multiple Testing Problems in Pharmaceutical Statistics*. Chapman and Hall; Boca Raton, FL: 2010.
7. Wang W. Estimating the association parameter for copula models under dependent censoring. *Journal of the Royal Statistical Society, Series B*. 2003; 65:257–273. DOI: 10.1111/1467-9868.00385
8. Hsu L, Prentice RL. On assessing the strength of dependency between failure time variates. *Biometrika*. 1996; 83:491–506. DOI: 10.1093/biomet/83.3.491
9. Kellerer AM, Chmelevsky D. Small-sample properties of censored-data rank tests. *Biometrics*. 1983; 39:675–682. DOI: 10.2307/2531095
10. Hsieh FY. Comparing sample size formulae for trials with unbalanced allocation using the logrank test. *Statistics in Medicine*. 1992; 11:1091–1098. DOI: 10.1002/sim.4780110810 [PubMed: 1496196]
11. Strawderman RL. An asymptotic analysis of the logrank test. *Lifetime Data Analysis*. 1997; 3:225–249. DOI: 10.1023/A:1009648914586 [PubMed: 9384654]
12. Westfall, PH., Tobias, RD., Rom, D., Wolfinger, RD., Hochberg, Y. *Multiple Comparisons and Multiple Tests Using the SAS System*. SAS; Cary NC: 2011.
13. Senn S, Bretz F. Power and sample size when multiple endpoints are considered. *Pharmaceutical Statistics*. 2007; 6:161–170. DOI: 10.1002/pst.301 [PubMed: 17674404]
14. Sugimoto T, Sozu T, Hamasaki T. A convenient formula for calculating sample size of clinical trials with multiple co-primary continuous endpoints. *Pharmaceutical Statistics*. 2012; 11:118–128. DOI: 10.1002/pst.505 [PubMed: 22415870]

15. Freedman LS. Table of the number of patients required in clinical trials using the logrank test. *Statistics in Medicine*. 1982; 1:121–129. DOI: 10.1002/sim.4780010204 [PubMed: 7187087]
16. Clayton DG. A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease. *Biometrika*. 1978; 65:141–151. DOI: 10.1093/biomet/65.1.141
17. Hougaard P. A class of multivariate failure time distribution. *Biometrika*. 1986; 73:671–678. DOI: 10.1093/biomet/71.1.75
18. Perren TJ, Swart AM, Pfisterer J, Ledermann JA, Pujade-Lauraine E, Kristensen G, Carey MS, Beale P, Cervantes A, Kurzeder C, du Bois A, Sehouli J, Kimmig R, Stähle A, Collinson F, Essapen S, Gourley C, Lortholary A, Selle F, Mirza MR, Leminen A, Plante M, Stark D, Qian W, Parmar MK, Oza AM. A phase 3 trial of bevacizumab in ovarian cancer. *New England Journal of Medicine*. 2011; 365:2484–2496. DOI: 10.1056/NEJMoa1103799 [PubMed: 22204725]
19. Wiens B, Dmitrienko A. On selecting a multiple comparison procedure for analysis of a clinical trial: fallback, fixed-sequence and related procedures. *Statistics in Biopharmaceutical Research*. 2010; 2:22–32. DOI: 10.1198/sbr.2010.08035
20. Bretz, F., Hothorn, T., Westfall, P. *Multiple Comparisons Using R*. Boca Raton, FL: Chapman and Hall; 2011.
21. Lagakos SW. A stochastic model for censored-survival data in the presence of an auxiliary variable. *Biometrics*. 1976; 32:551–559. [PubMed: 963170]
22. Lagakos SW. Using auxiliary variables for improved estimates of survival time. *Biometrics*. 1977; 33:399–404. [PubMed: 884198]
23. Lin DY, Robins JM, Wei LJ. Comparing two failure time distributions in the presence of dependent censoring. *Biometrika*. 1996; 83:381–393. DOI: 10.1093/biomet/83.2.381
24. Shih JH. A goodness-of-fit for association in a bivariate survival model. *Biometrika*. 1998; 85:189–200. DOI: 10.1093/biomet/85.1.189
25. Chang SH. A two-sample comparison for multiple ordered event data. *Biometrics*. 2000; 56:183–189. [PubMed: 10783794]
26. Wang R, Lagakos SW, Gray RJ. Testing and interval estimation for two-sample survival comparisons with small sample sizes and unequal censoring. *Biostatistics*. 2010; 11:676–692. DOI: 10.1093/biostatistics/kxq021 [PubMed: 20439258]
27. Siannis F, Farewell VT, Head J. A multi-state model for joint modelling of terminal and non-terminal events with application to Whitehall II. *Statistics in Medicine*. 2007; 26:426–442. DOI: 10.1002/sim.2342 [PubMed: 16220522]
28. Parast L, Tian L, Cai T. Landmark estimation of survival and treatment effect in a randomized clinical trial. *Journal of the American Statistical Association*. 2014; 109:384–394. DOI: 10.1080/01621459.2013.842488 [PubMed: 24659838]
29. Heinze G, Gnant M, Schemper M. Exact log-rank test for unequal follow-up. *Biometrics*. 2003; 59:1151–1157. DOI: 10.1111/j.0006-341X.2003.00132.x [PubMed: 14969496]

