Sample size and robust marginal methods for cluster-randomized trials with censored event times

YUJIE ZHONG

Department of Statistics and Actuarial Science,

University of Waterloo, Waterloo, ON, N2L 3G1, Canada

RICHARD J COOK

Department of Statistics and Actuarial Science,

University of Waterloo, Waterloo, ON, N2L 3G1, Canada

E-mail: rjcook@uwaterloo.ca

Summary

In cluster-randomized trials, intervention effects are often formulated by specifying marginal models, fitting them under a working independence assumption, and using robust variance estimates to address the association in the responses within clusters. We develop sample size criteria within this framework, with analyses based on semiparametric Cox regression models fitted with event times subject to right censoring. At the design stage, copula models are specified to enable derivation of the asymptotic variance of estimators from a marginal Cox regression model and to compute the number of clusters necessary to satisfy power requirements. Simulation studies demonstrate the validity of the sample size formula in finite samples for a range of cluster sizes, censoring rates and degrees of within-cluster association among event times. The power and relative efficiency implications of copula misspecification is studied, as well as the effect of within-cluster dependence in the censoring times. Sample size criteria and other design issues are also addressed for the setting where the event status is only ascertained at periodic assessments and times are interval censored.

Keywords: censored data; cluster-randomized trials; copula models; robust inference; sample size

1 INTRODUCTION

Cluster randomization is employed in clinical trials when it is appropriate on ethical [1], practical [2], or contextual [3] grounds to assign groups of individuals (e.g. families, schools, hospitals, or communities) to receive one of two or more interventions. In studies aiming to reduce the spread of infectious disease, for example, prevention strategies are most naturally administered to large groups of individuals (e.g. municipalities), and the resulting evidence of impact thereby reflects direct effects

(susceptibility), indirect effects (infectiousness of others), and the effect of herd immunity [4]. Cluster randomization also offers a way of minimizing contamination across treatment groups and can often enhance compliance [5, 6]. In some fields of research, the units providing the response are paired or otherwise grouped, as is the case in ophthalmology or audiology. In such settings, interventions that are administered and act through the blood stream (e.g. medications) necessitate randomizing individuals, and the units providing the response are clustered within individuals. The many advantages of cluster randomization have led to its increased use in recent years in diverse areas of research including health promotion [7], education for disease management [8], clinical research [9], and health policy and program evaluation [10]. Donner and Klar [11] give a thorough account of the practical and methodological issues in the conduct of cluster-randomized trials.

While much of the methodological work on cluster-randomized trials to date has been for continuous or binary responses, in many settings interest lies in evaluating the effect of an intervention in delaying or preventing the occurrence of an event. In patients with insulin-dependent diabetes mellitus, for example, interest may lie in the effect of medical therapies on the time to severe vision loss in each eye [12]; such times are correlated within individuals due to shared exposure to blood sugar levels, blood pressures, and other systemic features. Chronic otitis media is a condition arising in children characterized by poor drainage of fluid from the inner ear. A common intervention involves the surgical insertion of a ventilating tube and interest then may lie in assessing an experimental post-surgery medical therapy designed to prolong the function of the ventilating tubes. The child is then the unit of randomization [13, 14], and the times to failure of the tubes in the left and right ears would naturally be correlated. Settings involving time to event responses with larger cluster sizes include studies of fall prevention in retirement homes [15], studies of primary care practices and survival in patients with depression [16], and studies of pediatric clinics and time to discontinuation of breast-feeding [17].

Cox regression models involving random effects, or frailty terms, are widely used for analysing correlated time to event data [18, 19]. In this framework failure times are typically considered to be independent conditional on a latent variable representing unexplained differences between clusters, and the association among responses within clusters arises by marginalizing over the random effects. There are several important limitations of this approach for the analysis of event times in cluster-randomized trials. First, specification of a proportional hazard model given cluster-level random effects is unappealing when the treatment indicator is fixed at the cluster level [20, 21]. Second, regression coefficients reflect the multiplicative effects of the intervention, conditional on the latent variable; the proportional hazards assumption does not hold in the marginal model obtained by integrating out the random effects, making the interpretation of the intervention effect challenging. Third, while the dependence within clusters is accommodated in the marginal joint distribution, the association is not modeled in an appealing way. Simple measures of within-cluster dependence do not in general arise from the random effects formulation with censored failure time data, so it is difficult to extract useful information for the design of future similar trials.

Methods involving intervention effects specified on the basis of marginal Cox models feature none of these limitations and are therefore much more appealing. For large numbers of small groups of correlated failure time, Lee et al. [12] developed very useful methods for robust inference about regression coefficients in marginal Cox models fitted under a "working independence" assumption, similar in spirit to the working independence assumption adopted when clustered categorical data are analysed via generalized estimating equations [22] or when multivariate failure time data are analysed by the marginal approach of Wei et al. [23, 24]. Robust "sandwich" variance estimates provided by Lee et al. [12] ensure valid inference when there is within-cluster dependence in event times. The simple marginal interpretation of intervention effects and use of robust variance estimation make this a useful and simple framework for the analysis of event times in cluster-randomized trials.

A considerable amount of attention has been directed at the development of sample size formu-

lae for the cluster-randomized trials with continuous and discrete outcomes [5, 25–28], but relatively little work has been done for trials involving censored event times; in what follows, the term sample size is used to mean the number of clusters. Jahn-Eimermacher et al. [29] developed sample size criteria on the basis of a frailty model for the within-cluster dependence, but as mentioned earlier the frailty approach is unappealing for use in cluster-randomized trials. Manatunga and Chen [14] derived sample size formula for bivariate event times under a parametric proportional hazards model with exponential margins. Jung [30] proposed a simulation-based sample size calculation procedure involving a weighted rank test for clustered survival data, which allows variable cluster size. Moerbeek [31] studied the effect of sample size on precision of parameter estimates and statistical power for clustered randomized trials with discrete event times based on a generalized linear mixed model. Xie and Waksman [32] adapted the usual sample size criteria for log-rank tests by the introduction of a design effect involving the average cluster size and the intraclass correlation coefficient of the censoring (i.e., status) indicator of the response times. While the formula is relatively simple, the sample size criterion is based on an approximation of the asymptotic distribution of regression coefficients. More importantly, because the intraclass correlation coefficient in their design effect is for the censoring indicators rather than the underlying failure times, its magnitude is driven by both the dependence in the failure times within clusters and the within-cluster dependence in the censoring times. As a result, the event times may be independent within clusters, for example, but the censoring indicators may be highly correlated within clusters if the censoring times are dependent. Moreover, the correlation in the censoring indicators depends on both the administrative censoring time and the distribution of the random censoring time, so any plans to modify a study by extending follow-up or attempting to reduce loss to follow-up will render the measure of within-cluster dependence invalid.

We derive sample size criteria for cluster-randomized trials with censored time to event responses when the intervention effect is specified through a marginal semiparametric proportional hazards model fitted under a working independence assumption, and robust variance estimates are used as in Lee et al. [12]. Of course at the design stage a fully parametric model is required so a Weibull proportional hazard model is adopted to accommodate trend in the marginal hazard. Within-cluster dependence is conveniently modeled using copula functions [33, 34] as intervention effects may be specified in terms of the marginal distributions and within-cluster dependence is modeled by a separate association parameter. The resulting joint model is used to evaluate the components of the robust variance formula [12] for a variety of practical settings, and our approach does not involve any approximations apart from the usual ones used in large sample theory. We also study the effect of copula misspecification and the impact of within-cluster dependence in the random right censoring times. Sample size criteria are also developed for cluster-randomized trials with interval-censored event times that arise when the events are only detectable upon periodic inspection (e.g., radiographic examination, based on blood tests, urinalysis, etc.).

The remainder of this paper is organized as follows. In Section 2 we define notation and review copula models and the robust marginal method of Lee et al. [12]. The asymptotic distribution of the test statistic is then derived to facilitate the development of sample size criteria, and simulation studies are carried out to validate the derivations. In Section 3 we explore the impact of misspecification of the copula function and the impact of within-cluster dependence in the censoring times. Design criteria for cluster-randomized trials with type II interval-censored failure times are developed in Section 4. Section 5 contains an illustrative example, and concluding remarks and topics for future research are given in Section 6.

2 SAMPLE SIZE FOR TRIALS WITH CLUSTERED EVENT TIMES SUBJECT TO RIGHT-CENSORING

2.1 NOTATION AND ROBUST MARGINAL METHODS

We consider the setting in which n clusters, each comprised of J individuals, are randomly assigned to receive either an experimental or standard intervention. We let T_{ij} denote an event time of interest for individual j in cluster i, $j = 1, \ldots, J$, $i = 1, \ldots, n$, and assume interest lies in examining the effect of the experimental intervention by fitting a Cox regression model. Let Z_i be a binary covariate where $Z_i = 1$ indicates that cluster i is assigned to the experimental intervention and $Z_i = 0$ otherwise; we let $P(Z_i = 1) = p$. It is possible to generalize the methods that follow to accommodate a $p \times 1$ cluster-level covariate vector as we discuss in Section 6.

Suppose the plan is to observe individuals over the interval $(0, C^\dagger]$ where $C^\dagger$ is an administrative censoring time, and let C_{ij}^* denote a random (possibly latent) time of withdrawal from the study for individual j in cluster i with survivor function $\mathcal{G}^*(s) = P(C_{ij}^* \geq s)$. Then $C_{ij} = \min(C_{ij}^*, C^\dagger)$ denotes the resultant right-censoring time. We then let $X_{ij} = \min(T_{ij}, C_{ij})$, $Y_{ij}^\dagger(t) = \mathrm{I}(t \leq T_{ij})$, $Y_{ij}(t) = \mathrm{I}(t \leq C_{ij})$, and $\bar{Y}_{ij}(t) = Y_{ij}(t) Y_{ij}^\dagger(t)$ where the latter indicates individual j in cluster i is under observation and at risk of the event at time t; thus here, and in what follows, quantities with a vinculum (overbar) are observable in the presence of right censoring. Let $N_{ij}(t) = I(T_{ij} \leq t)$ indicate that individual j in cluster i experienced the event at or before time t, and $dN_{ij}(t) = I(T_{ij} = t)$. When viewed as a random function of time, $\{N_{ij}(s), 0 < s\}$ is a right-continuous stochastic process. If $d\bar{N}_{ij}(t) = \bar{Y}_{ij}(t) dN_{ij}(t)$ and $\bar{N}_{ij}(t) = \int_0^t d\bar{N}_{ij}(s)$, then $\{\bar{N}_{ij}(s), 0 < s\}$ is the observed counting process for individual j in cluster i. Finally we let $\bar{N}_i(t) = (\bar{N}_{i1}(t), \ldots, \bar{N}_{iJ}(t))'$, $\bar{Y}_i(t) = (\bar{Y}_{i1}(t), \ldots, \bar{Y}_{iJ}(t))'$ and let $\{\bar{Y}_i(\cdot), \bar{N}_i(\cdot), Z_i\}$ denote the data from cluster i.

Marginal proportional hazard models are based on the assumption that given Z_i, T_{ij} has a hazard function of the form

$$\lambda_{ij}(t|Z_i) = \lambda_0(t; \alpha) \exp(Z_i \beta) \tag{2.1}$$

where $\lambda_0(t; \alpha)$ is a baseline hazard function indexed by a vector of parameters α, and β is a scalar regression coefficient; let $\theta = (\alpha', \beta)'$. The marginal Cox regression model is obtained by leaving $\lambda_0(t; \alpha)$ of an unspecified form, making it a semiparametric model.

Lee et al. [12] considered the semiparametric Cox model and proposed estimation of β under a working independence assumption by which observations in each cluster are treated as independent of one another. This gives a partial score function for β, written as $U(\beta) = \sum_{i=1}^n U_i(\beta)$, where

$$U_i(\beta) = \sum_{j=1}^J \int_0^\infty \left\{ Z_i - \frac{S_1(t; \beta)}{S_0(t; \beta)} \right\} d\bar{N}_{ij}(t) , \tag{2.2}$$

with $S_r(t; \beta) = \sum_{j=1}^J S_{rj}(t; \beta)$, $S_{rj}(t; \beta) = n^{-1} \sum_{i=1}^n \bar{Y}_{ij}(t) Z_i^r \exp(Z_i \beta)$, $r = 0, 1$, $Z_i^0 = 1$ and $Z_i^1 = Z_i$; the root of $U(\beta) = 0$ is $\widehat{\beta}$, the estimate.

If the marginal Cox regression model is correctly specified, $n^{-1/2} U(\beta)$ is asymptotically normally distributed with mean zero and variance [12]

$$\mathcal{B} = E[U_i^2(\beta)] , \tag{2.3}$$

estimated by

$$\widehat{B} = \frac{1}{n} \sum_{i=1}^n U_i^2(\beta) \bigg|_{\beta = \widehat{\beta}} .$$

Lee et al. [12] showed that $\widehat{\beta}$ is consistent with $n^{1/2}(\widehat{\beta} - \beta) \xrightarrow{D} N(0, \Gamma)$ asymptotically, where $\Gamma = \mathcal{B}/\mathcal{A}^2$ and

$$\mathcal{A} = -E[\partial U_i(\beta)/\partial \beta] . \tag{2.4}$$

Note that (2.4) can be consistently estimated by

$$\widehat{A} = -\frac{1}{n} \sum_{i=1}^{n} \partial U_i(\beta)/\partial \beta \bigg|_{\beta=\widehat{\beta}} ,$$

and so robust inferences are based on $\widehat{\Gamma} = \widehat{B}/\widehat{A}^2$ for a given sample.

2.2 SAMPLE SIZE CALCULATIONS VIA COPULA MODELS FOR CLUSTERED FAILURE TIMES

While the robust analyses based on marginal Cox models in the previous section can be carried out once data are collected, model assumptions are required to derive the sample size (number of clusters) on the basis of large sample theory. In the context of clustered event time data, copula functions offer a convenient way of constructing joint distributions with proportional marginal hazards [33, 34]. In what follows, we use J to denote the dimension of the multivariate vector to coincide with the size of the clusters in the previous section.

A copula function in J dimensions is a multivariate distribution on $[0, 1]^J$ whose margins are uniform over $[0, 1]$ [34]. Thus, for a $J-$dimensional uniform random vector $U = (U_1, \ldots, U_J)'$, the joint probability function

$$\mathcal{C}(u_1, \ldots, u_J; \phi) = P(U_1 \leq u_1, \ldots, U_J \leq u_J; \phi) ,$$

defines a copula indexed by the parameter ϕ. The family of Archimedean copulas [35] can be written as

$$\mathcal{C}(u_1, \ldots, u_J; \phi) = \mathcal{H}^{-1} \left(\mathcal{H}(u_1; \phi) + \cdots + \mathcal{H}(u_J; \phi); \phi \right) ,$$

where $\mathcal{H} : [0, 1] \to [0, \infty)$ is a continuous, strictly decreasing and convex generator function satisfying $\mathcal{H}(1; \phi) = 0$. Kendall's τ, a widely used measure of association with event time data, can be written as

$$\tau = 1 + 4 \int_0^1 \frac{\mathcal{H}(u; \phi)}{\mathcal{H}'(u; \phi)} du$$

for Archimedean copulas.

