Statistica Sinica **34** (2024), 181-199
doi:https://doi.org/10.5705/ss.202021.0349

TEST FOR INFORMATIVE CLUSTER SIZE WITH RIGHT CENSORED SURVIVAL DATA

Alessandra Meddis*[1,2] and Aurélien Latouche[2,3]

[1] *University of Copenhagen,* [2] *INSERM, U900*
and [3] *Conservatoire National des Arts et Métier*

Abstract: Clustered survival data often arise in biomedical research. When the outcome depends on the size of the cluster, the cluster size is said to be informative. Many studies assume a noninformative cluster size, even though it may not always be true. We propose a test for the assumption of informative cluster size in clustered survival data with right censoring. We use standard martingale results to obtain the asymptotic distribution of the test statistic. Simulation studies show that the proposed test works well under various scenarios. To illustrate the proposed approach, we consider several applications: periodontal data, a multicentric study of patients with liver disease, and a recent data set of patients with metastatic cancer treated using immunotherapy.

Key words and phrases: Clustered data, hypothesis testing, informative cluster size, survival analysis.

1. Introduction

Clustered data are often encountered in biomedical research. Observations are organized within groups of various sample sizes, and while clusters are assumed to be independent, units within the same cluster are correlated because of some common shared features. Several methods have been proposed to handle clustered data, such as the frailty model and marginal models, but they assume that the outcome is unrelated to the sample size of the clusters. However, this assumption is not always valid, in which case the cluster size is said to be informative. For instance, consider the time to tooth loss in one individual with periodontal disease. Because the subject may already have lost some teeth from the disease, the time to loss in the individual (cluster) is linked to that individual's number of teeth (cluster size). Another example can be found in studies of men with lymphatic filariasis, which is characterized by one or more nests of adult filarial worms in the scrotum (Williamson et al. (2008)). Ideally, effective treatment would kill the worms in all of the nests. The nest-specific time to clear the worms is longer in men with multiple nests than in men with one nest. Moreover, informative cluster size (ICS) might be detected in a multicentric study or meta-analysis in which the size of the study is linked to the magnitude

*Corresponding author.

of the treatment effect.

The recent increase in interest on how to handle informative cluster size (Zhang et al. (2015); Chiang and Lee (2008); Pavlou and R. (2013)) had yielded several approaches. Hoffman, Sen and Weinberg (2001) proposed the within-cluster resampling (WCR) method, in which independent data sets are created by randomly sampling one observation from each cluster, with replacement. Williamson, Datta and Satten (2003) considered a generalized estimating equation method inversely weighted by the cluster sample sizes. Cong, Yin and Shen (2007) investigated the WCR method for clustered survival data by analyzing the resampled data sets using a Cox model. They also generalized the marginal models weighting the score function by using the inverse of the cluster sample size. Williamson et al. (2008) estimated the marginal distribution for multivariate survival data with ICS using cluster-weighted Weibull and Cox models. However, these methods all rely on the ICS assumption, without testing it in the study.

In practice, it is usually not possible to know in advance whether the ICS assumption is suitable for a particular data set. Although appropriate methods exist to handle this issue, unnecessarily allowing for ICS leads to a substantial loss of efficiency (Benhin, Rao and Scott (2005)). Therefore, assessing whether the cluster size is informative is fundamentally important for the decision on the statistical approach analysis. Although the nature of the the link between the cluster sample size and the outcome is interesting, we focus on a test for detecting informative cluster size to avoid possible bias in the analysis.

Benhin, Rao and Scott (2005) employed a Wald-type test for the ignorability of the cluster size in the estimating equations framework for linear and logistic regression models. Nevalainen, Oja and Datta (2017) introduced a test for ICS using a balanced bootstrap to estimate the null distribution. However, in survival analysis, testing procedures are limited to ad-hoc procedures that compare the marginal distributions between strata defined by the cluster size (Meddis et al. (2020)). We aim to provide a more general method for testing for ICS in right censored survival data. To do this, we consider two definitions for the estimator of the cumulative hazard function. The asymptotic distribution of the test statistic is obtained using standard martingale results.