APPEENDIX A. Asymptotic forms in the bivariate log-rank statistic

We provide the details of asymptotic forms of the bivariate log-rank statistic discussed in Section 2.3. See [3] for the case when both endpoints are non-fatal. We obtain the covariance process between two incremental differences $dM_{ik}(t) = d\mathcal{N}_{ik}(t) - \mathcal{Y}_{ik}(t)d\tilde{\Lambda}_k^{(g_i)}(t)$, $k = 1, 2$, where $M_{11}(t)$ and $M_{22}(t)$ are martingale processes relative to the filtration generated by the history prior to time t under the null hypothesis. This corresponds to calculating the expectation of

$$\begin{aligned} dM_{i1}(t)dM_{i2}(s) &= d\mathcal{N}_{i1}(t)d\mathcal{N}_{i2}(s) - \mathcal{Y}_{i1}(t)d\tilde{\Lambda}_1^{(g_i)}(t)d\mathcal{N}_{i2}(s) \\ &\quad - \mathcal{Y}_{i2}(s)d\tilde{\Lambda}_2^{(g_i)}(s)d\mathcal{N}_{i1}(t) + \mathcal{Y}_{i1}(t)\mathcal{Y}_{i2}(s)d\tilde{\Lambda}_1^{(g_i)}(t)d\tilde{\Lambda}_2^{(g_i)}(s). \end{aligned} \quad (\text{A.1})$$

One fatal case

Consider the non-composite setting with the censoring indicator $\Delta_{ik} = (T_{ik} = T_{ik}^*)$, $k = 1, 2$. Because $T_{i1} = \min(T_{i1}^*, T_{i2}^*, C_i)$ and $T_{i2} = \min(T_{i2}^*, C_i)$, the expectations of the i th at-risk processes given $g_i = j$, that is, $y_k^{(j)}(t)$, $k = 1, 2$ are

$$y_1^{(j)}(t) = E[\mathcal{Y}_{i1}(t)|g_i=j] = \Pr(T_{i1}^* \geq t, T_{i2}^* \geq t, C_i \geq t|g_i=j) = S^{(j)}(t, t)G(t)$$

for the non-fatal endpoint and

$$y_2^{(j)}(t) = E[\mathcal{Y}_{i2}(t)|g_i=j] = \Pr(T_{i2}^* \geq t, C_i \geq t|g_i=j) = S_2^{(j)}(t)G(t)$$

for the fatal endpoint (the latter form is identical to that of both non-fatal case). So each element in $\hat{H}_k(t)$ converges to the corresponding expectation almost surely by Glivenko–Cantelli’s theorem under some regular conditions, so that we have the asymptotic forms of $\hat{H}_k(t)$, $k = 1, 2$ as

$$H_k(t) = N^{-1} \frac{E[\mathcal{Y}_k^{(1)}(t)] E[\mathcal{Y}_k^{(2)}(t)]}{E[\mathcal{Y}_k^{(1)}(t)] + E[\mathcal{Y}_k^{(2)}(t)]} = \begin{cases} G(t) \frac{r^{(1)}r^{(2)}S^{(1)}(t,t)S^{(2)}(t,t)}{r^{(1)}S^{(1)}(t,t)+r^{(2)}S^{(2)}(t,t)}, & k=1, \\ G(t) \frac{r^{(1)}r^{(2)}S_2^{(1)}(t)S_2^{(2)}(t)}{r^{(1)}S_2^{(1)}(t)+r^{(2)}S_2^{(2)}(t)}, & k=2. \end{cases}$$

Similarly, the conditional expectations of the i th counting processes given $\mathcal{Y}_{ik}(t) = 1$ are

$$\begin{aligned} d\tilde{\Lambda}_1^{(g_i)}(t) &= E[d\mathcal{N}_{i1}(t)|\mathcal{Y}_{i1}(t)=1, g_i] \\ &= \frac{\Pr(t \leq T_{i1}^* < t+dt, t \leq T_{i2}^*, t \leq C_i|g_i)}{\Pr(\mathcal{Y}_{i1}(t)=1|g_i)} = -\frac{S^{(g_i)}(dt, t)}{S^{(g_i)}(t, t)} = \Lambda^{(g_i)}(dt, t) \end{aligned}$$

for the non-fatal endpoint and $d\tilde{\Lambda}_2^{(g_i)}(t) = E[d\mathcal{N}_{i2}(t)|\mathcal{Y}_{i2}(t)=1, g_i] = d\Lambda_2^{(g_i)}(t)$ for the fatal endpoint. Hence, the expectation of (A.1) is

$$\begin{aligned}
& E[dM_{i1}(t)dM_{i2}(s)|g_i] = \Pr(\mathcal{Y}_{i1}(t)\mathcal{Y}_{i2}(s)=1|g_i)E[dM_{i1}(t)dM_{i2}(s)|\mathcal{Y}_{i1}(t)\mathcal{Y}_{i2}(s)=1, g_i] \\
& = G(t \vee s) \left\{ (t \leq s)S^{(g_i)}(dt, ds) + (t \leq s)S^{(g_i)}(t, ds)\Lambda^{(g_i)}(dt, t) + S^{(g_i)}(dt, t \vee s)d\Lambda_2^{(g_i)}(s) + S^{(g_i)}(t, t \vee s)\Lambda^{(g_i)}(dt, t)d\Lambda_2^{(g_i)}(s) \right\}, \\
& \tag{A.2}
\end{aligned}$$