If $T_i = (T_{i1}, \ldots, T_{iJ})'$ is a $J \times 1$ vector of failure times, a joint model for $T_i|Z_i$ is obtained via the probability integral transforms $U_{ij} = \mathcal{F}(T_{ij}|Z_i; \theta)$, $j = 1, \ldots, J$, and linking all marginal survivor functions via the copula as

$$\mathcal{F}(t_i|Z_i; \psi) = P(T_{i1} > t_{i1}, \ldots, T_{iJ} > t_{iJ}|Z_i; \psi) = \mathcal{C}(\mathcal{F}(t_{i1}|Z_i; \theta), \ldots, \mathcal{F}(t_{iJ}|Z_i; \theta); \phi) , \tag{2.5}$$

where $\mathcal{F}(\cdot|Z_i; \theta)$ is the survivor function for T_{ij} given the covariate Z_i and $\psi = (\theta', \phi)'$. Because Kendall's τ is invariant to monotonic transformations, it also measures the association between the event times defined by the conditional (given Z_i) probability integral transform [35].

The Clayton copula is widely used in survival analysis and has generator function $\mathcal{H}(u; \phi) = \phi^{-1}(u^{-\phi} - 1)$, and then yields a joint survivor function for $T_i|Z_i$ of the form

$$\mathcal{F}(t_i|Z_i; \psi) = \left(\mathcal{F}(t_{i1}|Z_i; \theta)^{-\phi} + \cdots + \mathcal{F}(t_{iJ}|Z_i; \theta)^{-\phi} - (J - 1) \right)^{-1/\phi} . \tag{2.6}$$

The Frank copula with generator $\mathcal{H}(u; \phi) = -\log((\exp(-\phi u) - 1)/(\exp(-\phi) - 1))$ and the Gumbel copula with generator $\mathcal{H}(u; \phi) = (-\log u)^\phi$ are two other members of the Archimedean family that we consider shortly.

Returning to the issue of sample size determination, we consider the null and alternative hypotheses $H_0 : \beta = \beta_0 = 0$ and $H_A : \beta \neq \beta_0$, respectively, and let β_A denotes the clinically important effect of interest. Under a two-sided test at the γ_1 level of significance, the number of clusters required to ensure $1 - \gamma_2$ power to reject H_0 at β_A can be determined on the basis of a Wald test. The asymptotic robust variance of this Wald statistic involves the variance of the score statistic $\mathcal{B}$ and the information $\mathcal{A}$. To derive the expressions for these two quantities (2.3) and (2.4), we evaluate their asymptotic expressions under a fully specified parametric model at the design stage. The variance of the score statistic also depends on the within-cluster association of failure times and the form of the joint distribution is implied by the copula function (2.5). Explicit expressions for (2.3) and (2.4) are given in (A.9) and (A.10) of Appendix A. Note that (A.9) is derived for a more general case, in which censoring times are also correlated within clusters, but if we further assume independent within-cluster censoring times, then (A.13) can be used instead. Let $\Gamma = \mathcal{B}/\mathcal{A}^2$ denote the asymptotic variance of the estimator $\widehat{\beta}$, then the required sample size (number of clusters) is

$$
n \geq \left\{ \frac{z_{\gamma_1/2}\sqrt{\Gamma_0} + z_{\gamma_2}\sqrt{\Gamma_A}}{\beta_A} \right\}^2 \tag{2.7}
$$

where z_u is the $100(1 - u)\%$ percentile of the standard normal distribution and Γ_0 and Γ_A are the asymptotic variances of $\hat{\beta}$ evaluated under the null and alternative hypotheses.

2.3 Empirical Validation of Sample Size Formula under Correct Model Specification

We consider a two-arm cluster-randomized trial with equal allocation probabilities where the binary treatment indicator takes the value $Z_i = 1$ if cluster i is randomized to the experimental intervention and $Z_i = 0$ otherwise; and $P(Z_i = 1) = P(Z_i = 0) = 0.5$. We assume $T_{ij}|Z_i$ has a proportional hazards structure as in (2.1), where the cumulative baseline hazard is of a Weibull form with $\Lambda_0(t; \alpha) = \int_0^t \lambda_0(s; \alpha)ds = (\lambda_0 t)^\kappa$ and $\alpha = (\lambda_0, \kappa)'$. The parameter κ accommodates a decreasing ($\kappa < 1$), constant ($\kappa = 1$) or increasing ($\kappa > 1$) hazard; here we focus primarily on the case with $\kappa = 0.75$ to reflect modest decreasing trend in risk. If the plan is to observe individuals over $(0, C^\dagger]$, without loss of generality we let $C^\dagger = 1$ denote the administrative censoring time. The parameter λ_0 is then chosen as the solution to $P(T_{ij} > C^\dagger|Z_i = 0) = p_a$ to give the desired administrative censoring rate for the control group, where $p_a = 0.2$. A random censoring time for the jth individual in cluster i is denoted by C_{ij}^* and assumed to be exponentially distributed with rate ρ; we assume here that $C_{ij} \perp C_{ik}|Z_i$ so censoring is independent within clusters. The effective right-censoring time is then $C_{ij} = \min(C_{ij}^*, C^\dagger)$, and the value ρ that solves $P(T_{ij} > C_{ij}|Z_i = 0) = p_0$ gives p_0, the net censoring rate in the control arm; we consider $p_0 = 0.2$ to correspond to the case of strictly administrative censoring and $p_0 = 0.5$ to correspond to the case of 30% random and 20% administrative censoring.

Suppose the within-cluster association in the failure time is induced by the Clayton copula with parameter ϕ, so the joint survivor function for $T_i = (T_{i1}, \ldots, T_{iJ})'$ is given by (2.6), where J is the cluster size. The copula parameter is chosen to give Kendall's τ of 0.05, 0.1, and 0.25 for small, mild and moderate within-cluster associations, respectively. We consider cluster sizes of $J = 2, 5, 20,$ and 100 that represent from small to large cluster sizes. For each parameter combination, we compute the required number of clusters (n) on the basis of (2.7) to give power $1 - \gamma_2 = 0.8$ using a two-sided test with a type I error rate $\gamma_1 = 0.05$. We then generate the corresponding clustered event times and (independent) censoring times, fit the marginal Cox model, and obtain the robust variance estimate derived

by Lee et al. [12] to test the null hypothesis of no treatment effect. We report empirical standard error (ESE) and average robust standard error (ASE) for $\hat{\beta}$, empirical rejection rate (REJ%) defined as the percentage of samples in which the null hypothesis $H_0 : \beta = 0$ is rejected by a two-sided Wald test at the nominal 5% level, and the empirical coverage probability (ECP%) of nominal 95% confidence intervals for β (the proportion of simulated samples for which the nominal 95% confidence interval contained the true value of β). Because the ECP% is the complement of the REJ% when $\beta = 0$, we do not report it in this case; see Table 1.

For each parameter configuration we generate 2000 samples, so the half-width of a 95% confidence interval for the type I error rate would be approximately $1.96(0.05 \times 0.95/2000)^{1/2} = 0.01$, and one could expect the REJ% to fall outside the range $[0.04, 0.06]$ in one out of 20 settings by chance; by similar arguments, one would expect the ECP% to fall within the range 94 and 96% 19 times out of 20. If the nominal power 0.80 is correct, then the empirical power would be expected to fall outside the range $[0.78, 0.82]$ for one out of every 20 configurations. From Table 1, it is apparent that the REJ% under $\beta = 0$ are within the acceptable range for most cases. Under the alternative hypothesis, the empirical coverage probabilities are within the acceptable range of 94-96%, and the REJ% are broadly compatible with the nominal level. All of these findings support the validity of the derived sample size formula. Similar results are seen when the baseline hazard is increasing ($\kappa = 1.25$) or constant ($\kappa = 1.0$); see Section S2 of the Supporting information.

3   ASYMPTOTIC CALCULATIONS INVESTIGATING DESIGN ROBUSTNESS AND RELATIVE EFFICIENCY

3.1   ROBUSTNESS OF POWER TO MISSPECIFICATION OF THE COPULA FUNCTION

Choosing a suitable copula at the design stage is challenging, so here we explore the sensitivity of study power to misspecified copula functions. We consider the same parameter configurations as in Section 2, where $\kappa = 0.75$ and the administrative censoring rate is $p_a = 0.2$. The sample size is estimated under the Clayton copula with Kendall's $\tau = 0.1$ and 0.25 with $\beta_A = \log 0.8$. Under the derived number of clusters, we construct the corresponding power curves under the Frank or Gumbel copula functions with the same value of Kendall's τ. Figure 1 shows these power curves for different copula functions for $J = 20$ under different net censoring rates ($p_0 = 0.2, 0.5$ and 0.7) in the control arm. When the censoring rate is mild and due strictly to administrative censoring ($p_0 = 0.2$), misspecification of the copula function impacts power, but use of the Clayton copula ensures power is maintained under the Frank or Gumbel copula functions. When the net censoring rate increases to 50%, the impact of copula misspecification is negligible, however, when the net censoring rate increases to 70%, the impact on power is again appreciable; in this case, the Clayton copula leads to samples sizes which are too small. These findings suggest that the misspecification of copula functions can have significant impact on study power and the impact depends on the censoring rate. The findings are broadly similar for cluster sizes of $J = 2, 5$ and 100.

To examine the effect of copula misspecification more fully we next consider the asymptotic relative efficiencies of the estimators through the functions

$$ARE_{F:C} = \frac{\text{asvar}_F(\hat{\beta})}{\text{asvar}_C(\hat{\beta})} \ , \quad ARE_{G:C} = \frac{\text{asvar}_G(\hat{\beta})}{\text{asvar}_C(\hat{\beta})} \ , \quad \text{and} \quad ARE_{F:G} = \frac{\text{asvar}_F(\hat{\beta})}{\text{asvar}_G(\hat{\beta})} \ , \qquad (3.1)$$

where *asvar()* denotes an asymptotic variance and 'C', 'F', and 'G' denote the Clayton, Frank and Gumbel copulas, respectively. We set $\kappa = 0.75$ and $\beta = \log 0.8$ and set the control administrative censoring rate to $p_a = 0.2$ at $C^\dagger = 1$; again λ_0 is found to satisfy $P(T_{ij} > C^\dagger | Z_{ij} = 0) = p_a$. The random censoring times are assumed to be independently exponentially distributed with rate ρ, which is selected to ensure a net censoring rate for the control arm through the constraint $P(T_{ij} >$

Table 1: Sample size estimation and empirical properties of estimators under cluster-randomized designs when within-cluster association between event times is induced by the Clayton copula; $\kappa = 0.75$, $\beta_A = \log 0.8$, $p_a = 0.2$, nsim= 2000.

| | | 20% Censoring ($p_0 = 0.2$) | | | | | | | | 50% Censoring ($p_0 = 0.5$) | | | | | | | |
| | | $\beta = 0$ | | | | $\beta = \beta_A$ | | | | $\beta = 0$ | | | | $\beta = \beta_A$ | | | |
J	τ	n	ESE	ASE	REJ%	ESE	ASE	ECP%	REJ%	n	ESE	ASE	REJ%	ESE	ASE	ECP%	REJ%
2	0.05	433	0.078	0.079	4.8	0.080	0.081	95.2	80.3	677	0.078	0.079	5.1	0.083	0.081	94.8	78.0
	0.10	464	0.080	0.079	5.4	0.081	0.081	95.0	79.0	708	0.078	0.079	4.5	0.080	0.081	95.9	79.0
	0.25	556	0.077	0.079	5.1	0.081	0.081	94.2	79.8	803	0.079	0.079	4.8	0.081	0.081	94.5	78.8
5	0.05	211	0.081	0.079	5.2	0.083	0.080	94.1	78.0	309	0.079	0.079	5.3	0.081	0.081	95.0	78.4
	0.10	262	0.080	0.079	5.4	0.083	0.080	94.0	79.0	359	0.079	0.079	4.4	0.082	0.081	95.0	79.7
	0.25	409	0.080	0.079	5.0	0.079	0.080	94.8	79.2	511	0.081	0.079	5.5	0.081	0.081	94.8	79.1
20	0.05	100	0.080	0.078	5.4	0.080	0.079	95.0	78.3	125	0.080	0.078	5.8	0.082	0.080	94.4	80.2
	0.10	160	0.079	0.079	4.9	0.078	0.079	95.2	81.0	185	0.077	0.079	4.7	0.082	0.080	94.5	78.5
	0.25	335	0.081	0.079	5.7	0.081	0.080	95.0	81.5	365	0.078	0.079	4.8	0.079	0.080	95.0	81.8
100	0.05	71	0.079	0.078	5.8	0.080	0.078	94.1	81.2	76	0.080	0.078	6.0	0.080	0.078	94.1	80.7
	0.10	133	0.079	0.079	4.9	0.081	0.079	94.1	79.0	139	0.079	0.079	4.6	0.080	0.079	94.3	81.6
	0.25	316	0.081	0.079	6.0	0.080	0.080	94.9	80.0	326	0.081	0.079	5.8	0.080	0.080	95.3	79.8
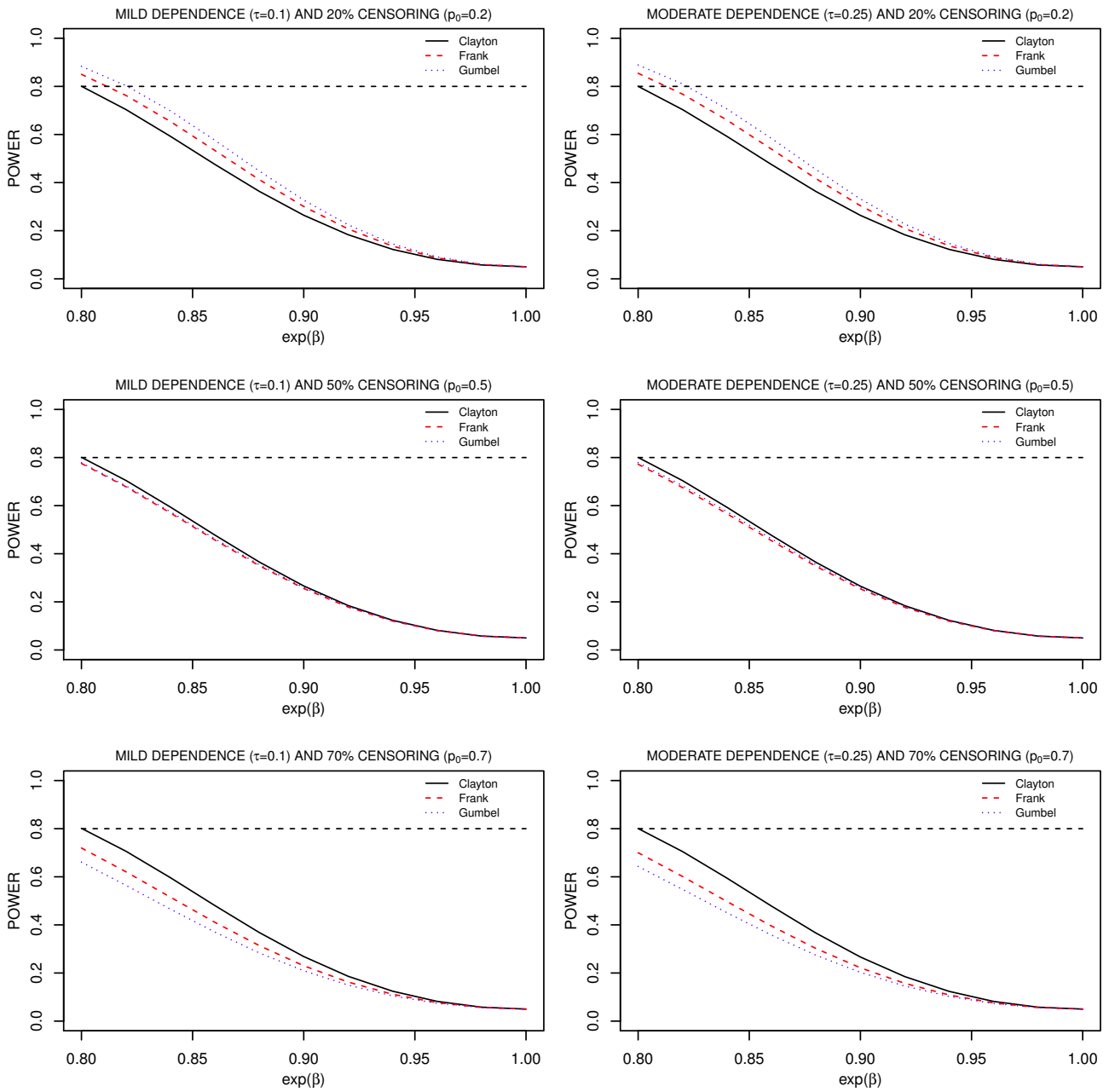
Figure 1: Power curves for different copula functions when sample size is estimated on the basis of the the Clayton copula with $\tau = 0.10$ (left column) and $\tau = 0.25$ (right column) under 20% (top row), 50% (middle row) and 70% (bottom row) net censoring for the control arm; $\kappa = 0.75$, $\beta_A = \log 0.8$, $p_a = 0.2$, $J = 20$.