The rest of this paper is organized as follows. In Section 2, we discuss the problem of informative cluster size and describe possible target populations in clustered data. We further introduce a new method for tetsing for ICS in right censored survival data, and provide the asymptotic distribution of the test statistic. In Section 3, we present a simulation study to assess the power of the test. In Section 4, we illustrate the usefulness of the method in several applications. Section 5 concludes the paper.

2. Methodology

2.1. Notations and assumptions

Let T_{ik} and C_{ik} be the time-to-event and the censoring time for unit i in cluster k, with K clusters with sample size N_k and $N = \sum_k N_k$. We observe the failure time $\tilde{T}_{ik} = \min\{T_{ik}, C_{ik}\}$ and the indicator of the event $\Delta_{ik} = \mathbf{I}(T_{ik} \leq C_{ik})$, where $\Delta_{ik} = 1$ if the event occurs, and zero otherwise. Let $(G_1, G_2, \ldots, G_K)$ be a sample of K independent and identically distributed (i.i.d.) observations, where each G_k represents a cluster consisting of $\{N_k, (\tilde{T}_{1k}, \Delta_{1k}), \ldots, (\tilde{T}_{N_k k}, \Delta_{N_k k})\}$. We assume that T_{ik} and C_{ik} are independent for all i, k and that T_{ik} are independent between clusters, but that within cluster k, $(T_{1k}, T_{2k}, \ldots, T_{N_k k})$ can be correlated.

2.2. Target population

When we have clustered data, the cluster sample size may provide information about the outcome. The cluster sizes may vary because of the study design used to collect the data, that is, because of an inherent feature of the data, or because of missing data. In the case of missing observations, we might be interested in the effect on the outcome for the complete cluster (i.e., observed and missing observations), and thus assumptions on the censoring are needed for inference. We assume there is independent censoring, and do not discuss the problem of missing data. In this context two marginal analyses that might be of interest (Hoffman, Sen and Weinberg (2001); Seaman, Pavlou and Copas (2014b)). One makes an inference for the population of *all observed members* (AOM), thus referring to a random individual in the observed population. The second, makes an inference for the population of *typical observed members of a typical cluster* (TOM), thus referring to a random individual belonging to a random cluster of the observed population. In the first case, larger clusters contribute more to inference, because equal weights are given to all observed members. In the second, clusters are equally weighted. For the AOM, the analysis provides an interpretation for a unit randomly sampled from the overall observed population. A TOM analysis has a cluster-based interpretation, that is an interpretation for a randomly selected unit sampled from a randomly selected cluster. Asymptotically, the two marginal analyses reach the same conclusion if the cluster size is unrelated to the outcome (Seaman, Pavlou and Copas (2014a)). However, they differ, in general, in the presence of ICS.

For each cluster k, let r be the index of a randomly selected member of the observed cluster. As in Seaman, Pavlou and Copas (2014b), we define $e_{AOM} = \mathbf{E}[N_k T_r | N_k \geq 1] / \mathbf{E}[N_k | N_k \geq 1]$ and $e_{TOM} = \mathbf{E}[T_r | N_k \geq 1]$. When $\mathbf{E}[T_r | N_k = n] = \mathbf{E}[T_r | N_k \geq 1]$ we have noninformative cluster size (NICS), otherwise the cluster size is informative (Hoffman, Sen and Weinberg (2001)). In general, ICS refers to any violation of the condition $\mathbf{P}(T_{ik} \leq t | N_k = n) = \mathbf{P}(T_{ik} \leq t)\ \forall n$. Under NICS, the two marginal analyses coincide ($e_{TOM} = e_{AOM}$). However, this

is not true, in general, when the cluster size is informative. Thus, when choosing a method for analysis, it is important to specify in advance which target population would best address the scientific question.