because, noting that $T_{i1} \leq T_{i2}$ always holds, we have

$$\begin{aligned}
& \Pr(\mathcal{Y}_{i1}(t)\mathcal{Y}_{i2}(s)=1|g_i) = \Pr(t \leq T_{i1}^*, t \vee s \leq T_{i2}^*, t \vee s \leq C_i | g_i) = G(t \vee s)S^{(g_i)}(t, t \vee s), \\
& E[d\mathcal{N}_{i1}(t)d\mathcal{N}_{i2}(s)|\mathcal{Y}_{i1}(t)\mathcal{Y}_{i2}(s)=1, g_i] = \frac{\Pr(t \leq T_{i1}^* < t+dt, s \leq T_{i2}^* < s+ds | g_i)}{\Pr(t \leq T_{i1}^*, t \vee s \leq T_{i2}^* | g_i)} = (t \leq s) \frac{S^{(g_i)}(dt, ds)}{S^{(g_i)}(t, t \vee s)}, \\
& E[\mathcal{Y}_{i1}(t)d\mathcal{N}_{i2}(s)|\mathcal{Y}_{i1}(t)\mathcal{Y}_{i2}(s)=1, g_i] = \frac{\Pr(t \leq T_{i1}^*, s \leq T_{i2}^* < s+ds, t \vee s \leq T_{i2}^* | g_i)}{\Pr(t \leq T_{i1}^*, t \vee s \leq T_{i2}^* | g_i)} = -(t \leq s) \frac{S^{(g_i)}(t, ds)}{S^{(g_i)}(t, t \vee s)} \text{ and} \\
& E[d\mathcal{N}_{i1}(t)\mathcal{Y}_{i2}(s)|\mathcal{Y}_{i1}(t)\mathcal{Y}_{i2}(s)=1, g_i] = \frac{\Pr(t \leq T_{i1}^* < t+dt, t \vee s \leq T_{i2}^* | g_i)}{\Pr(t \leq T_{i1}^*, t \vee s \leq T_{i2}^* | g_i)} = -\frac{S^{(g_i)}(dt, t \vee s)}{S^{(g_i)}(t, t \vee s)}.
\end{aligned}$$

Therefore, we have the form of $dA^{(j)}(t, s) = G(t \vee s)^{-1}E[dM_{i1}(t)dM_{i2}(s) | g_i = j]$ given in (3a) of Section 2.3.

Next, assume the composite setting $\Delta_{i1} = (T_{i1} = \min(T_{i1}^*, T_{i2}^*))$ and $\Delta_{i2} = (T_{i2} = T_{i2}^*)$. That is, the definition of $d\mathcal{N}_{i1}(t)$ is only changed from the non-composite version. Because observing the 2nd endpoint is composed in the i th counting processes $\mathcal{N}_{i1}(t)$ for the first endpoint, the conditional expectations of $d\mathcal{N}_{i1}(t)$ given $\mathcal{Y}_{i1}(t) = 1$ is

$$\begin{aligned}
d\tilde{\Lambda}_1^{(g_i)}(t) &= E[d\mathcal{N}_{i1}(t)|\mathcal{Y}_{i1}(t)=1, g_i] = \frac{\sum_{k=1,2} \Pr(t \leq T_{ik}^* < t+dt, t \leq T_{ik}^*, t \leq C_i | g_i)}{\Pr(\mathcal{Y}_{i1}=1 | g_i)} \\
&= \Lambda^{(g_i)}(dt, t) + \Lambda^{(g_i)}(t, dt),
\end{aligned}$$

where $k' = 3-k$. So, the terms of $d\tilde{\Lambda}_1^{(g_i)} = \Lambda^{(g_i)}(dt, t)$ in (A.2) are replaced by $\Lambda^{(g_i)}(dt, t) + \Lambda^{(g_i)}(t, dt)$ in this composite setting. Similarly, the other corrections in (A.2) are of terms of the conditional expectations including $d\mathcal{N}_{i1}(t)$, and derived as

$$\begin{aligned}
E[d\mathcal{N}_{i1}(t)d\mathcal{N}_{i2}(s)|\mathcal{Y}_{i1}(t)\mathcal{Y}_{i2}(s)=1, g_i] &= \frac{(t \leq s)S^{(g_i)}(dt, ds) - (t=s)S^{(g_i)}(t, ds)}{S^{(g_i)}(t, t \vee s)} \text{ and} \\
E[d\mathcal{N}_{i1}(t)\mathcal{Y}_{i2}(s)|\mathcal{Y}_{i1}(t)\mathcal{Y}_{i2}(s)=1, g_i] &= -\frac{S^{(g_i)}(dt, t \vee s) + (t \leq s)S^{(g_i)}(dt, t \vee s)}{S^{(g_i)}(t, t \vee s)}.
\end{aligned}$$

We achieve (3b) applying these results into the expectation of (A.1). In this composite setting, we can see that the intensity information on the 2nd endpoint (death event) is added into $d\tilde{\Lambda}_1^{(j)}(t)$ and $dA^{(j)}(t, s)$ compared with the non-composite version.