$C_{ij}|Z_i = 0) = p_0$, where p_0 ranges from 0.2 to 0.8. Figure 2 displays the contour plots of the asymptotic relative efficiencies in (3.1) as a function of the degree of within-cluster association in the event times (Kendall's τ) and the net censoring rate (p_0) for both $J = 20$ (left panels) and $J = 100$ (right panels); we restrict attention to values of Kendall's τ ranging from 0 to 0.4 to cover realistic scenarios. For $J = 20$, if the net censoring rate is less than 40%, the Gumbel copula leads to a more efficient estimator, followed by the Frank copula and then the Clayton copula; the Clayton copula should therefore be used for the sample size calculations to ensure adequate power among this set of copulas. If the net censoring rate in this setting is higher than 40-50%, the Gumbel copula should be adopted at the design stage since it yields the estimator with the greater variance. The trend is broadly similar for $J = 100$.
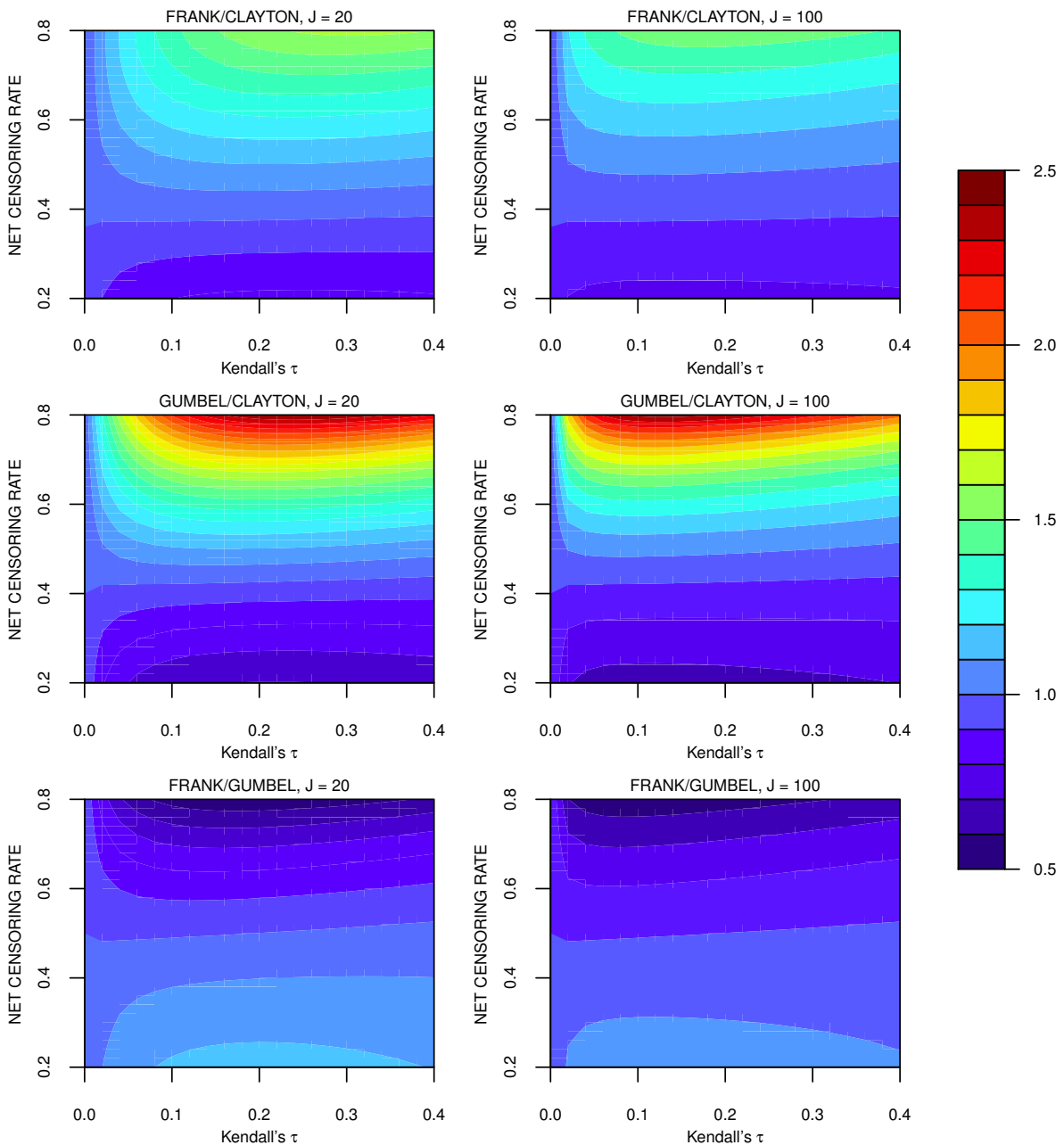

Figure 2: Contour plots of the asymptotic relative efficiencies in (3.1) for estimators defined as the solution to (2.2) when clustered failure times are generated on the basis of different copula functions; $\kappa = 0.75$, $\beta = \log 0.8$, $p_a = 0.2$.

Both Figure 1 and Figure 2 show that the proposed formula for calculating required sample size

is sensitive to both the copula function and censoring rate. A simple pragmatic approach to deal with this sensitivity is to consider a class of copula functions and a range of administrative and random censoring rates. The required sample sizes can be computed for each configuration by (2.7), and the largest sample size can then be chosen to ensure the pre-specified power requirements are met for any copula model and censoring pattern among those considered. Of course there may also be uncertainty in the value of Kendall's τ, but in this case the largest plausible value of τ will lead to the largest sample size within a given copula family and at a given censoring rate.

3.2   IMPACT OF WITHIN-CLUSTER DEPENDENCE IN THE RANDOM CENSORING TIMES

Although the assumption of independent censoring times within clusters is commonly, the factors inducing the association in the failure times within clusters may also induce an association in the censoring times. Here we examine the impact of within-cluster dependence in the censoring times on study power. We consider a trial designed to have 80% power to detect $\beta = \log 0.8$ on the basis of a two-sided Wald test at the 5% significance level under the assumption that random censoring times are independent within clusters and a Clayton copula model is used for the response. In this case, the minimal required sample size is estimated under the within-cluster independent censoring assumption (2.7) in which (A.13) is used to compute $\mathcal{B}$. We then calculate the theoretical power when the within-cluster censoring times are correlated, and (A.9) is used to compute $\mathcal{B}$. We let $\kappa = 0.75$, $p_a = 0.2$, and consider $J = 2$ and $J = 20$ with net censoring rates ranging from 0.2 to 0.8. The Clayton, Frank and Gumbel copula functions are considered for jointly modeling the distribution of the censoring times within clusters. While it is more general to allow different degrees of within-cluster associations for the failure and censoring times, for parsimony we restrict attention to the case that the value of Kendall's τ is the same for the failure times (τ) and censoring times (τ_c).

Figure 3 suggests that the naive assumption of within-cluster independence in the censoring times can lead to sample sizes which are too small and hence studies with inadequate power. As the net censoring rate increases (and hence the proportion of event times censored by the random censoring time increases) this effect becomes more pronounced. For example, for $J = 20$ and $\tau = 0.25$, the power is 0.8 for all the copula functions when $p_0 = 0.2$ as in this case there is no dependent random censoring time. However, when the net censoring rate increases to 80%, the power decreases to 0.756, 0.765 and 0.766 under the Clayton, Frank and Gumbel copula models for the censoring times. Further, if we compare the left panel to the right panels of Figure 3, we find that when the association in the censoring times increases, the power implications of ignoring the within-cluster dependence become more serious for all copula functions. The power is also more seriously impacted with larger cluster sizes; compare the top panels to the respective bottom panels of Figure 3.

A reviewer suggested examining the effect of misspecifying the shape of the baseline hazard function in the marginal event time distribution in the setting where the administrative and random censoring rates are correct; this ensures that the expected number of events is comparable in the assumed and true parameter settings but would mean, naturally, that the times of the events would be misspecified. The details on how this was investigated, along with the associated findings, are given in Section S3 of the Supporting information. We find that there is negligible impact on power of misspecifying the shape parameter in this setting when there is only administrative censoring. When the event times are subject to random censoring, there can be an increase or decrease in the power compared to the nominal level, and the extent of the effect depends on the copula function modeling the within-cluster dependence; this is not surprising as it is well known that the different copula functions model the association between event times differently over the range of possible values.
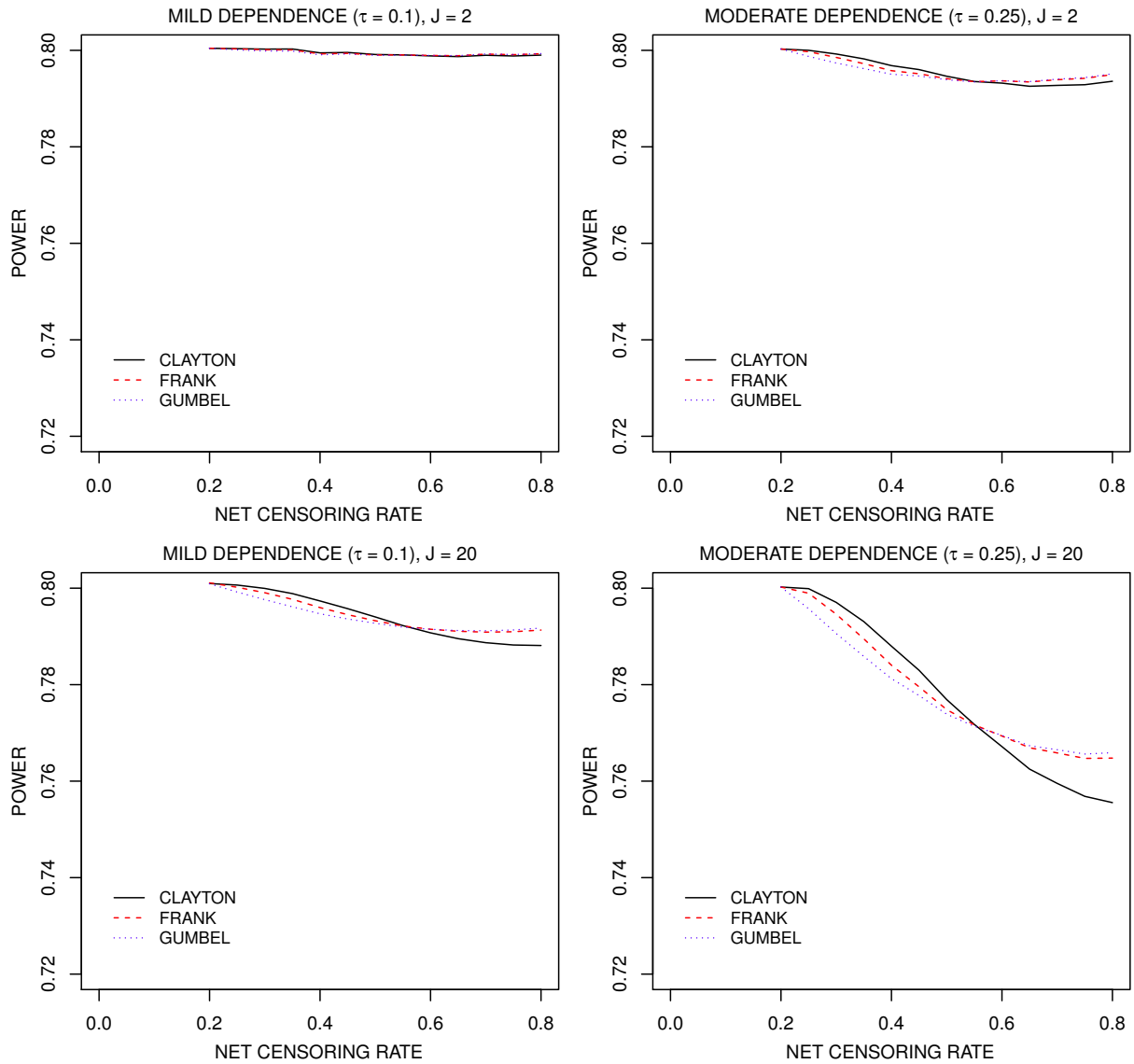
Figure 3: Power implications of within-cluster association in the random censoring times under joint censoring models induced by different copula functions where the within-cluster association in the failure and censoring times are constrained to be the same ($\tau = \tau_c$); the original sample size is computed on the basis of a Clayton copula model for the failure times and the assumption of independent censoring times; $\kappa = 0.75$, $\beta_A = \log 0.8$, $p_0 = 0.2$.

4 SAMPLE SIZE CALCULATIONS FOR TRIALS WITH CLUSTERED INTERVAL-CENSORED EVENT TIMES

4.1 ESTIMATING EQUATIONS AND SAMPLE SIZE CRITERIA

Interval-censored event times arise when it is only possible to determine whether events have occurred at periodic assessments [36]. In rheumatology studies, for example, interest lies in the time to the development of joint damage, but the extent of joint damage is only possible to determine when patients undergo radiographic examination [37]. In this case the time of joint damage will only be known to fall between the time of the first radiograph showing evidence of damage and the time of the preceding radiographic examination. Other examples include trials aiming to evaluate osteoporosis treatments for the prevention of asymptomatic fractures, studies of the development of new metastatic lesions, and studies in nephrology on the development of kidney stones.