2.3. Definition of the test

Let $\mathcal{N}_{ik}(t) = \mathbf{I}(\tilde{T}_{ik} \leq t, \Delta_{ik} = 1)$ be the counting process at time t, with intensity $\lambda_{ik}(t) = \alpha_{ik}(t)Y_{ik}(t)$, where $Y_{ik}(t) = I(\tilde{T}_{ik} \geq t)$ represents the at-risk process and $\alpha(t)$ is the hazard function. Given the cumulative intensity function $\Lambda_{ik}(t) = \int_0^t \lambda_{ik}(s)ds$, we define $M_{ik}(t) = \mathcal{N}_{ik}(t) - \Lambda_{ik}(t)$, or equivalently $d\mathcal{N}_{ik}(t) = \alpha_{ik}(t)Y_{ik}(t)dt + dM_{ik}(t)$. The quantity $M_{ik}(t)$ is not a martingale with respect to the joint filtration generated by all the times, because of the correlation within the clusters. However, it is a martingale with respect to the filtration $\mathcal{F}_{ik}(t) = \sigma\{\mathcal{N}_{ik}(u), Y_{ik}(u) : 0 \leq u \leq t\}$. Moreover, we define the Nelson-Aalen estimator of the cumulative hazard function $A(t) = \int_0^t \alpha(s)ds$ for the two marginal analyses as follows:

$$\hat{A}_{TOM}(t) = \int_0^t \frac{d\mathcal{N}_{TOM}(s)}{Y_{TOM}(s)}$$
$$\hat{A}_{AOM}(t) = \int_0^t \frac{d\mathcal{N}_{AOM}(s)}{Y_{AOM}(s)}$$

where $\hat{A}_{TOM}(t)$ estimates the number of events for a typical observed member, and $\hat{A}_{AOM}(t)$ estimates the number of events for all observed member populations. In fact, the weighted counting process and at-risk process are defined as:

$$\mathcal{N}_{TOM}(t) = \frac{1}{K}\sum_k \frac{1}{N_k}\sum_i \mathcal{N}_{ik}(t)$$
$$Y_{TOM}(t) = \frac{1}{K}\sum_k \frac{1}{N_k}\sum_i Y_{ik}(t)$$

where units within a cluster are equally weighted by the inverse of the cluster sample size, and

$$\mathcal{N}_{AOM}(t) = \frac{1}{N}\sum_k \sum_i \mathcal{N}_{ik}(t)$$
$$Y_{AOM}(t) = \frac{1}{N}\sum_k \sum_i Y_{ik}(t)$$

where equal weights are given to each unit, regardless of the cluster to which they belong. The above estimators are consistent estimators for the cumulative hazard functions even though the data are clustered and the observations are dependent within each cluster (Ying and Wei (1994)).

Let τ be the follow-up time. In order to define the null hypothesis of the

test, we rely on the fact that under NICS, the two marginal analyses coincide:

$$H_0 : \alpha_{TOM}(t) = \alpha_{AOM}(t)\ \forall t \in [0,\tau]$$
$$H_1 : \alpha_{TOM}(t^*) \neq \alpha_{AOM}(t^*) \text{ in } t^* \in [0,\tau].$$

The proposed test statistic is

$$Z(\tau) = \int_0^\tau L(t)\{d\hat{A}_{TOM}(t) - d\hat{A}_{AOM}(t)\},$$

where $L(t) = Y_{AOM}(t)Y_{TOM}(t)/K$ is a weight function defined to ensure convergence. Under the null hypothesis, $Z(\tau)/\sqrt{K}$ asymptotically tends to a Gaussian distribution with mean zero and covariance matrix V.

2.3.1. Proof of asymptotic distribution

By definition:

$$Z(\tau) = \int_0^\tau L(t)\left(\frac{d\mathcal{N}_{TOM}(t)}{Y_{TOM}(t)} - \frac{d\mathcal{N}_{AOM}(t)}{Y_{AOM}(t)}\right),$$

where $d\mathcal{N}_h(t) = dM_h(t) + \alpha_h(t)Y_h(t)dt$.