Both fatal case

We consider the non-composite setting, $\Delta_{ik} = (T_{ik} = T_{ik}^*)$, $k = 1, 2$ and omit the composite setting for simplicity. Following $T_{i1} = \min(T_{i1}^*, T_{i2}^*, C_i)$ and $T_{i2} = \min(T_{i1}^*, T_{i2}^*, C_i)$, the expectations of the i th at-risk processes given $g_i = j$ are the same forms as

$$y_k^{(j)}(t) = E[\mathcal{Y}_{ik}(t) | g_i = j] = \Pr(t \leq T_{i1}^*, t \leq T_{i2}^*, t \leq C_i | g_i) = S^{(j)}(t, t)G(t),$$

for both endpoints, identical to that obtained for the one fatal case. Hence, by Glivenko–Cantelli’s theorem under some regular conditions, the asymptotic forms of $\hat{H}_k(t)$, $k = 1, 2$ are

$$H_k(t) = N^{-1} \frac{E[\mathcal{Y}_k^{(1)}(t)] E[\mathcal{Y}_k^{(2)}(t)]}{E[\mathcal{Y}_k^{(1)}(t)] + E[\mathcal{Y}_k^{(2)}(t)]} = G(t) \frac{r^{(1)}r^{(2)}S^{(1)}(t, t)S^{(2)}(t, t)}{r^{(1)}S^{(1)}(t, t) + r^{(2)}S^{(2)}(t, t)}.$$

The conditional expectations of the i th counting processes given $\mathcal{Y}_{ik}(t) = 1$ are

$$\begin{aligned} d\tilde{\Lambda}_k^{(g_i)}(t) &= E[d\mathcal{N}_{ik}(t) | \mathcal{Y}_{ik}(t) = 1, g_i] = \frac{\Pr(t \leq T_{ik}^* < t + dt, t \leq T_{ik}^*, t \leq C_i | g_i)}{\Pr(\mathcal{Y}_{ik}(t) = 1 | g_i)} \\ &= \begin{cases} \frac{S^{(g_i)}(dt, t)}{S^{(g_i)}(t, t)} &= \Lambda^{(g_i)}(dt, t), \quad k=1 \\ \frac{S^{(g_i)}(t, dt)}{S^{(g_i)}(t, t)} &= \Lambda^{(g_i)}(t, dt), \quad k=2 \end{cases} \end{aligned}$$

similar to the result for the non-fatal endpoint in the one fatal case, where $k' = 3 - k$. Hence, the expectation of (A.1) can be derived as

$$\begin{aligned} E[dM_{i1}(t)dM_{i2}(s) | g_i] &= G(t \vee s) \left\{ (t=s)S^{(g_i)}(dt, ds) + (t \leq s)S^{(g_i)}(t \vee s, ds)\Lambda^{(g_i)}(dt, t) \right. \\ &\quad \left. + (t \geq s)S^{(g_i)}(dt, t \vee s)d\Lambda_2^{(g_i)}(s) + S^{(g_i)}(t \vee s, t \vee s)\Lambda^{(g_i)}(dt, t)\Lambda^{(g_i)}(s, ds) \right\}, \end{aligned}$$

(A.3)

because

$$\begin{aligned}
\Pr(\mathcal{Y}_{i1}(t)\mathcal{Y}_{i2}(s)=1|g_i) &= \Pr(t \vee s \leq T_{i1}^*, t \vee s \leq T_{i2}^*, t \vee s \leq C_i | g_i) = G(t \vee s) S^{(g_i)}(t \vee s, t \vee s), \\
E[d\mathcal{N}_{i1}(t)d\mathcal{N}_{i2}(s)|\mathcal{Y}_{i1}(t)\mathcal{Y}_{i2}(s)=1, g_i] &= \frac{\Pr(t \leq T_{i1}^* < t+dt, s \leq T_{i2}^* < s+ds, t \vee s \leq T_{i1}^*, t \vee s \leq T_{i2}^* | g_i)}{\Pr(t \vee s \leq T_{i1}^*, t \vee s \leq T_{i2}^* | g_i)} \\
&= (t=s) \frac{S^{(g_i)}(dt, ds)}{S^{(g_i)}(t \vee s, t \vee s)}, \\
E[\mathcal{Y}_{i1}(t)d\mathcal{N}_{i2}(s)|\mathcal{Y}_{i1}(t)\mathcal{Y}_{i2}(s)=1, g_i] &= \frac{\Pr(t \leq T_{i1}^*, s \leq T_{i2}^* < s+ds, t \vee s \leq T_{i1}^*, t \vee s \leq T_{i2}^* | g_i)}{\Pr(t \vee s \leq T_{i1}^*, t \vee s \leq T_{i2}^* | g_i)} \\
&= -(t \leq s) \frac{S^{(g_i)}(t \vee s, ds)}{S^{(g_i)}(t \vee s, t \vee s)}, \\
E[d\mathcal{N}_{i1}(t)\mathcal{Y}_{i2}(s)|\mathcal{Y}_{i1}(t)\mathcal{Y}_{i2}(s)=1, g_i] &= \frac{\Pr(t \leq T_{i1}^* < t+dt, s \leq T_{i2}^*, t \vee s \leq T_{i1}^*, t \vee s \leq T_{i2}^* | g_i)}{\Pr(t \vee s \leq T_{i1}^*, t \vee s \leq T_{i2}^* | g_i)} \\
&= -(t \geq s) \frac{S^{(g_i)}(dt, t \vee s)}{S^{(g_i)}(t \vee s, t \vee s)}.
\end{aligned}$$