We assume again that $T_{ij}|Z_i$ follows a proportional hazards model (2.1) with a $q \times 1$ parameter α indexing the baseline hazard and β the regression parameter of interest. The marginal survivor function $\mathcal{F}(t|Z_i; \theta) = P(T_{ij} \geq t|Z_i; \theta)$ is then indexed by a $(q+1) \times 1$ parameter $\theta = (\alpha', \beta)'$. In the present setting, we consider a cluster-randomized trial in which the plan is to observe each individual at R prespecified assessment times $a_1, \ldots, a_R$; we let $a_0 = 0$ and $a_{R+1} = \infty$. Under this observation scheme, we observe $Y_{ijr} = \mathrm{I}(a_{r-1} < T_{ij} \leq a_r)$, $r = 1, \ldots, R+1$. The response data provided by individual j in cluster i is $Y_{ij} = (Y_{ij1}, \ldots, Y_{ijR})'$, where $Y_{ij,R+1} = 1 - \sum_{r=1}^R Y_{ijr}$, and $Y_i = (Y_{i1}', \ldots, Y_{iJ}')'$ contains all response data from cluster i, $i = 1, \ldots, n$. Let $\mu_{ij} = (\mu_{ij1}, \ldots, \mu_{ijR})'$, where

$$\mu_{ijr} = E[Y_{ijr}|Z_i; \theta] = P(a_{r-1} < T_{ij} \leq a_r|Z_i; \theta) = \mathcal{F}(a_{r-1}|Z_i; \theta) - \mathcal{F}(a_r|Z_i; \theta), r = 1, \ldots, R.$$

Let $\mu_i = (\mu_{i1}', \ldots, \mu_{iJ}')'$ denote a $JR \times 1$ vector and $D_i = [\partial \mu_i/\partial \alpha', \partial \mu_i/\partial \beta]$ a $JR \times (q+1)$ matrix of associated derivatives. We consider the generalized estimating equations for θ, similar to those of Kor et al. [38],

$$U(\theta) = \sum_{i=1}^n U_i(\theta) = \sum_{i=1}^n \begin{bmatrix} U_i(\alpha) \\ U_i(\beta) \end{bmatrix} = \sum_{i=1}^n D_i' V_i^{-1}(Y_i - \mu_i) . \tag{4.1}$$

where V_i is a $JR \times JR$ working covariance matrix. A working independence assumption is particularly convenient in this context, and provided robust variance estimates are used at the time of analysis, inferences can be valid when there is within-cluster dependence in the event times. In this case V_i is block diagonal with $R \times R$ block diagonal matrices $V_{ij} = \mathrm{Cov}(Y_{ij}, Y_{ij}'|Z_i)$, $j = 1, \ldots, J$, which account for the correlation of responses at different assessment times within individuals; that is,

$$V_i = \begin{pmatrix} V_{i1} & & \\ & \ddots & \\ & & V_{iJ} \end{pmatrix} = \begin{pmatrix} \mathrm{Cov}(Y_{i1}, Y_{i1}'|Z_i) & & 0 \\ & \ddots & \\ 0 & & \mathrm{Cov}(Y_{iJ}, Y_{iJ}'|Z_i) \end{pmatrix}, \tag{4.2}$$

and the (r,s)th entry of V_{ij} is

$$\mathrm{Cov}(Y_{ijr}, Y_{ijs}|Z_i) = \begin{cases} \mu_{ijr}(1 - \mu_{ijr}), & r = s ; \\ -\mu_{ijr}\mu_{ijs}, & r \neq s . \end{cases} \tag{4.3}$$

Note that if the marginal regression models are correctly specified, $n^{-1/2}U(\theta)$ is asymptotically multivariate normal with mean zero and covariance given analogously to (2.3) by

$$\mathcal{B} = E[U_i(\theta)U_i(\theta)'] , \tag{4.4}$$

estimated as

$$\widehat{B} = \frac{1}{n} \sum_{i=1}^{n} U_i(\theta) U_i'(\theta) \Big|_{\theta = \widehat{\theta}}.$$

The estimator $\widehat{\theta}$ is the root of $U(\theta) = 0$ and is consistent for θ with $n^{1/2}(\widehat{\theta} - \theta) \xrightarrow{D} N(0, \Gamma)$ asymptotically, where $\Gamma = \mathcal{A}^{-1} \mathcal{B}[\mathcal{A}^{-1}]'$, and $\mathcal{A} = -E[\partial U_i(\theta)/\partial \theta']$. Again the matrix $\mathcal{A}$ can be consistently estimated by

$$\widehat{A} = -\frac{1}{n} \sum_{i=1}^{n} \partial U_i(\theta)/\partial \theta' \Big|_{\theta = \widehat{\theta}}. \tag{4.5}$$

Model assumptions are required to derive the sample size formula on the basis of the preceding asymptotic variance formula for clustered interval-censored data. Copula functions can be used to construct the joint distribution with any specified marginal properties. Consider a cluster-randomized trial in which the treatment is randomly allocated to clusters. Suppose we aim to test whether the treatment has an effect on the time to a certain event. The null hypothesis is $H_0: \beta = \beta_0 = 0$, and the alternative hypothesis is $H_A: \beta \neq \beta_0$, and let β_A denote the clinically important effect.

As in Section 2, the limiting distribution of a Wald statistic can be used to select the required sample size (number of clusters) for a two-sided test with significance level γ_1 and power $1 - \gamma_2$. The key point is to derive the formulae for $\mathcal{A} = E[-\partial U_i(\theta)/\partial \theta']$ and $\mathcal{B} = E[U_i(\theta)U_i'(\theta)]$, and hence the form of $\Gamma = \mathcal{A}^{-1}\mathcal{B}[\mathcal{A}^{-1}]'$, so the required sample size can be obtained on the basis of $\Psi = \Gamma_{q+1, q+1}$, the element from the covariance matrix; the formulae are outlined in Appendix B. The resulting sample size n necessary to detect the effect of treatment with the specified power is

$$n \geq \left\{ \frac{z_{\gamma_1/2}\sqrt{\Psi_0} + z_{\gamma_2}\sqrt{\Psi_A}}{\beta_A} \right\}^2, \tag{4.6}$$

where Ψ_0 and Ψ_A are the elements of Γ computed under the null and alternative settings. At the design stage of clinical trials, to estimate the required number of clusters, specifications of the effect of interest β_A, cluster size J, inspection times $a_1, \ldots, a_R$, parametric baseline hazard function, and especially the joint distribution for clustered event times are required.

4.2 EMPIRICAL VALIDATION OF SAMPLE SIZE FORMULA FOR CLUSTERED INTERVAL-CENSORED EVENT TIMES

Here we examine the performance of the proposed sample size formula for clustered interval-censored data. Consider an equal allocation cluster-randomized trial with binary treatment covariate Z_i, $P(Z_i = 1) = P(Z_i = 0) = 0.5$. Assume that T_{ij} follows the proportional hazards model given by (2.1) with Weibull baseline cumulative hazard $\Lambda_0(s) = (\lambda_0 s)^\kappa$, where $\alpha = (\log \lambda_0, \log \kappa)'$, $q = 2$, and $\theta = (\alpha', \beta)'$, $j = 1, \ldots, J$, $i = 1, \ldots, n$; we consider cluster sizes of $J = 2, 5, 20$ and 100. Suppose $\kappa = 0.75$ and choose λ_0 so that $P(T_{ij} > 1 | Z_i = 0) = p_a$ to give a specified administrative censoring rate; we set $p_a = 0.2$. Suppose the plan is to assess each individual R times over the interval $[0, 1]$ at prespecified assessment times $a_1, \ldots, a_R$ evenly spaced over the observation interval, that is, $a_r = r/R$, $r = 1, \ldots, R$, with $R = 2, 4$ or 12. Let $Y_{ij} = (Y_{ij1}, \ldots, Y_{ijR})'$ denote the event information provided by individual j in cluster i, where $Y_{ijr} = I(a_{r-1} < T_{ij} \leq a_r)$.

Suppose the within-cluster association in the underlying failure times is induced by the Clayton copula with Kendall's τ of 0.05, 0.10, and 0.25 for small, mild and moderate within-cluster association respectively. For each parameter combination, we estimate the sample size (number of clusters) by (4.6) given $\beta_A = \log 0.8$, the type I error rate $\gamma_1 = 0.05$ and power $1 - \gamma_2 = 0.8$. After obtaining the required minimum sample size, we generate the corresponding covariate Z_i and clustered

response Y_i. Parameter estimate are then obtained via the estimating equation (4.1). For each parameter combination, nsim $= 2000$ datasets are simulated and analysed to yield 2000 estimates of β and respective robust variance estimates. The ESE, ASE, REJ% and 95% ECP% are summarized in Table 2.

The REJ% is close to the nominal type I error rate when $\beta = 0$ and close to the nominal power when $\beta = \log 0.8$, with the latter supporting the validity of the sample size formula. The empirical biases (not shown) are all negligible, so it is not surprising that the empirical coverage probabilities are all compatible with the nominal 95% level. As the number of assessments increases, the required sample size is found to decrease, but the extent of this decrease from the case of $R = 4$ to $R = 12$ is quite small, particularly when cluster sizes are large. To clearly understand the impact of the number of assessments on the efficiency, we computed the asymptotic relative efficiency of estimators for the marginal parameters, defined as

$$\mathrm{ARE}_{r,k} = \frac{\mathrm{asvar}(\tilde{\theta}_k)}{\mathrm{asvar}_r(\hat{\theta}_k)} \ ,$$

where $\mathrm{asvar}(\tilde{\theta}_k)$ is the asymptotic variance of θ_k for $R = 100$; this value is large enough to mimic the case that the event times are known precisely, that is, the case of clustered right-censored event times. The term $\mathrm{asvar}_r(\hat{\theta}_k)$ represents the asymptotic variance of $\hat{\theta}_k$ for the case $R = r$, corresponding to clustered interval-censored failure time data, where $k = 1, 2, 3$.

Figure 4 shows the trend of asymptotic relative efficiency for estimators of the marginal parameters with cluster sizes of $J = 2, 5$, and 20, respectively. From these figures, we note that when the number of assessments increases to $R = 8$, the asymptotic relative efficiencies for both λ_0 and β are close to 1 in all cases considered. This also supports the empirical findings that the number of clusters required decreases very little when the number of assessments increases from $R = 4$ to 12. Figure 4 also shows that the impact of the number of assessments is more severe for small cluster sizes, which agrees with what we found from Table 2. As one might expect, however, the number of assessments seriously affects the efficiency of the estimator for the trend parameter κ, so when the entire marginal distribution is of interest, increasing the number of assessments certainly can improve efficiency for some features of the distribution. There is of course a trade-off between the statistical goals of precision and power and the economic and other costs. The development of optimal design criteria which enables one to weigh the merits of increasing the number of clusters or the number of follow-up assessments to be scheduled, subject to prespecified budgetary constraints, represents an important area of future research.

5 ILLUSTRATIVE EXAMPLE INVOLVING TREATMENT FOR OTITIS MEDIA

Otitis media is inflammation of the inner ear which puts patients at risk of permanent damage and loss of hearing. We illustrate the steps in trial design by considering the study discussed in Manatunga and Chen [14] in which children from 6 months to 8 years of age with otitis media requiring surgical insertion of tubes in the auditory canal are randomized to receive either 2 weeks of medical therapy with prednisone and sulfamethoprim or no medical therapy (standard care). The trial is conceived on the basis of the data in Le and Lindgren [13] in which all children except one had bilateral inflammation, and so we consider clusters of size two with $J = 2$. In the absence of information on the trend, we set $\kappa = 1$. The median time to failure of the inserted tube was estimated to be 7 months, assuming 30 days per month yields $\lambda_0 = -\log 0.5/210 \approx 0.0033$. As in Manatunga and Chen [14] we set $\tau = 0.56$ to reflect moderate to strong within-child association in the failure times. Since follow-up is planned for 1.5 years, we set $C^\dagger = 540$ and anticipate an administrative censoring rate of 17% for the control arm. To accommodate study withdrawal, we adopt an exponential model for

Table 2: Sample size estimation and empirical properties of estimator $\hat\beta$ under cluster-randomized design for interval-censored data when the Clayton copula is used to induce the within-cluster association between event times; $\kappa = 0.75$, $\beta_A = \log 0.8$, $p_a = 0.2$, nsim= 2000.