Therefore

$$\begin{aligned} Z(\tau) &= \int_0^\tau L(t)\left(\frac{dM_{TOM}(t) + \alpha_{TOM}(t)Y_{TOM}(t)}{Y_{TOM}(t)}\right) \\ &\quad - \left(\frac{dM_{AOM}(t) + \alpha_{AOM}(t)Y_{AOM}(t)}{Y_{AOM}(t)}\right) \\ &= \int_0^\tau L(t)\left(\frac{dM_{TOM}(t)}{Y_{TOM}(t)} - \frac{dM_{AOM}(t)}{Y_{AOM}(t)}\right) + \int_0^\tau L(t)\{\alpha_{TOM}(t) - \alpha_{AOM}(t)\}dt. \end{aligned}$$

Under the null hypothesis $\alpha_{TOM}(t) = \alpha_{AOM}(t)\ \forall t \in [0,\tau]$, and by the definition of $\mathcal{N}_h(t)$, $dM_{TOM}(t) = \sum_k (1/N_k)\sum_i dM_{ik}(t)$ and $dM_{AOM}(t) = \sum_k \sum_i dM_{ik}(t)$.

We specify $L(t) = Y_{AOM}(t)Y_{TOM}(t)/K$, and obtain

$$\begin{aligned} Z(\tau) &= \int_0^\tau \frac{L(t)}{Y_{TOM}(t)} \sum_{k=1}^{K} \frac{1}{N_k} \sum_{i=1}^{N_k} dM_{ik}(t) - \int_0^\tau \frac{L(t)}{Y_{AOM}(t)} \sum_{k=1}^{K}\sum_{i=1}^{N_k} dM_{ik}(t) \\ &= \int_0^\tau \frac{Y_{AOM}(t)}{K} \sum_{k=1}^{K} \frac{1}{N_k} \sum_{i=1}^{N_k} dM_{ik}(t) - \int_0^\tau \frac{Y_{TOM}(t)}{K} \sum_{k=1}^{K}\sum_{i=1}^{N_k} dM_{ik}(t) \end{aligned}$$

We can interchange the sums and the integral:

$$Z(\tau) = \sum_{k=1}^{K} \frac{1}{N_k} \sum_{i=1}^{N_k} \int_0^\tau \frac{Y_{AOM}(t)}{K} dM_{ik}(t) \ - \sum_{k=1}^{K}\sum_{i=1}^{N_k} \int_0^\tau \frac{Y_{TOM}(t)}{K} dM_{ik}(t)$$

$$= \sum_{k=1}^{K} \frac{1}{N_k} \int_0^\tau \frac{Y_{AOM}(t)}{K} dM_k(t) \; - \sum_{k=1}^{K} \int_0^\tau \frac{Y_{TOM}(t)}{K} dM_k(t)$$

where $M_k(t) = \sum_{i=1}^{N_k} M_{ik}(t)$. Thus, the statistic can be rewritten as

$$Z(\tau)\frac{1}{\sqrt{K}} = \frac{1}{\sqrt{K}} \sum_{k=1}^{K} \int_0^\tau \left(\frac{Y_{AOM}(t)}{N_k K} - \frac{Y_{TOM}(t)}{K} \right) dM_k(t).$$

Because of the dependence between observations, we cannot use to the usual martingale theory to prove the asymptotic normality. However, we assume that observations are correlated within a cluster, and the N_k is finite. Thus $\{T_{ik}\}$ is an m-dependent sequence (with $m = \max_k\{N_k\}$) because $\{T_{i1}, T_{i2}, \ldots, T_{iN_k}\}$ and $\{T_{i1}, T_{i2}, \ldots, T_{iN_k'}\}$ are independent classes of random variables for $k \neq k'$. Applying the same argument as in the proof of Theorem 2 of Ying and Wei (1994), the process $(1/\sqrt{K}) \sum_{k=1}^{K} \int_0^\tau dM_k(t)$ converges weakly to a zero-mean Gaussian process $U^Z(t)$.