Now, let us add an assumption natural for continuous time-to-event data that we cannot observe T_{i1}^* and T_{i2}^* simultaneously in both fatal event times. Then, the factors that occur at only $t = s$, such as $\mathbb{1}(t = s) S^{(j)}(dt, ds)$, do not contribute to the integral of the form

$\int_0^\tau \int_0^\tau (\cdot) dA^{(j)}(t, s)$, so that we can ignore $\mathbb{1}(t = s) S^{(j)}(dt, ds)$ in (A.3) as the zero term. We can also apply the relations $S^{(j)}(s, ds) = -S^{(j)}(s, s) \Lambda^{(j)}(s, ds)$ and $S^{(j)}(dt, t) = -S^{(j)}(t, t) \Lambda^{(j)}(dt, t)$ into (A.3) further. Hence, (A.3) becomes to zero under continuous time-to-event data. As a result, the form of $dA^{(j)}(t, s) = G(t \vee s)^{-1} E[dM_{j1}(t)dM_{j2}(s) | g_i = j]$ is obtained as (4a) in Section 2.3.

APPENDIX B. Probability formula for the observed two endpoints

We provide the formula of $P_{ab} = \Pr(\Delta_i^{(1)} = a, \Delta_i^{(2)} = b)$ for the observation $(\Delta_i^{(1)}, \Delta_i^{(2)})$ of the four patterns. We may monitor a trial based on the number of events obtained from the required sample size using the probability formula. Let $P_{ab}^{(j)} = \Pr(\Delta_i^{(1)} = a, \Delta_i^{(2)} = b | g_i = j)$, $a, b = 0, 1, j = 1, 2$. Using this notation, we can write

$$P_{ab} = r^{(1)} P_{ab}^{(1)} + r^{(2)} P_{ab}^{(2)}$$

and then, for example, in the one fatal case, we can obtain

$$P_{11}^{(j)} = Q_1^{(j)}, P_{10}^{(j)} = R_2^{(j)} - Q_1^{(j)}, P_{01}^{(j)} = R_1^{(j)}, \text{ and } P_{00}^{(j)} = 1 - R_1^{(j)} - R_2^{(j)}, j = 1, 2, \text{ where}$$

$$\begin{aligned}
Q_1^{(j)} &= \int_0^\infty \left(\int_0^s f^{(j)}(t, s) dt \right) G(s) ds, & Q_2^{(j)} &= \int_0^\infty \left(\int_0^t f^{(j)}(t, s) ds \right) G(t) dt, \\
R_1^{(j)} &= \int_0^\infty \left(\int_s^\infty f^{(j)}(t, s) dt \right) G(s) ds, & R_2^{(j)} &= \int_0^\infty \left(\int_t^\infty f^{(j)}(t, s) ds \right) G(t) dt,
\end{aligned}$$

$f^{(j)}(x, y) \left(= \frac{d^2}{dx dy} S^{(j)}(x, y) \right)$ is the density function of (T_{i1}^*, T_{i2}^*) for the i th participant assigned as $g_i = j$. The changes of $P_{ab}^{(j)}$ in the other censoring scenarios under the non-composite setting are summarized as Table BI including the one fatal case. For an example with the composite setting, let P_{ab}^{TTP} and P_{ab}^{PFS} be P_{ab} when the first endpoint corresponds to

the TTP and the PFS, respectively, and the second endpoint is consistently of the OS. Then we have $P_{11}^{\text{PFS}} = P_{11}^{\text{TTP}} + P_{01}^{\text{TTP}}$, $P_{10}^{\text{PFS}} = P_{10}^{\text{TTP}}$, $P_{00}^{\text{PFS}} = P_{00}^{\text{TTP}}$, and $P_{01}^{\text{PFS}} = 0$. It would be enough for us to consider P_{ab} under the non-composite setting as the probability formula, because it includes more information than P_{ab} under the composite setting.