| J	τ	R	n	$\beta = 0$			$\beta = \log 0.8$			
				ESE	ASE	REJ%	ESE	ASE	REJ%	ECP%
2	0.05	2	458	0.081	0.079	5.6	0.083	0.081	79.0	94.3
		4	440	0.078	0.079	4.9	0.080	0.081	81.3	95.4
		12	433	0.079	0.079	5.1	0.081	0.081	79.6	95.4
	0.10	2	490	0.079	0.079	4.6	0.080	0.081	78.5	95.6
		4	471	0.079	0.079	5.2	0.081	0.081	78.0	94.9
		12	465	0.080	0.079	5.2	0.082	0.081	79.8	94.6
	0.25	2	584	0.077	0.079	4.2	0.079	0.080	78.1	95.5
		4	564	0.080	0.079	5.7	0.082	0.081	80.9	94.3
		12	557	0.081	0.079	5.0	0.082	0.081	80.0	94.4
5	0.05	2	222	0.076	0.079	4.8	0.078	0.080	82.2	95.7
		4	214	0.080	0.079	5.6	0.082	0.081	79.5	94.3
		12	212	0.079	0.079	5.7	0.081	0.080	79.6	94.4
	0.10	2	272	0.080	0.079	5.4	0.082	0.080	80.0	93.9
		4	265	0.080	0.079	5.4	0.081	0.080	79.1	94.4
		12	262	0.080	0.079	5.5	0.082	0.080	78.6	94.9
	0.25	2	422	0.079	0.079	4.8	0.079	0.080	81.1	95.1
		4	413	0.079	0.079	5.3	0.079	0.080	79.7	95.3
		12	409	0.081	0.079	5.1	0.082	0.080	79.5	95.0
20	0.05	2	103	0.078	0.079	4.8	0.078	0.079	80.6	95.7
		4	101	0.079	0.079	5.5	0.080	0.080	78.7	94.9
		12	101	0.079	0.078	5.6	0.080	0.079	78.9	95.1
	0.10	2	164	0.080	0.079	4.7	0.080	0.079	78.9	94.8
		4	161	0.080	0.079	4.9	0.080	0.080	80.1	94.8
		12	160	0.080	0.079	5.3	0.080	0.080	78.5	95.1
	0.25	2	342	0.080	0.079	5.3	0.080	0.080	79.0	95.1
		4	337	0.081	0.079	5.7	0.081	0.080	79.4	94.8
		12	336	0.078	0.079	4.9	0.079	0.080	79.4	95.3
100	0.05	2	72	0.080	0.078	5.1	0.082	0.079	79.7	93.5
		4	71	0.078	0.079	4.6	0.078	0.079	79.7	95.5
		12	71	0.076	0.078	4.6	0.077	0.078	81.3	95.3
	0.10	2	135	0.080	0.079	5.4	0.080	0.079	80.1	94.9
		4	134	0.080	0.079	5.5	0.080	0.079	81.5	94.9
		12	133	0.081	0.079	5.7	0.081	0.079	79.4	94.8
	0.25	2	320	0.079	0.080	4.9	0.079	0.080	81.0	95.4
		4	317	0.079	0.079	4.3	0.079	0.080	82.0	95.4
		12	316	0.081	0.079	5.4	0.082	0.080	79.2	94.3
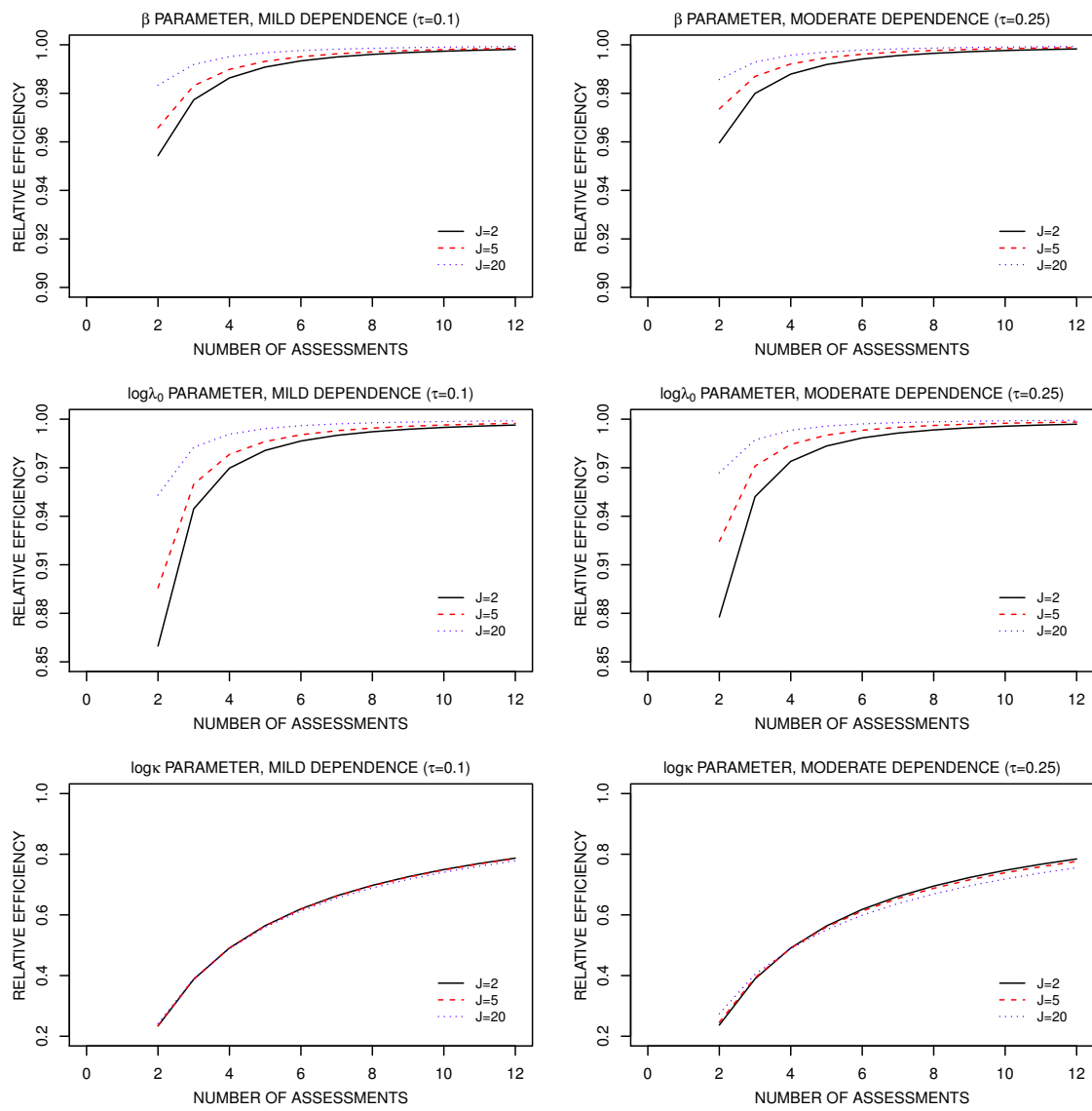
Figure 4: Asymptotic relative efficiency of estimators for marginal parameters for clustered interval-censored event times as a function of the number of assessments, degree of dependence and copula function.

loss to follow-up to give a net rate of censoring in the control arm of 40% or 60%. Note that this setting is slightly different than the setting discussed in Section 3 where different individuals within each cluster had different censoring times; here the clusters are defined by children and the times to failure of the left and right tubes would be censored at a common time. The formula in Appendix A can be easily modified to address this by defining $\tilde{\mathcal{G}}(\cdot)$ as the survival distribution for the cluster-level censoring time and replacing $\mathcal{G}(s, t)$ by $\tilde{\mathcal{G}}(\max(s, t))$ in (A.9). Under Clayton, Frank and Gumbel copulas, we compute the number of children required to randomize to ensure 80% power to detect a 30, 40 or 50% reduction in the marginal hazard for failure based on a two-sided test at the 5% level. The results displayed in Table 3 provide a simple illustration of how the most conservative copula depends on the rate of censoring. When the net censoring is expected to be 40%, the Clayton copula yields the largest sample size, but when it is 60%, the Frank copula yields the largest sample sizes.

Table 3: Number of clusters (children) required for otitis media study under Clayton, Frank and Gumbel copulas for different clinically important treatment effects and net censoring rates.

Censoring %	$\exp(\beta) = 0.7$			$\exp(\beta) = 0.6$			$\exp(\beta) = 0.5$		
	Clayton	Frank	Gumbel	Clayton	Frank	Gumbel	Clayton	Frank	Gumbel
40%	366	357	347	181	177	172	101	99	96
60%	521	530	519	258	263	259	144	147	145

6   DISCUSSION

We derived sample size formulae for cluster-randomized trials involving right- and interval-censored event times in which the analysis is based on a marginal proportional hazards assumption. For right-censored data, we derived expressions for the asymptotic robust variance of the Wald statistic on the basis of the approach of Lee et al. [12], and for clustered interval-censored data, we likewise adopted the structure of Kor et al. [38]. Both of these frameworks invoke a working independence assumption, so robust variance estimation is required to ensure valid inference in the presence of within-cluster association. The simulation studies conducted confirm that the formulae are valid. Code for computing the required sample size is available in R from the authors upon request. Robustness of these formulae to the misspecification of copula functions and to within-cluster dependence in the censoring times is also investigated using large sample theory for clustered failure times in the context of right-censored data.

As in other sample size formula for cluster-randomized trials, the derived formulae require specification of the degree of within-cluster dependence, measured in the failure time setting by Kendall's τ. A good approximation to the degree of within-cluster dependence is important [39], so it is therefore customary to rely on estimates reported in the literature. We recognize that it is unlikely that values of Kendall's τ would be reported in the clinical literature, and so we recommend the conduct of small pilot studies. More recently, there has been increased interest in planning trials with adaptive sample size re-estimation. This is carried out in its simplest form by having an internal pilot study, after which blinded data are used to estimate unknown parameters; these new estimates are then used to revise sample size calculations. This is a generally important area of research as these methods increase efficiency. We have developed such methods in another context [40] and plan to study this in the present setting in future work.

We have focussed on settings with a single binary treatment indicator, but the proposed methods extend naturally to deal with trials where analyses control for cluster-level covariates. A two-

dimensional covariate vector would arise if one designed a three-armed trial, in which case one might specify $Z_i = (Z_{i1}, Z_{i2})'$ where Z_{i1} and Z_{i2} indicate assignment to the first and second experimental treatments, respectively, and $Z_{i1} = Z_{i2} = 0$ if cluster i is assigned to the control intervention. More generally, other multidimensional descriptive cluster-level covariates can be incorporated into the analyses, but at the design stage their joint distribution would have to be specified to facilitate computation of the matrix expectations in the robust variance formula; see Appendices A and B. Individual-level covariates can also be controlled for in the analysis in principle, but assumptions would again be required regarding their joint distribution and in particular the extent to which these covariates are dependent within clusters.

We also restricted attention to the setting in which cluster sizes are fixed at a common value, denoted by J. While this is often reasonable, cluster sizes routinely vary within studies. When planning studies in this case, it is perhaps most common to use formulae derived for fixed cluster sizes but to use the anticipated average cluster size $\bar{J} = \sum_{i=1}^{n} J_i/n$ in the formula in place of J [29,32,41]. When cluster sizes vary and the response is continuous, use of the average size in the formula derived for common cluster sizes can lead to inadequate power; the loss in power can be small, however, if the cluster size tends to be large and the intraclass correlation coefficient is small [42,43]. Manatunga et al. [42] developed a refinement to the usual sample size formula for continuous outcomes to deal with variable cluster sizes, which involves adding a correction term (a function of the coefficient of variation of the cluster size) to the formula on the basis of a common cluster size. Van Breukelen et al. [43] investigated the consequences of unequal versus equal cluster sizes in terms of the precision of treatment effects estimators in cluster-randomized trials with continuous outcomes. They provide a formula for the approximate relative efficiency of the estimators, which can be used to adjust an initial estimate of the number of clusters required on the basis of a common cluster size. Candel and Van Breukelen [44] extended this approach to varying cluster size with binary outcomes when the analysis is based on the mixed logistic regression. We are not aware of any methods for dealing with variable cluster sizes for event time responses, and so extensions to deal with this represent an important area of future research.

In principle, the methods we develop can be adapted to the setting in which the number of clusters is fixed, and the goal is to determine the number of individuals within each cluster necessary to achieve the desired power. This may be a more appropriate framework in a health promotion study, for example, in which clinics are randomized to deliver one of two smoking cessation programs. If there are a fixed number of clinics available to take part but patients are continually being referred to these clinics, it is natural to want to know how many patients should be recruited from these clinics to ensure adequate power to detect a specified effect of an experimental cessation program. As pointed out by Hemming et al. [45], the limiting robust standard deviation of estimators obtained under the working independence assumption decreases as the cluster size increases, but there is a positive limiting value. As a result, for a given number of clusters, minimal clinically important effect, and type I error rate, there is a limit to the power that can be achieved by increasing the cluster size. Conversely, for a given number of clusters, power, and type I error rate, there is a limit to how small the effect can even as cluster sizes increase. In situations where small clinically important effects are specified, it may therefore be necessary to select the number of clusters and the cluster size in concert to ensure practical and statistical constraints are met.

When the clustered event times are interval-censored data, our sample size formula is derived based on the assumption that all the assessments on each individual are available. Individuals may of course prematurely drop-out of studies, leading to missed assessments. In this case, the response vectors are incompletely observed, but modifications to the estimating functions are straightforward if assumptions about the withdrawal process are made.

APPENDIX A   LIMITING DISTRIBUTION OF THE WALD STATISTIC BASED ON CLUSTERED RIGHT-CENSORED EVENT TIMES

In what follows, we consider the setting in which Z_i is a fixed binary treatment indicator and assume that the marginal distribution of T_{ij}, the event time for individual j in cluster i, satisfies the proportional hazard assumption with

$$\lambda_{ij}(t|Z_i) = \lambda_0(t; \alpha) \exp(Z_i \beta) ,$$

where $\lambda_0(t; \alpha)$ is the baseline hazard function indexed by a vector α, a β is the coefficient of interest, and $j = 1, \ldots, J$, $i = 1, \ldots, n$. If $C^\dagger$ is an administrative censoring time, the plan is to observe over $(0, C^\dagger]$, but C^*_{ij} is a random censoring time with survivor function $\mathcal{G}^*(c)$, representing a possible early withdrawal time. The net censoring time for individual j in cluster i is then $C_{ij} = \min(C^*_{ij}, C^\dagger)$, with survivor function $\mathcal{G}(c)$.

In counting process notation, we let $\{N_{ij}(t), 0 < t\}$ denote the right-continuous counting process for T_{ij}, where $N_{ij}(t) = \mathrm{I}(T_{ij} \leq t)$ indicates that the event occurred at or before time t for individual j in cluster i. Then $dN_{ij}(t) = 1$ if individual j in cluster i experiences the event at time t, and $dN_{ij}(t) = 0$ otherwise. Let $\bar{Y}_{ij}(t) = Y_{ij}(t)Y^\dagger_{ij}(t)$ be the indicator that the jth individual in cluster i is under observation and at risk of event at time t, where $Y^\dagger_{ij}(t) = \mathrm{I}(T_{ij} \geq t)$ and $Y_{ij}(t) = \mathrm{I}(C_{ij} \geq t)$.

Under working independence assumption, the partial score function for β is

$$U(\beta) = \sum_{i=1}^n \sum_{j=1}^J \int_0^\infty \bar{Y}_{ij}(t) \left(Z_i - \frac{\sum_{j=1}^J S_{1j}(t; \beta)}{\sum_{j=1}^J S_{0j}(t; \beta)} \right) dN_{ij}(t) ,$$

where $S_{rj}(t; \beta) = n^{-1} \sum_{i=1}^n \bar{Y}_{ij}(t) Z_i^r \exp(Z_i \beta)$, $r = 0, 1$ and $Z_i^0 = 1$ and $Z_i^1 = Z_i$. Lee et al. [12] show that the score function is asymptotically equivalent to a sum of independent identically distributed terms

$$n^{-1/2} U(\beta) = n^{-1/2} \sum_{i=1}^n \sum_{j=1}^J \zeta_{ij}$$

where

$$\zeta_{ij} = \int_0^\infty \bar{Y}_{ij}(t)(Z_i - W(t))dM_{ij}(t) ,$$

where we suppress the functional dependence on β in the terms $W(t) = \sum_{j=1}^J s_{1j}(t; \beta) / \sum_{j=1}^J s_{0j}(t; \beta)$ with $s_{rj}(t; \beta)$ the limit of $S_{rj}(t; \beta)$, and

$$M_{ij}(t) = N_{ij}(t) - \int_0^t \bar{Y}_{ij}(u) \exp(Z_i \beta) \lambda_0(u) du$$

where $\{M_{ij}(t), 0 < t\}$ is a martingale. By the Central Limit Theorem, $n^{-1/2} U(\beta)$ converges to a normal random variable with mean 0 and variance $\mathcal{B}$, where

$$\mathcal{B} = n^{-1} \sum_{i=1}^n \mathrm{Var}(\zeta_{i\cdot}) = \sum_{j,k=1}^J \mathrm{Cov}\left(\zeta_{ij}, \zeta_{ik}\right) = \sum_{j,k=1}^J E(\zeta_{ij}\zeta_{ik}) , \tag{A.1}$$

where $\zeta_{i\cdot} = \sum_{j=1}^J \zeta_{ij}$, $i = 1, \ldots, n$.