Table BI

The probability formula on observing $(\Delta_i^{(1)}, \Delta_i^{(2)})$ in the three censoring scenarios under the non-composite setting

$P_{ab}^{(j)}$	Both non-fatal case (Non-competing model)	One fatal case (Semi-competing model)	Both fatal case (Full-competing model)
$P_{11}^{(j)}$	$Q_1^{(j)} + Q_2^{(j)}$	$Q_1^{(j)}$	0
$P_{10}^{(j)}$	$R_2^{(j)} - Q_1^{(j)}$	$R_2^{(j)} - Q_1^{(j)}$	$R_2^{(j)}$
$P_{01}^{(j)}$	$R_1^{(j)} - Q_2^{(j)}$	$R_1^{(j)}$	$R_1^{(j)}$
$P_{00}^{(j)}$	$1 - R_1^{(j)} - R_2^{(j)}$	$1 - R_1^{(j)} - R_2^{(j)}$	$1 - R_1^{(j)} - R_2^{(j)}$

Figure 1.

Behavior of the total sample sizes N_{mp} as a function of the correlation $\rho^{(j)}$ for $\psi_2^{-1}=1.3, 1.5$ and 1.7, arranged from the left for $S_2^{(1)}(\tau)=0.4, 0.5, 0.6$ and from the top following to $N_{mp}^{(nc)}$ from the both non-fatal case, $N_{mp}^{(nc)}$ and $N_{mp}^{(c)}$ of one fatal case and $N_{mp}^{(c)}$ of both fatal case, given $1 - \beta = 0.8$, $\alpha = 0.025$, $\tau_a = 2$, $\tau_f = 3$, $\psi_1^{-1}=1.7$, $S_1^{(1)}(\tau)=0.3$ and early dependency

Table I

Classification of event types by the composite and non-composite examples (six scenarios)

Type	Non-composite examples	Composite examples
Both non-fatal events	HIV trial <ul style="list-style-type: none"> Time to virologic failure Time to discontinuation of treatment due to toxicity 	HIV trial <ul style="list-style-type: none"> Time to virologic failure or to discontinuation of treatment due to toxicity
	Cardiovascular trial: <ul style="list-style-type: none"> Time to revascularization Time to major bleeding 	
	Oncology trial <ul style="list-style-type: none"> Time to progression (TTP) Time to all cause death (OS: overall survival) 	Oncology trial <ul style="list-style-type: none"> Time to progression or all cause death (PFS: progression free survival) Time to all cause death (OS)
One fatal event (one non-fatal event)	Cardiovascular trial <ul style="list-style-type: none"> Time to cardiovascular hospital admissions Time to all cause death (OS) 	Cardiovascular trial <ul style="list-style-type: none"> Time to myocardial infarction Time to major adverse cardiac events (MACE) (acute myocardial infarction, ischaemic stroke, coronary arterial occlusion, death)
Both fatal events	Cardiovascular trial <ul style="list-style-type: none"> Time to cardiac-specific death Time to the other cause death 	Cardiovascular trial <ul style="list-style-type: none"> Time to cardiac-specific death Time to all cause death (OS)

Table II

Elements of $U_k(\tau)$, $V_{kk}^{(0)}(\tau)$, $V_{kk}(\tau)$, and $V_{12}(\tau)$ for the three censoring scenarios under the non-composite setting. $t \vee s = \max(t, s)$.

Elements	Both non-fatal case	One fatal case	Both fatal case
k			
$d\tilde{\Lambda}_k^{(j)}(t)$	1	$\Lambda^{(j)}(dt, t)$	
	2		$\Lambda^{(j)}(t, dt)$
$y_k^{(j)}(t)$	$d\tilde{\Lambda}_k^{(j)}(t)$		
	1	$S^{(j)}(t, t)G(t)$	
	2		
$H_k(t)$	$S_k^{(j)}(t)G(t)$		
	1	$G(t) \frac{r^{(1)}r^{(2)}S^{(1)}(t,t)S^{(2)}(t,t)}{r^{(1)}S^{(1)}(t,t)+r^{(2)}S^{(2)}(t,t)}$	
	2		
$dA^{(j)}(t, s)$	$G(t) \frac{r^{(1)}r^{(2)}S_k^{(1)}(t)S_k^{(2)}(t)}{r^{(1)}S_k^{(1)}(t)+r^{(2)}S_k^{(2)}(t)}$		
	$S^{(j)}(dt, ds)$	$(t \leq s)S^{(j)}(dt, ds)$	0
	$+S^{(j)}(t, ds)d\Lambda_1^{(j)}(t)$	$+(t \leq s)S^{(j)}(t, ds)\Lambda^{(j)}(dt, t)$	
	$+S^{(j)}(dt, s)d\Lambda_2^{(j)}(s)$	$+S^{(j)}(dt, t \vee s)d\Lambda_2^{(j)}(s)$	
	$+S^{(j)}(t, s)d\Lambda_1^{(j)}(t)d\Lambda_2^{(j)}(s)$	$+S^{(j)}(t, t \vee s)\Lambda^{(j)}(dt, t)d\Lambda_2^{(j)}(s)$	

Table III

The case of one fatal outcome (ii) (semi-competing risk): total numbers of participants N_{cp} calculated from (8), the corresponding empirical powers $\tilde{P}_{cp}$ (%) and alternative sizing solutions N_{sim}^{-1} , $S_{TTP}^{(1)}$ (τ)=0.3 and $S_{OS}^{(1)}$ (τ)=0.4.