The root of $U(\beta) = 0$ is a consistent estimator $\widehat{\beta}$ with $n^{1/2}(\widehat{\beta} - \beta) \xrightarrow{D} N(0, \Gamma)$, where $\Gamma = \mathcal{B}/\mathcal{A}^2$ and $\mathcal{A} = -E[\partial U_i(\beta)/\partial \beta]$. The sample size formula is derived on the basis of this limiting distribution with the $\mathcal{B}$ and $\mathcal{A}$ computed on the basis of parametric models. We give the results of these derivations in the following two sections under the assumption of dependent within-cluster censoring times and independent censoring within clusters.

A.1 GENEARL DERIVATION OF $\mathcal{B}$

We first consider a general case in which the censoring times could be correlated within clusters. Assume $(C_{i1}, \ldots, C_{iJ})' \perp Z_i$ and let $\mathcal{G}(u) = P(C_{ij} \geq u)$ be the survivor function for the censoring time C_{ij}, and $\mathcal{G}(s,t) = P(C_{ij} \geq s, C_{ik} \geq t)$ denote the joint survivor function for the censoring times (C_{ij}, C_{ik}) within cluster i; both are assumed common across the two groups. The joint survivor function $\mathcal{G}(s,t)$ describes the association between within-cluster censoring times.

To derive an expression for (A.1) we first consider the case where $j = k$ and note

$$
\begin{aligned}
E\left[\zeta_{ij}^2\right] &= E\left[\int_0^C \bar{Y}_{ij}(s)(Z_i - W(s))^2 \lambda_{ij}(s) ds\right] \\
&= E_{Z_i}\left[E_{Y_{ij}^\dagger(s)|Z_i}\left\{E_{Y_{ij}(s)|Y_{ij}^\dagger(s),Z_i}\left[\int_0^C \bar{Y}_{ij}(s)(Z_i - W(s))^2 \lambda_{ij}(s) ds\right]\right\}\right] \\
&= E_{Z_i}\left[E_{Y_{ij}^\dagger(s)|Z_i}\left\{\int_0^C \mathcal{G}(s) Y_{ij}^\dagger(s)(Z_i - W(s))^2 \lambda_{ij}(s) ds\right\}\right] \\
&= E_{Z_i}\left[\int_0^C \mathcal{G}(s) P(T_{ij} \geq s|Z_i)(Z_i - W(s))^2 \lambda_{ij}(s) ds\right] \\
&= E_{Z_i}\left[\int_0^C \mathcal{G}(s)(Z_i - W(s))^2 f_j(s|Z_i) ds\right]
\end{aligned}
\tag{A.2}
$$

where $f_j(s|Z_i)$ is the conditional density of the event time for individual j in cluster i. And $E_{Z_i}[\cdot]$ depends on the trial allocation probability.

For $j \neq k$, since

$$
E[\zeta_{ij}\zeta_{ik}] = E\left[\iint_{(0,C]^2} \bar{Y}_{ij}(s)\bar{Y}_{ik}(t)(Z_i - W(s))(Z_i - W(t)) dM_{ij}(s) dM_{ik}(t)\right],
$$

and

$$
\begin{aligned}
dM_{ij}(s)dM_{ik}(t) =\ & dN_{ij}(s)dN_{ik}(t) - dN_{ij}(s)\bar{Y}_{ik}(t)d\Lambda_{ik}(t) - \bar{Y}_{ij}(s)d\Lambda_{ij}(s)dN_{ik}(t) \\
& - \bar{Y}_{ij}(s)\bar{Y}_{ik}(t)d\Lambda_{ij}(s)d\Lambda_{ik}(t),
\end{aligned}
$$

then

$$
\begin{aligned}
E[\zeta_{ij}\zeta_{ik}] =\ & E\left[\iint_{(0,C]^2} \bar{Y}_{ij}(s)\bar{Y}_{ik}(t)(Z_i - W(s))(Z_i - W(t)) dN_{ij}(s) dN_{ik}(t)\right] \\
& - E\left[\iint_{(0,C]^2} \bar{Y}_{ij}(s)\bar{Y}_{ik}(t)(Z_i - W(s))(Z_i - W(t)) dN_{ij}(s) d\Lambda_{ik}(t)\right] \\
& - E\left[\iint_{(0,C]^2} \bar{Y}_{ij}(s)\bar{Y}_{ik}(t)(Z_i - W(s))(Z_i - W(t)) d\Lambda_{ij}(s) dN_{ik}(t)\right] \\
& + E\left[\iint_{(0,C]^2} \bar{Y}_{ij}(s)\bar{Y}_{ik}(t)(Z_i - W(s))(Z_i - W(t)) d\Lambda_{ij}(s) d\Lambda_{ik}(t)\right].
\end{aligned}
\tag{A.3}
$$

The first term in (A.3) is then computed as

$$E\left[\iint_{(0,C]^2} \bar{Y}_{ij}(s)\bar{Y}_{ik}(t)(Z_i - W(s))(Z_i - W(t))dN_{ij}(s)dN_{ik}(t)\right]$$

$$= E_{Z_i}\left[E_{Y_{ij}^\dagger(s),Y_{ik}^\dagger(t)|Z_i}\left\{E_{dN_{ij}(s),dN_{ik}(t)|Y_{ij}^\dagger(s),Y_{ik}^\dagger(t),Z_i}\right.\right.$$

$$\left.\left[E_{Y_{ij}(s),Y_{ik}(t)|Z_i,Y_{ij}^\dagger(s),Y_{ik}^\dagger(t),dN_{ij}(s),dN_{ik}(t)}\left\{\iint_{(0,C]^2} \bar{Y}_{ij}(s)\bar{Y}_{ik}(t)(Z_i - W(s))(Z_i - W(t))dN_{ij}(s)dN_{ik}(t)\right\}\right]\right\}\right]$$

$$= E_{Z_i}\left[E_{Y_{ij}^\dagger(s),Y_{ik}^\dagger(t)|Z_i}\left\{E_{dN_{ij}(s),dN_{ik}(t)|Y_{ij}^\dagger(s),Y_{ik}^\dagger(t),Z_i}\left[\right.\right.\right.$$

$$\left.\left.\left.\iint_{(0,C]^2} \mathcal{G}(s,t)Y_{ij}^\dagger(s)Y_{ik}^\dagger(t)(Z_i - W(s))(Z_i - W(t))dN_{ij}(s)dN_{ik}(t)\right]\right\}\right]$$

$$= E_{Z_i}\left[E_{Y_{ij}^\dagger(s),Y_{ik}^\dagger(t)|Z_i}\left\{\iint_{(0,C]^2} \mathcal{G}(s,t)Y_{ij}^\dagger(s)Y_{ik}^\dagger(t)(Z_i - W(s))(Z_i - W(t))P(T_{ij} = s, T_{ik} = t|Y_{ij}^\dagger(s),Y_{ik}^\dagger(t),Z_i)dsdt\right\}\right]$$

$$= E_{Z_i}\left[\iint_{(0,C]^2} \mathcal{G}(s,t)(Z_i - W(s))(Z_i - W(t))f_{jk}(s,t|Z_i)dsdt\right], \tag{A.4}$$

where $f_{jk}(s,t|Z_i)$ is the pairwise conditional density for (T_{ij}, T_{ik}) obtained through the specification of a copula function. Using the same strategy for remaining terms of (A.3), we obtain

$$E\left[\iint_{(0,C]^2} \bar{Y}_{ij}(s)\bar{Y}_{ik}(t)(Z_i - W(s))(Z_i - W(t))dN_{ij}(s)d\Lambda_{ik}(t)\right]$$

$$= E_{Z_i}\left[\iint_{(0,C]^2} \mathcal{G}(s,t)(Z_i - W(s))(Z_i - W(t))\left(-\frac{\partial \mathcal{F}_{jk}(s,t|Z_i)}{\partial s}\right)\lambda_0(t)e^{Z_i\beta}dsdt\right], \tag{A.5}$$

$$E\left[\iint_{(0,C]^2} \bar{Y}_{ij}(s)\bar{Y}_{ik}(t)(Z_i - W(s))(Z_i - W(t))d\Lambda_{ij}(s)dN_{ik}(t)\right]$$

$$= E_{Z_i}\left[\iint_{(0,C]^2} \mathcal{G}(s,t)(Z_i - W(s))(Z_i - W(t))\left(-\frac{\partial \mathcal{F}_{jk}(s,t|Z_i)}{\partial t}\right)\lambda_0(s)e^{Z_i\beta}dsdt\right], \tag{A.6}$$

and

$$E\left[\iint_{(0,C]^2} \bar{Y}_{ij}(s)\bar{Y}_{ik}(t)(Z_i - W(s))(Z_i - W(t))d\Lambda_{ij}(s)d\Lambda_{ik}(t)\right]$$

$$= E_{Z_i}\left[\iint_{(0,C]^2} \mathcal{G}(s,t)(Z_i - W(s))(Z_i - W(t))\mathcal{F}_{jk}(s,t|Z_i)\lambda_0(s)e^{Z_i\beta}\lambda_0(t)e^{Z_i\beta}dsdt\right]. \tag{A.7}$$

where $\mathcal{F}_{jk}(s,t|Z_i)$ is the pairwise conditional survivor function for (T_{ij}, T_{ik}) obtained through the specification of a copula function. Plugging (A.4 - A.7) into (A.3), we obtain

$$E[\zeta_{ij}\zeta_{ik}] = E_{Z_i}\left\{\iint_{(0,C]^2} \mathcal{G}(s,t)(Z_i - W(s))(Z_i - W(t))f_{jk}(s,t|Z_i)dsdt\right.$$

$$-\iint_{(0,C]^2} \mathcal{G}(s,t)(Z_i - W(s))(Z_i - W(t))\left(-\frac{\partial \mathcal{F}_{jk}(s,t|Z_i)}{\partial s}\right)\lambda_0(t)e^{Z_i\beta}dsdt$$

$$-\iint_{(0,C]^2} \mathcal{G}(s,t)(Z_i - W(s))(Z_i - W(t))\left(-\frac{\partial \mathcal{F}_{jk}(s,t|Z_i)}{\partial t}\right)\lambda_0(s)e^{Z_i\beta}dsdt$$

$$\left.+\iint_{(0,C]^2} \mathcal{G}(s,t)(Z_i - W(s))(Z_i - W(t))\mathcal{F}_{jk}(s,t|Z_i)\lambda_0(s)e^{Z_i\beta}\lambda_0(t)e^{Z_i\beta}dsdt\right\}. \tag{A.8}$$

The asymptotic variance of $n^{-1/2}U(\beta)$ can then be calculated as

$$
\begin{aligned}
\mathcal{B} = \sum_{j=1}^{J} E_{Z_i} &\left[\int_0^C \mathcal{G}(s)(Z_i - W(s))^2 f_j(s|Z_i)ds \right] \\
+ \sum_{j \neq k} &\left[E_{Z_i} \left\{ \iint_{(0,C]^2} \mathcal{G}(s,t)(Z_i - W(s))(Z_i - W(t))f_{jk}(s,t|Z_i)dsdt \right.\right. \\
&- \iint_{(0,C]^2} \mathcal{G}(s,t)(Z_i - W(s))(Z_i - W(t)) \left(-\frac{\partial \mathcal{F}_{jk}(s,t|Z_i)}{\partial s} \right) \lambda_0(t)e^{Z_i\beta}dsdt \\
&- \iint_{(0,C]^2} \mathcal{G}(s,t)(Z_i - W(s))(Z_i - W(t)) \left(-\frac{\partial \mathcal{F}_{jk}(s,t|Z_i)}{\partial t} \right) \lambda_0(s)e^{Z_i\beta}dsdt \\
&\left.\left. + \iint_{(0,C]^2} \mathcal{G}(s,t)(Z_i - W(s))(Z_i - W(t))\mathcal{F}_{jk}(s,t|Z_i)\lambda_0(s)e^{Z_i\beta}\lambda_0(t)e^{Z_i\beta}dsdt \right\} \right].
\end{aligned} \tag{A.9}
$$

The expression for $\mathcal{A}$ is likewise computed as

$$
\begin{aligned}
\mathcal{A} &= E \left\{ \sum_{j=1}^{J} \int_0^\infty \bar{Y}_{ij}(t) \left[\frac{\left(\sum_k s_{2k}(t;\beta) \right) \left(\sum_k s_{0k}(t;\beta) \right) - \left(\sum_k s_{1k}(t;\beta) \right)^2}{\left(\sum_k s_{0k}(t;\beta) \right)^2} \right] dN_{ij}(t) \right\} \\
&= E_{Z_i} \left\{ \sum_{j=1}^{J} \int_0^C \left[\frac{\left(\sum_k s_{2k}(t;\beta) \right) \left(\sum_k s_{0k}(t;\beta) \right) - \left(\sum_k s_{1k}(t;\beta) \right)^2}{\left(\sum_k s_{0k}(t;\beta) \right)^2} \right] \mathcal{G}(t)f_j(t|Z_i)dt \right\},
\end{aligned} \tag{A.10}
$$

where

$$
s_{0k}(t;\beta) = E\left(\bar{Y}_{ik}(t)\exp(Z_i\beta) \right) = E_{Z_i}\left(\mathcal{G}(t)\mathcal{F}_k(t|Z_i)\exp(Z_i\beta) \right) \tag{A.11}
$$

and

$$
s_{1k}(t;\beta) = s_{2k}(t;\beta) = E\left(\bar{Y}_{ik}(t)\exp(Z_i\beta)Z_i \right) = E_{Z_i}\left(\mathcal{G}(t)\mathcal{F}_k(t|Z_i)\exp(Z_i\beta)Z_i \right). \tag{A.12}
$$

Having expressions for $\mathcal{B}$ and $\mathcal{A}$ the asymptotic variance of $\widehat{\beta}$ can then be obtained and used for power and sample size calculations.

A.2 DERIVATION OF $\mathcal{B}$ WHEN CENSORING TIMES ARE INDEPENDENT WITHIN CLUSTERS

In the special case in which the censoring times are independent within clusters, the term $\mathcal{A}$ is unaffected. The computation of $E[\zeta_{ij}\zeta_{ik}]$ for $j \neq k$ and hence the derivation of $\mathcal{B}$ is, however, affected. In this case, we obtain

$$
\begin{aligned}
\mathcal{B} = \sum_{j=1}^{J} E_{Z_i} &\left[\int_0^C \mathcal{G}(s)(Z_i - W(s))^2 f_j(s|Z_i)ds \right] \\
+ \sum_{j \neq k} &\left[E_{Z_i} \left\{ \iint_{(0,C]^2} \mathcal{G}(s)\mathcal{G}(t)(Z_i - W(s))(Z_i - W(t))f_{jk}(s,t|Z_i)dsdt \right.\right. \\
&- \iint_{(0,C]^2} \mathcal{G}(s)\mathcal{G}(t)(Z_i - W(s))(Z_i - W(t)) \left(-\frac{\partial \mathcal{F}_{jk}(s,t|Z_i)}{\partial s} \right) \lambda_0(t)e^{Z_i\beta}dsdt \\
&- \iint_{(0,C]^2} \mathcal{G}(s)\mathcal{G}(t)(Z_i - W(s))(Z_i - W(t)) \left(-\frac{\partial \mathcal{F}_{jk}(s,t|Z_i)}{\partial t} \right) \lambda_0(s)e^{Z_i\beta}dsdt \\
&\left.\left. + \iint_{(0,C]^2} \mathcal{G}(s)\mathcal{G}(t)\mathcal{F}_{jk}(s,t|Z_i)(Z_i - W(s))(Z_i - W(t))\lambda_0(s)e^{Z_i\beta}\lambda_0(t)e^{Z_i\beta}dsdt \right\} \right],
\end{aligned} \tag{A.13}
$$

where the pairwise survivor function of the censoring times $\mathcal{G}(s,t)$ in (A.9) is simply replaced by $\mathcal{G}(s)\mathcal{G}(t)$ under the independent within-cluster censoring assumption.

APPENDIX B   LIMITING DISTRIBUTION OF WALD STATISTICS WITH CLUSTERED INTERVAL-CENSORED DATA

We assume again that $T_{ij}|Z_i$ satisfies the proportional hazards assumption in (2.1) with marginal distribution indexed by $\theta = (\alpha', \beta)'$ where α is a $q \times 1$ parameter vector. Consider a trial in which individuals are event-free at $a_0 = 0$, and are scheduled to be observed at R assessment times $a_1, \ldots, a_R$ over $(0, C^\dagger]$ where $a_R = C^\dagger$ and $a_{R+1} = \infty$. Let $Y_{ij} = (Y_{ij1}, \ldots, Y_{ijR})'$ denote the event time information provided by individual j in cluster i, where $Y_{ijr} = \mathrm{I}(a_{r-1} < T_{ij} \le a_r)$ indicates that the event was determined to have occurred in $(a_{r-1}, a_r]$; let $Y_i = (Y_{i1}', \ldots, Y_{iJ}')'$. Adopted the strategy in Kor et al. [38], the estimating function for parameter θ can be written as

$$U(\theta) = \sum_{i=1}^n U_i(\theta) = \sum_{i=1}^n D_i' V_i^{-1}(Y_i - \mu_i) \,,$$

where $\mu_i = E[Y_i|Z_i]$ is the conditional mean of $Y_i|Z_i$, $D_i = \partial \mu_i / \partial \theta'$, and V_i is the working matrix. Under the working independence assumption, V_i is a block diagonal matrix with the blocks $V_{ij} = \mathrm{Cov}(Y_{ij}, Y_{ij}'|Z_i)$, $j = 1, \ldots, J$, which accounts for the negative dependence between responses at different assessment times for each individual; that is,

$$V_i = \begin{pmatrix} \mathrm{Cov}(Y_{i1}, Y_{i1}'|Z_i) & & 0 \\ & \ddots & \\ 0 & & \mathrm{Cov}(Y_{iJ}, Y_{iJ}'|Z_i) \end{pmatrix} . \tag{B.1}$$

As stated in Section 4, the estimator $\widehat{\theta}$ is the root of $U(\theta) = 0$ and has asymptotically normal distribution,

$$n^{1/2}(\widehat{\theta} - \theta) \to N(0, \Gamma) \,,$$

where $\Gamma = \mathcal{A}^{-1}\mathcal{B}\left[\mathcal{A}^{-1}\right]'$. Hence the asymptotic distribution for β is

$$n^{1/2}(\widehat{\beta} - \beta) \to N(0, \Psi) \,, \tag{B.2}$$

where $\Psi = \Gamma[q+1, q+1]$.

The null and alternative hypotheses are $H_0 : \beta = \beta_0 = 0$ and $H_A : \beta \ne \beta_0$, respectively, and let β_A be the clinically important effect of interest. To derive the expression for $\mathcal{A}$ and $\mathcal{B}$, we note that

$$\mathcal{A} = E[-\partial U_i(\theta)/\partial \theta'] = E_{Z_i}[D_i' V_i^{-1} D_i] \,,$$

$$\mathcal{B} = E[U_i(\theta)U_i'(\theta)] = E[D_i' V_i^{-1}(Y_i - \mu_i)(Y_i - \mu_i)' V_i^{-1} D_i] = E_{Z_i}[D_i' V_i^{-1} W_i V_i^{-1} D_i] \,,$$

where $W_i = \mathrm{Cov}(Y_i, Y_i'|Z_i)$ is the full covariance matrix of Y_i, which accounts for both the within-cluster association between Y_{ij} and Y_{ik}, $j, k = 1, \ldots, J$ and the association within-individuals over time (i.e. between Y_{ijr} and Y_{ijs}, $r, s = 1, \ldots, R$) such that

$$W_i = \begin{pmatrix} \mathrm{Cov}(Y_{i1}, Y_{i1}'|Z_i) & \mathrm{Cov}(Y_{i1}, Y_{i2}'|Z_i) & \cdots & \mathrm{Cov}(Y_{i1}, Y_{iJ}'|Z_i)) \\ & \mathrm{Cov}(Y_{i2}, Y_{i2}'|Z_i) & \cdots & \mathrm{Cov}(Y_{i2}, Y_{iJ}'|Z_i) \\ & & \ddots & \vdots \\ & & & \mathrm{Cov}(Y_{iJ}, Y_{iJ}'|Z_i) \end{pmatrix} . \tag{B.3}$$

Note that

$$\mathrm{Cov}(Y_{ij}, Y'_{ij}|Z_i) = \begin{pmatrix} \mathrm{Cov}(Y_{ij1}, Y_{ij1}|Z_i) & \mathrm{Cov}(Y_{ij1}, Y_{ij2}|Z_i) & \cdots & \mathrm{Cov}(Y_{ij1}, Y_{ijR}|Z_i) \\ & \mathrm{Cov}(Y_{ij2}, Y_{ij2}|Z_i) & \cdots & \mathrm{Cov}(Y_{ij2}, Y_{ijR}|Z_i) \\ & & \ddots & \vdots \\ & & & \mathrm{Cov}(Y_{ijR}, Y_{ijR}|Z_i) \end{pmatrix} , \quad (\mathrm{B}.4)$$

where

$$\mathrm{Cov}(Y_{ijr}, Y_{ijr}|Z_i) = \mu_{ijr}(1 - \mu_{ijr}) , \quad \text{and} \quad \mathrm{Cov}(Y_{ijr}, Y_{ijs}|Z_i) = E[Y_{ijr}Y_{ijs}|Z_i] - \mu_{ijr}\mu_{ijs} = -\mu_{ijr}\mu_{ijs} ,$$
$$(\mathrm{B}.5)$$

$j = 1, \ldots, J$. The covariance between Y_{ij} and Y'_{ik}, $j \neq k$, is more involved and makes use of the copula assumptions. Specifically,

$$\mathrm{Cov}(Y_{ij}, Y'_{ik}|Z_i) = E[Y_{ij}Y'_{ik}|Z_i] - \mu_{ij}\mu'_{ik}$$

$$= \begin{pmatrix} E[Y_{ij1}Y_{ik1}|Z_i] & E[Y_{ij1}Y_{ik2}|Z_i] & \cdots & E[Y_{ij1}Y_{ikR}|Z_i] \\ & E[Y_{ij2}Y_{ik2}|Z_i] & \cdots & E[Y_{ij2}Y_{jkR}|Z_i] \\ & & \ddots & \vdots \\ & & & E[Y_{ijR}Y_{ikR}|Z_i] \end{pmatrix} - \mu_{ij}\mu'_{ik} , \quad (\mathrm{B}.6)$$

where

$$E[Y_{ijr}Y_{iks}|Z_i] = \mathcal{F}(a_{r-1}, a_{s-1}|Z_i) - \mathcal{F}(a_{r-1}, a_s|Z_i) - \mathcal{F}(a_r, a_{s-1}|Z_i) + \mathcal{F}(a_r, a_s|Z_i) , \quad (\mathrm{B}.7)$$

can be calculated based on the copula model. By plugging (B.4) and (B.6) into (B.1) and (B.3), we obtain the expression for V_i and W_i, and hence we can obtain $\mathcal{A}$ and $\mathcal{B}$. On the basis of the asymptotic property of the Wald statistic (B.2), we derive the sample size criteria (4.6).

ACKNOWLEDGEMENTS

REFERENCES

[1] Edwards SJ, Braunholtz DA, Lilford RJ and Stevens AJ. Ethical issues in the design and conduct of cluster randomised controlled trials. *British Medical Journal* 1999; **318**: 1407–1409.

[2] Torgerson DJ. Contamination in trials: is cluster randomisation the answer? *British Medical Journal* 2001; **322**: 355–357.

[3] Silverman WK, Kurtines WM, Ginsburg GS, Weems CF, Lumpkin PW, Carmichael D. Treating anxiety disorders in children with group cognitive-behavioral therapy: A randomized clinical trial. *Journal of Consulting and Clinical Psychology* 1999, **67**, 995–1003.

[4] Hayes RJ, Alexander NDE, Bennett S, Cousens SN. Design and analysis issues in cluster-randomized trials of interventions against infectious diseases. *Statistical Methods for Medical Research* 2000, **9**, 95–116.

[5] Donner A, Klar N. *Cluster randomization trials in epidemiology: Theory and application. Journal of Statistical Planning and Inference* 1994; **42**: 37–56.

[6] Moerbeek M. Randomization of clusters versus randomization of persons within clusters: Which is preferable? *The American Statistician* 2005; **59**; 72–78.

[7] Cameron R, Brown KS, Best JA, Pelkman CL, Madill CL, Manske SR, Payne M. Effectiveness of a social influences smoking prevention program as a function of provider type, training method, and school risk. *American Journal of Public Health* 1999; **89**; 1827–1831.

[8] Shah S, Peat JK, Mazurski EJ, Wang H, Sindhusake D, Bruce C, Henry RL, Gibson PG. Effect of peer led programme for asthma education in adolescents: cluster randomised controlled trial. *British Medical Journal* 2001; **322**, 583.

[9] Martin CM, Doig GS, Heyland DK, Morrison T, Sibbald WJ. Multicentre, cluster-randomized clinical trial of algorithms for critical-care enteral and parenteral therapy (ACCEPT). *Canadian Medical Association Journal* 2004; **170**; 197–204.

[10] Campbell M, Fitzpatrick R, Haines A, Kinmonth AL, Sandercock P, Spiegelhalter D, Tyrer P. Framework for design and evaluation of complex interventions to improve health. *British Medical Journal* 2000; **321**: 694–696.

[11] Donner A, Klar N. The Design and Analysis of Cluster-Randomization Trials in Health Research. Oxford University Press, 2000, New York.

[12] Lee EW, Wei LJ, Amato DA. Cox-type Regression analysis for large numbers of small groups of correlated failure time observations. In *Survival Analysis: State of the Art*, Klein, JP., Goel, PK. (eds). Kluwer Academic Publishers: Dordrect, 1992; 237–247.

[13] Le CT, Lindgren BR. Duration of ventilating tubes: a test for comparing two clustered samples of censored data. *Biometrics* 1996; **52**: 328–334.

[14] Manatunga AK, Chen S. Sample size estimation for survival outcomes in cluster-randomized studies with small cluster sizes. *Biometrics* 2000; **56**: 616–621.

[15] Lord SR, Castell S, Corcoran J, Dayhew J, Matters B, Shan A, Williams P. The effect of group exercise on physical functioning and falls in frail older people living in retirement villages: a randomized, controlled trial. *Journal of the American Geriatrics Society* 2003; **51**: 1685–1692.

[16] Bogner HR, Morales KH, Post EP, Bruce ML. Diabetes, depression, and death : A randomized controlled trial of a depression treatment program for older adults based in primary care (PROSPECT). *Diabetes Care* 2007; **30**: 3005–3010.

[17] Kramer MS, Chalmers B, Hodnett ED, Sevkovskaya Z, Dzikovich I, Shapiro S, Collet JP, Vanilovich I, Mezen I, Ducruet T, Shishko G, Zubovich V, Mknuik D, Gluchanina E, Dombrovskiy V, Ustinovitch A, Kot T, Bogdanovich N, Ovchinikova L, Helsing E. for the PROBIT Study Group. Promotion of breastfeeding intervention trial (PROBIT). *Journal of the American Medical Association* 2001; **285**: 413–420.

[18] Bellamy SL, Li Y, Ryan LM, Lipsitz S, Canner MJ, Wright R. Analysis of clustered and interval censored data from a community-based study in asthma. *Statistics in Medicine* 2004; **23**: 3607–3621.

[19] Glidden DV, Vittinghoff E. Modelling clustered survival data from multicentre clinical trials. *Statistics in Medicine* 2004; **23**: 369–388.

[20] Neuhaus JM, Kalbfleisch JD, Hauck WW. A comparison of cluster-specific and population-averaged approaches for analyzing correlated binary data. *International Statistical Review* 1991; 25–35.

[21] Neuhaus JM, Kalbfleisch JD. Between-and within-cluster covariate effects in the analysis of clustered data. *Biometrics* 1998; **54**: 638–645.

[22] Zeger SL, Liang KY. Longitudinal data analysis for discrete and continuous outcomes. *Biometrika* 1986; **73**: 121–130.

[23] Wei LJ, Lin DY, Weissfeld L. Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *Journal of the American Statistical Association* 1989; **84**: 1065–1073.

[24] Lin DY. Cox regression analysis of multivariate failure time data: the marginal approach. *Statistics in Medicine* 1994; **13**: 2233–2247.

[25] Cornfield J. Randomization by group: A formal analysis. *American Journal of Epidemiology* 1978; **108**: 100–102.

[26] Donner A, Birkett N, Buck C. Randomization by cluster: Sample size requirements and analysis. *American Journal of Epidemiology* 1981; **114**: 906–914.

[27] Lee EW, Dubin N. Estimation and sample size considerations for clustered binary responses. *Statistics in Medicine* 1994; **13**: 1241–1252.

[28] Hayes RJ, Bennett S. Simple sample size calculation for cluster-randomized trials. *International Journal of Epidemiology* 1999; **28**: 319–326.

[29] Jahn-Eimermacher A, Ingel K, Schneider A. Sample size in cluster-randomized trials with time to event as the primary endpoint. *Statistics in Medicine* 2013; **32**: 739–751.

[30] Jung SH. Sample size calculation for weighted rank tests comparing survival distributions under cluster randomization: a simulation method. *Journal of Biopharmaceutical Statistics* 2007; **17**: 939–849.

[31] Moerbeek M. Sample size issues for cluster randomized trials with discrete-time survival endpoints. *Methodology* 2012; **8**: 146–158.

[32] Xie T, Waksman J. Design and sample size estimation in clinical trials with clustered survival times as the primary endpoint. *Statistics in Medicine* 2003; **22**: 2835–2846.

[33] Joe H. *Multivariate Models and Multivariate Dependence Concepts*. Chapman and Hall. 1997, London.

[34] Nelsen RB. *An Introduction to Copulas*. Springer. 2006, New York.

[35] Genest C, MacKay J. The joy of copulas: bivariate distributions with uniform marginals. *The American Statistician* 1986; **40(4)**: 280–283.

[36] Sun J. *The Statistical Analysis of Interval-Censored Failure Time Data*. Springer. 2006, New York.