Dependence	Composite setting				Non-composite setting			
	Structure	ψ_{OS}^{-1}	$\rho^{(k)}$	N_{sim}	$N_{cp}^{(c)}$	$\tilde{P}_{cp}$	N_{PFS}	N_{OS}
Late (Clayton copula)		1.3	0.0	968	972	80.1	262	972
						80.1	262	972
		1.3	0.3	968	972	80.2	312	972
						80.2	312	972
		1.3	0.5	968	972	80.2	348	972
						80.2	348	972
		1.3	0.8	968	972	79.8	418	972
						79.8	418	972
		1.5	0.0	430	436	80.5	196	432
						80.5	196	432
Early (Gumbel copula)		1.5	0.3	436	440	80.6	232	432
						80.6	232	432
		1.5	0.5	440	444	80.6	260	432
						80.6	260	432
		1.5	0.8	450	456	80.5	316	432
						80.5	316	432
		1.7	0.0	272	278	81.2	160	266
						81.2	160	266
		1.7	0.3	280	286	80.9	190	266
						80.9	190	266
		1.7	0.5	288	294	81.0	214	266
						81.0	214	266
		1.7	0.8	308	314	80.9	262	266
						80.9	262	266
		1.3	0.0	968	972	80.2	262	972
						80.2	262	972
		1.3	0.3	968	972	80.3	282	972
						80.3	282	972
		1.3	0.5	968	972	80.3	294	972
						80.3	294	972
		1.3	0.8	970	972	80.1	306	972
						80.1	306	972
		1.5	0.0	432	436	80.4	196	432
						80.4	196	432
		1.5	0.3	430	434	80.5	212	432
						80.5	212	432
		1.5	0.5	430	434	80.5	224	432
						80.5	224	432
		1.5	0.8	428	432	80.3	238	432
						80.3	238	432
		1.7	0.0	272	278	81.0	160	266
						81.0	160	266
		1.7	0.3	274	278	80.9	176	266
						80.9	176	266
		1.7	0.5	274	278	80.8	186	266
						80.8	186	266
		1.7	0.8	274	278	80.8	206	266
						80.8	206	266
						81.2	280	266
						81.2	280	266

The case of both fatal outcomes (iii) (Competing risk): total numbers of participants N_{cp} calculated from (8), the corresponding empirical powers $\tilde{P}_{cp}$ (%) and alternative sizing solutions N_{sim} , N_1 and N_2 under $\psi_2^{-1}=1.7$ and $S_1^{(1)}(\tau)=S_2^{(1)}(\tau)=0.5$.

Dependence		Composite setting					Non-composite setting			
Structure	ψ_1^{-1}	$\rho^{(k)}$	$N_{\text{cp}}^{(c)}$	$\tilde{\mathbf{r}}_{\text{cp}}$	$N_1^{(c)}$	N_2	$N_{\text{cp}}^{(\text{nc})}$	$\tilde{\mathbf{r}}_{\text{cp}}$	$N_1^{(\text{nc})}$	N_2
Late (Clayton copula)	1.5	0	418	80.7	254	404	710	80.7	648	404
	1.5	0.3	474	80.8	294	456	834	80.6	774	456
	1.5	0.5	518	80.4	328	498	944	80.5	886	498
	1.5	0.8	626	80.7	414	592	1258	80.2	1211	592
	1.7	0	404	80.8	200	400	528	81.0	400	400
	1.7	0.3	466	80.8	232	462	608	81.0	462	462
Early (Gumbel copula)	1.7	0.5	518	80.6	258	514	676	80.8	514	514
	1.7	0.8	652	80.7	324	646	850	80.7	646	646
	1.5	0	418	80.8	254	404	710	80.8	648	404
	1.5	0.3	444	81.0	288	422	854	80.7	814	422
	1.5	0.5	452	80.9	312	422	1010	80.5	992	422
	1.5	0.8	436	81.0	360	364	1840	80.6	1840	364
	1.7	0	404	80.7	200	400	528	81.1	400	400
	1.7	0.3	458	80.8	228	454	598	81.3	454	454
	1.7	0.5	496	80.8	246	492	648	81.1	492	492
	1.7	0.8	566	81.0	282	562	740	81.2	562	562

Table V

Total sample size N_{cp} and the empirical power for detecting the joint reduction in the TTP or PFS and the OS, where N_{cp} is designed with the target power 80% at $\alpha = 0.025$, based on the assumption of ICON7 study, $\psi_{TTP}^{-1} = 1.31$, $\psi_{OS}^{-1} = 1.23$, $\tau_a = 24$ and $\tau_t = 36$ with $S_{TTP}^{(1)}(60) = 0.25$ and $S_{OS}^{(1)}(60) = 0.4$.

Dependence		Composite setting			Expected number of (TTP, OS)			
Structure	$\rho^{(k)}$	$N_{cp}^{(c)}$	$\tilde{P}_{cp}(\tilde{P}_{PFS}, \tilde{P}_{OS})$	N_{PFS}	N_{OS}	D_{11}	D_{10}	D_{01}
Late (Clayton copula)	0.0	1510	80.1 (99.0, 80.3)	658	1498	240	476	487
	0.3	1528	80.0 (97.4, 80.7)	789	1498	275	429	461
	0.5	1544	80.0 (96.0, 81.1)	881	1498	302	395	441
	0.8	1564	80.0 (93.4, 81.7)	1024	1498	370	316	383
Early (Gumbel copula)	0.0	1510	80.3 (99.0, 80.6)	658	1498	240	476	487
	0.3	1508	80.1 (98.5, 80.3)	700	1498	315	384	411
	0.5	1506	80.2 (98.2, 80.4)	726	1498	374	327	351
	0.8	1498	80.0 (97.9, 80.1)	744	1498	510	243	211
Dependence			Non-composite setting					
Structure	$\rho^{(k)}$	$N_{cp}^{(nc)}$	$\tilde{P}_{cp}(\tilde{P}_{TTP}, \tilde{P}_{OS})$	N_{TTP}	N_{OS}	D_{11}	D_{10}	D_{01}
Late (Clayton copula)	0.0	1628	80.3 (96.3, 83.3)	913	1498	259	514	525
	0.3	1674	80.2 (94.9, 84.2)	1023	1498	301	470	505
	0.5	1658	80.2 (94.1, 84.7)	1076	1498	332	434	484
	0.8	1658	80.3 (94.1, 84.1)	1056	1498	392	335	406
Early (Gumbel copula)	0.0	1628	80.2 (96.4, 83.3)	913	1498	259	514	525
	0.3	1594	80.1 (96.7, 82.3)	878	1498	333	406	435
	0.5	1562	80.2 (97.0, 81.8)	830	1498	388	339	364
	0.8	1510	80.0 (98.6, 80.2)	695	1498	514	245	213

$\tilde{P}_{TTP}$, $\tilde{P}_{PFS}$, and $\tilde{P}_{OS}$ are 100 times empirical powers (%) when single hypotheses on the TTP, PFS and OS are tested, given total sample sizes N_{cp} .

The case of one fatal outcome (ii) (semi-competing risk): total numbers of participants N_{mp} calculated from (11), the corresponding empirical powers $\tilde{P}_{\text{mp}}$ (%) and alternative sizing solutions N_{PFS} , N_{TTP} , and N_{OS} under $\psi_{\text{TTP}}^{-1} = 1.7$, $S_{\text{TTP}}^{(1)}(\tau) = 0.3$ and $S_{\text{OS}}^{(1)}(\tau) = 0.5$.

Table VI

Dependence		Composite setting						Non-composite setting					
structure	ψ_{OS}^{-1}	$\rho^{(k)}$	ω_0	$N_{mp}^{(c)}$	$\tilde{P}_{mp}$	N_{PFS}	N_{OS}	ω_b	$N_{mp}^{(nc)}$	$\tilde{P}_{mp}$	N_{TTP}	N_{OS}	
Late (Clayton copula)	1.3	0.0	1.00	254	80.7	254	1194	0.95	268	80.8	270	1194	
	1.3	0.3	1.00	284	80.5	284	1194	0.95	288	81.0	288	1194	
	1.3	0.5	1.00	322	80.9	322	1194	0.95	296	80.5	298	1194	
	1.3	0.8	1.00	366	80.8	366	1194	1.00	288	80.7	288	1194	
	1.5	0.0	1.00	200	80.9	200	532	0.80	242	80.2	266	532	
	1.5	0.3	1.00	232	80.7	232	532	0.80	266	80.4	292	532	
	1.5	0.5	1.00	256	80.5	256	532	0.80	282	80.5	306	532	
	1.5	0.8	1.00	300	80.9	300	532	0.80	296	80.5	312	532	
	1.7	0.0	1.00	168	81.1	168	328	0.55	208	80.4	264	328	
	1.7	0.3	1.00	196	80.9	196	328	0.55	228	80.7	296	328	
Early (Gumbel copula)	1.7	0.5	0.95	218	80.6	218	328	0.55	242	80.7	314	328	
	1.7	0.8	0.95	260	80.8	262	328	0.50	264	80.7	332	328	
	1.3	0.0	1.00	254	80.7	254	1194	0.95	268	80.9	270	1194	
	1.3	0.3	1.00	266	80.6	266	1194	1.00	246	81.0	246	1194	
	1.3	0.5	1.00	272	80.6	272	1194	1.00	222	80.8	222	1194	
	1.3	0.8	1.00	258	80.3	258	1194	1.00	174	81.0	174	1194	
	1.5	0.0	1.00	200	80.7	200	532	0.80	242	80.5	266	532	
	1.5	0.3	1.00	216	81.0	216	532	0.90	250	80.5	260	532	
	1.5	0.5	1.00	224	80.7	224	532	0.95	246	80.6	248	532	
	1.5	0.8	1.00	230	80.9	230	532	1.00	214	81.0	214	532	
	1.7	0.0	1.00	168	81.3	168	328	0.55	208	80.4	264	328	
	1.7	0.3	1.00	186	81.0	186	328	0.60	228	80.3	272	328	
	1.7	0.5	1.00	196	80.7	196	328	0.65	240	80.3	280	328	
	1.7	0.8	1.00	214	80.7	214	328	1.00	248	81.0	248	328	