[37] Gladman DD, Farewell VT, Nadeau C. Clinical indicators of progression in psoriatic arthritis: Multivariate relative risk model. *The Journal of Rheumatology* 1995; **22**: 675–679.

[38] Kor CT, Cheng KF, Chen YH. A method for analyzing clustered interval-censored data based on Cox's model. *Statistics in Medicine* 2013; **32**: 822–832.

[39] Korendijk EJH, Moerbeek M, Maas CJM. The robustness of designs for trials with nested data against incorrect initial intracluster correlation coefficient estimates. *Journal of Educational and Behavioral Statistics* 2010; **35**: 556–585.

[40] Cook RJ, Bergeron PJ, Boher JM, Liu Y. Two-stage design of clinical trials involving recurrent events. *Statistics in Medicine* 2009; **28**: 2617–2638.

[41] Donner A. Approaches to sample size estimation in the design of clinical trials-A review. *Statistics in Medicine* 1984; **3**: 199–214.

[42] Manatunga AK, Hudgens MG, Chen S. Sample size estimation in cluster randomized studies with varying cluster size. *Biometrical Journal* 2001; **43**: 75–86.

[43] Van Breukelen GJ, Candel MJ, Berger MP. Relative efficiency of unequal versus equal cluster sizes in cluster randomized and multicentre trials. *Statistics in Medicine* 2007; **26**:2589–2603.

[44] Candel MJ, Van Breukelen GJ. Sample size adjustments for varying cluster sizes in cluster randomized trials with binary outcomes analyzed with second-order PQL mixed logistic regression. *Statistics in Medicine* 2010; **29**: 1488–1501.

[45] Hemming K, Girling AJ, Sitch AJ, Marsh J, Lilford RJ. Sample size calculations for cluster randomised controlled trials with a fixed number of clusters. *BMC Medical Research Methodology* 2011; **11**: 102.

Web-based Supplementary Materials for

Sample size and robust marginal methods for cluster-randomized trials with censored event times

YUJIE ZHONG

Department of Statistics and Actuarial Science,

University of Waterloo, Waterloo, ON, N2L 3G1, Canada

RICHARD J COOK

Department of Statistics and Actuarial Science,

University of Waterloo, Waterloo, ON, N2L 3G1, Canada

E-mail: rjcook@uwaterloo.ca

S1. Survivor and hazard functions for proportional hazard model

The baseline survivor and hazard functions of the proportional hazards model adopted in Section 2.3 are displayed in Figure 1S.

Figure 1S: The marginal survivor and hazard functions for the control group ($Z = 0$) when the baseline hazard is decreasing ($\kappa = 0.75$), constant ($\kappa = 1.0$), and increasing ($\kappa = 1.25$).
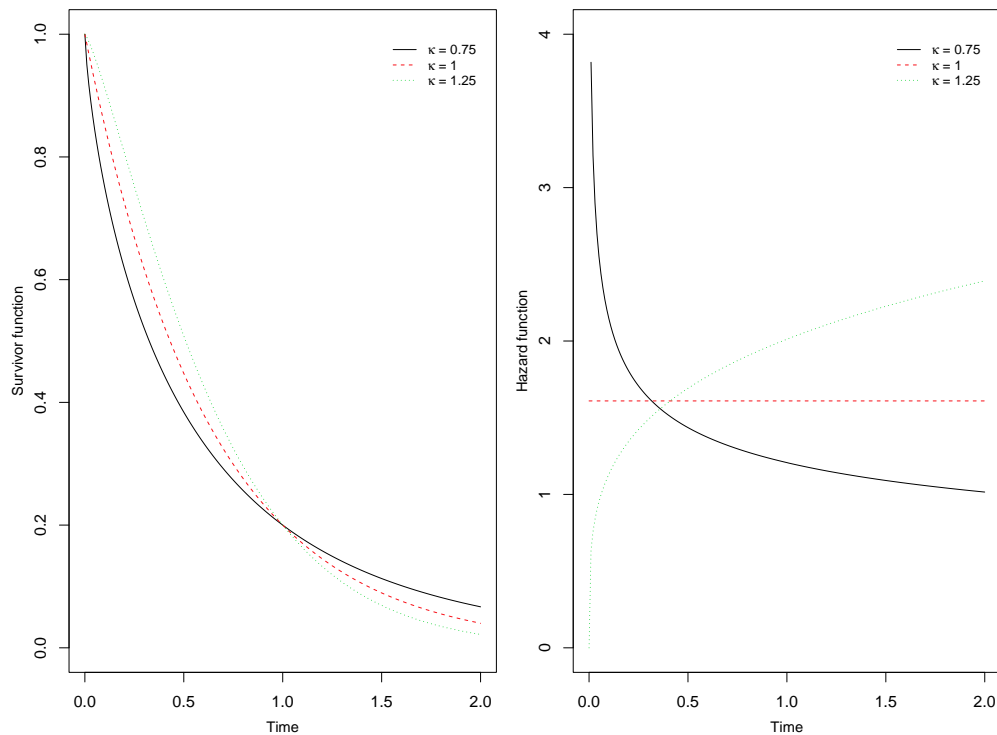
S2. Additional Simulation Results

The main body of the manuscript reported results when the sample size calculations and empirical investigations were carried out with a decreasing Weibull baseline hazard with $\kappa = 0.75$. Analogous results are provided here for the case of an increasing ($\kappa = 1.25$) and constant ($\kappa = 1.0$) baseline hazard; see Table 1S.

| | | | $p_0 = 0.2$ | | | | | | | $p_0 = 0.5$ | | | | | | |
| | | | $\beta = 0$ | | | $\beta = \beta_A$ | | | | $\beta = 0$ | | | $\beta = \beta_A$ | | | |
J	τ	n	ESE	ASE	REJ%	ESE	ASE	ECP%	REJ%	n	ESE	ASE	REJ%	ESE	ASE	ECP%	REJ%
									$\kappa = 1.25$								
2	0.05	433	0.079	0.079	5.2	0.080	0.081	95.8	79.8	676	0.078	0.079	4.7	0.081	0.081	94.7	79.8
	0.10	464	0.080	0.079	5.4	0.081	0.081	94.8	79.3	707	0.081	0.079	5.8	0.083	0.081	93.7	77.5
	0.25	556	0.077	0.079	4.5	0.081	0.081	94.5	80.5	800	0.081	0.079	5.5	0.080	0.081	95.8	80.0
5	0.05	211	0.078	0.079	5.2	0.079	0.080	94.8	79.8	308	0.079	0.079	5.4	0.083	0.081	94.5	78.5
	0.10	262	0.080	0.079	5.0	0.080	0.080	94.5	80.5	359	0.079	0.079	4.5	0.081	0.081	95.2	78.2
	0.25	409	0.080	0.079	5.2	0.081	0.080	95.0	79.8	508	0.078	0.079	5.2	0.082	0.081	94.4	78.7
20	0.05	100	0.080	0.079	5.1	0.080	0.079	94.4	80.5	125	0.080	0.078	5.5	0.081	0.080	93.8	79.8
	0.10	160	0.078	0.079	5.5	0.081	0.079	94.3	80.3	185	0.079	0.079	5.3	0.080	0.080	95.0	79.8
	0.25	335	0.080	0.079	5.2	0.081	0.080	94.3	79.9	361	0.078	0.079	5.2	0.080	0.080	95.6	78.2
100	0.05	71	0.078	0.078	5.3	0.080	0.079	95.0	78.9	76	0.080	0.078	5.7	0.079	0.078	94.6	80.4
	0.10	133	0.080	0.079	5.3	0.079	0.079	95.0	79.3	138	0.078	0.079	5.4	0.081	0.079	94.3	80.7
	0.25	316	0.081	0.079	5.3	0.080	0.079	94.8	80.7	322	0.080	0.079	4.6	0.079	0.080	95.5	80.0
									$\kappa = 1.0$								
2	0.05	433	0.077	0.079	4.5	0.080	0.081	95.3	79.5	676	0.079	0.079	4.5	0.082	0.081	94.6	77.6
	0.10	464	0.080	0.079	5.3	0.081	0.081	95.0	77.1	708	0.079	0.079	5.7	0.081	0.081	95.0	78.6
	0.25	556	0.081	0.079	5.6	0.082	0.081	94.2	78.0	801	0.078	0.079	5.0	0.083	0.081	94.7	77.1
5	0.05	211	0.079	0.079	4.7	0.080	0.080	95.5	79.2	309	0.079	0.079	4.8	0.081	0.081	94.8	78.5
	0.10	262	0.081	0.079	5.8	0.082	0.080	94.2	80.0	359	0.078	0.079	5.2	0.081	0.081	95.2	78.5
	0.25	409	0.081	0.079	5.8	0.083	0.080	94.3	79.2	509	0.080	0.079	5.2	0.082	0.081	94.4	78.7
20	0.05	100	0.078	0.079	4.9	0.079	0.079	94.8	79.8	125	0.081	0.078	5.6	0.078	0.080	95.8	79.8
	0.10	160	0.079	0.079	4.4	0.080	0.080	95.1	79.0	185	0.079	0.079	5.0	0.080	0.080	94.8	78.2
	0.25	335	0.081	0.079	5.5	0.081	0.080	94.3	79.1	363	0.080	0.079	5.1	0.082	0.080	94.5	79.5
100	0.05	71	0.080	0.078	5.4	0.080	0.078	94.3	80.5	76	0.078	0.078	5.2	0.077	0.078	95.3	80.3
	0.10	133	0.079	0.079	4.8	0.079	0.079	95.0	80.2	138	0.080	0.079	5.2	0.081	0.079	94.2	79.5
	0.25	316	0.080	0.079	5.1	0.080	0.080	94.7	79.8	324	0.079	0.079	4.7	0.082	0.080	94.1	78.5

Table 1S: Sample size estimation and empirical properties of estimators under cluster-randomized designs when within-cluster association is induced by Clayton copula; $p_a = 0.20$, $\beta_A = \log 0.8$, nsim$= 2000$.

S3. Impact of Misspecification of Marginal Parameter

At the suggestion of a referee we further explore the effect of misspecification here by considering whether there is any impact of misspecifying the extent of trend in the baseline hazard function on sample size, when the expected number of events is correctly specified. We assume that the marginal distribution of $T_{ij}|Z_i$ is of the proportional hazards form (2.1), where the baseline hazard is $\lambda_0(s; \alpha)ds = d\Lambda_0(s; \alpha)$ with Weibull cumulative hazard $\Lambda_0(s; \alpha) = (\lambda_0 s)^\kappa$, $\alpha = (\lambda_0, \kappa)'$. As in Section 2, we focus on the test of $H_0 : \beta = 0$ vs. $H_A : \beta \neq 0$ and let β_A denote the minimal clinically important effect of interest. The sample size is determined to ensure $100(1 - \gamma_2)\% = 80\%$ power to reject H_0 at β_A, given the type I error rate $100\gamma_1\% = 5\%$.

If the administrative censoring rate p_a and net censoring rate p_0 are correctly specified but there is no useful pilot data on what κ values are appropriate, one might use $\kappa = 1.0$ to compute the

required number of clusters by (2.7) at $\beta_A = \log 0.8$. To explore sensitivity of the power to the parameter κ, with the derived number of clusters we next examine the theoretical power at different values of κ under an administrative censoring rate of $p_a = 0.2$ and net censoring rate of $p_0 = 0.2$ or 0.5 for the control group, we consider values of κ ranging from 0.5 to 1.5 and examine the impact of misspecification under the Clayton, Frank, and Gumbel copula functions. Figure 2S and Figure 3S show the power of such test when the sample size is calculated based on formula (2.7) by using $\kappa = 1.0$ for 20% and 50% net censoring rates, respectively. Settings with cluster sizes of 2 and 100 and weak ($\tau = 0.1$) and moderate ($\tau = 0.25$) degrees of within cluster association are considered. As can be seen from Figure 2S, when there is only administrative censoring there is no impact on power from misspecification of κ; all power functions are horizontal lines with value 0.8 for all copula models. When there is random censoring, Figure 3S indicates the effect of misspecifying the shape parameter. The effect of κ misspecification is smaller when $J = 2$ than when $J = 100$. Moreover the power is more robust to misspecification of the shape parameter under the Clayton copula than it is under the Frank and Gumbel copulas.

Figure 2S: Theoretical power as a function of κ for trials designed with the correct values of $p_a = p_0 = 0.2$ but sample size is determined under the assumption $\kappa = 1$.
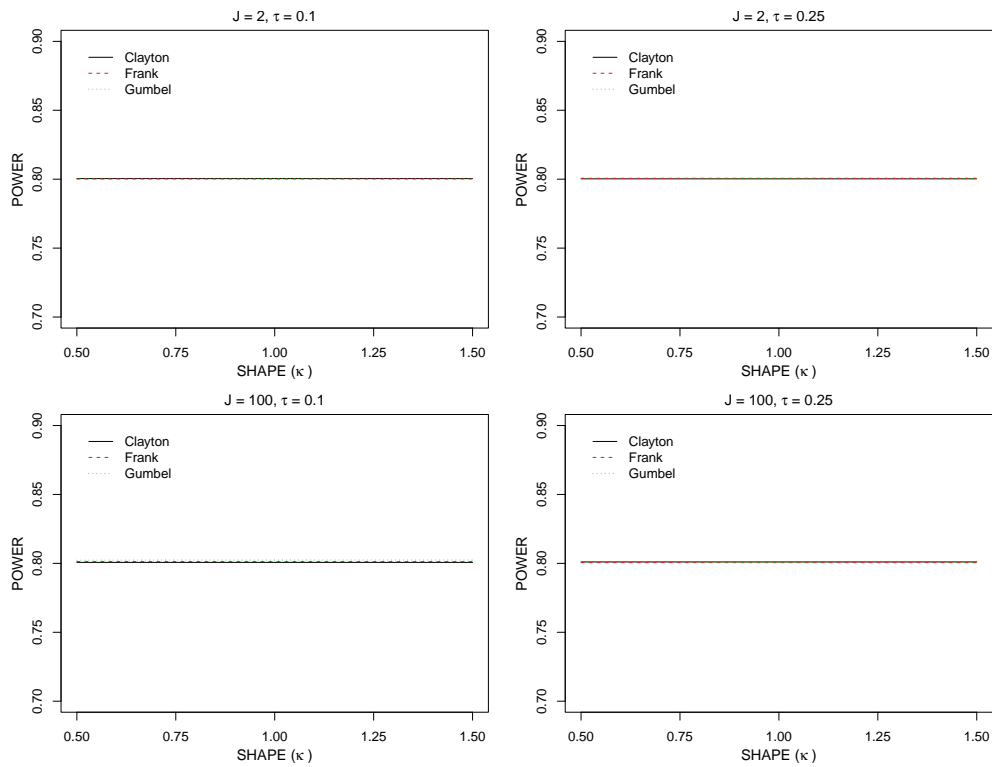
Figure 3S: Theoretical power as a function of κ for trials designed with the correct values of $p_a = 0.2$ and random censoring yielding a 50% net censoring rate ($p_0 = 0.50$), but sample size is determined under the assumption $\kappa = 1$.