Define $y_{AOM}(t)$ and $y_{TOM}(t)$ the limits of $Y_{AOM}(t)/N_k K$ and $Y_{TOM}(t)/K$ respectively, when $N \to \infty$. The quantity $\int_0^\tau |Y_{AOM}(t)/(N_k K) - Y_{TOM}(t)/K|$ is bounded away from infinity in N, and

$$Z(\tau)\frac{1}{\sqrt{K}} = \frac{1}{\sqrt{K}} \sum_{k=1}^{K} \int_0^\tau \left(\frac{Y_{AOM}(t)}{N_k K} - \frac{Y_{TOM}(t)}{K} \right) dM_k(t)$$

and

$$Z^*(\tau)\frac{1}{\sqrt{K}} = \frac{1}{\sqrt{K}} \sum_{k=1}^{K} \sum_{i=1}^{N_k} \int_0^\tau \{y_{AOM}(t) - y_{TOM}(t)\} dM_{ik}(t)$$

converge almost surely to the same limit $\int_0^\tau \{y_{AOM}(t) - y_{TOM}(t)\} dU^Z(t)$ (as in Lee, Wei and Ying (1993)).

Hence, the statistic converges to a Gaussian with mean zero and covariance matrix V, asymptotically equivalent to $V^* = (1/K) \sum_k \sum_j \sum_{j'} \epsilon_{jk} \epsilon_{j'k}$ with $\epsilon_{jk} = \int_0^\tau (y_{AOM}(t) - y_{TOM}(t)) dM_{jk}(t)$, where $dM_{jk}(t) = dN_{jk}(t) - dA(t) Y_{jk}(t)$. We can estimate the covariance by replacing $dA(t)$ with $d\{\sum_{m=1}^{K} \sum_{f=1}^{N_m} N_{fm}(t)\}$ $\{\sum_{m=1}^{K} \sum_{f=1}^{N_m} Y_{fm}(t)\}^{-1}$ and $(y_{AOM}(t) - y_{TOM}(t))$ with $\hat{\omega}_k(t) = \{Y_{AOM}(t)/(KN_k) - Y_{TOM}(t)/K\}$:

$$\hat{\epsilon}_{jk} = \Delta_{jk} \hat{\omega}_k(T_{jk}) - \sum_{i=1}^{K} \sum_{l=1}^{N_i} \frac{\Delta_{li} \hat{\omega}_k(T_{li}) Y_{jk}(T_{li})}{\sum_{m=1}^{K} \sum_{f=1}^{N_m} Y_{fm}(T_{li})}.$$

2.4. Extension to regression setting

One might be interested in investigating the assumption of the dependence of the failure times on the cluster sample size given a set of covariates. Let

X_{ik} denote the covariate values for individual i in cluster k. We define NICS when $\mathbb{P}(T_{ik} \leq t|X_{ik}, N_k = n) = \mathbb{P}(T_{ik} \leq t|X_{ik})\ \forall n$. The covariates X_{ik} can include a set of cluster- and/or individual-level covariates. In this context, we assume T_{ik} is independent of C_{ik} given X_{ik}, and possible correlation in $(T_{1k}, T_{2k}, \ldots, T_{N_k k})$ within each cluster k given the set of covariates. To model the hazard conditional on the covariates, we consider the Cox model $\alpha_{ik}(t) = \alpha_0(t)\exp(\beta' X_{ik})$, and define $M_{ik}(t) = \mathcal{N}_{ik}(t) - \int_0^t \alpha_0(s)Y_{ik}(s)\exp(\beta' X_{ik})ds$. The proposed nonparametric test can than be extended to a regression setting by replacing the Nelson-Aalen estimator with the Breslow estimator of the cumulative baseline hazard function for the two marginal analyses:

$$\hat{A}_{TOM}(t, \hat{\beta}) = \int_0^t \frac{d\mathcal{N}_{TOM}(s)}{\overline{Y}_{TOM}(s, \hat{\beta})}$$
$$\hat{A}_{AOM}(t, \hat{\beta}) = \int_0^t \frac{d\mathcal{N}_{AOM}(s)}{\overline{Y}_{AOM}(s, \hat{\beta})},$$

where $\overline{Y}_{AOM}(t, \beta) = (1/N)\sum_{k=1}^{K}\sum_{i=1}^{N_k} Y_{ik}(t)\exp(\beta' X_{ik})$ and $\overline{Y}_{TOM}(t, \beta) = (1/K)\sum_{k=1}^{K}(1/N_k)\sum_{i=1}^{N_k} Y_{ik}(t)\exp(\beta' X_{ik})$. The regression coefficients β are estimated by solving the score function weighted by the inverse of the cluster sample size (Cong, Yin and Shen (2007)):

$$U(\beta) = \sum_{k=1}^{K}\frac{1}{N_k}\sum_{i=1}^{N_k}\int_0^\tau \left[X_{ik} - \frac{\sum_{j=1}^{K}(1/N_j)\sum_{l=1}^{N_j} Y_{lj}(t)X_{lj}\exp(\beta' X_{lj})}{\sum_{j=1}^{K}(1/N_j)\sum_{l=1}^{N_j} Y_{lj}(t)\exp(\beta' X_{lj})}\right] dN_{ik}(t)$$
$$= 0.$$

The test statistic is

$$Z^x(\tau) = \int_0^\tau L^x(t, \hat{\beta})\{d\hat{A}_{TOM}(t, \hat{\beta}) - d\hat{A}_{AOM}(t, \hat{\beta})\},$$
$$L^x(t, \hat{\beta}) = \frac{\overline{Y}_{AOM}(s, \hat{\beta})\overline{Y}_{TOM}(s, \hat{\beta})}{K}.$$

Under the null hypothesis, we obtain

$$Z^x(\tau)\frac{1}{\sqrt{K}} = \frac{1}{\sqrt{K}}\sum_{k=1}^{K}\int_0^\tau \left(\frac{\overline{Y}_{AOM}(s, \hat{\beta})}{N_k K} - \frac{\overline{Y}_{TOM}(s, \hat{\beta})}{K}\right) dM_k(t)$$

where $M_k(t) = \sum_{i=1}^{N_k} M_{ik}(t) = \sum_{i=1}^{N_k}\mathcal{N}_{ik}(t) - \int_0^t \alpha_0(s)Y_{ik}(s)\exp(\beta' X_{ik})ds$.

As in the previous section, the quantity $(1/\sqrt{K})\sum_{k=1}^{K}\int_0^\tau dM_k(t)$ converges weakly to a Gaussian process $U^Z(t)$, and a similar argument leads to the

asymptotic equivalence between

$$\frac{1}{\sqrt{K}}\sum_{k=1}^{K}\int_0^\tau \left(\frac{\overline{Y}_{AOM}(s,\hat{\beta})}{N_k K}-\frac{\overline{Y}_{TOM}(s,\hat{\beta})}{K}\right)dM_k^x(t)$$

and

$$\frac{1}{\sqrt{K}}\sum_{k=1}^{K}\sum_{i=1}^{N_k}\int_0^\tau w_k^\beta(t)dM_{ik}^x(t),$$

where $w_k^\beta(t)$ is the limit of $\overline{Y}_{AOM}(s,\hat{\beta})/(N_kK)-\overline{Y}_{TOM}(s,\hat{\beta})/K$.

Therefore, the statistic $Z^x(\tau)(1/\sqrt{K})$ converges to a Gaussian with mean zero and a covariance matrix asymptotically equivalent to $(1/K)\sum_k\sum_j\sum_{j'}\epsilon_{jk}\epsilon_{j'k}$, with $\epsilon_{jk}=\int_0^\tau\omega_k^\beta(t)dM_{jk}(t)$, estimated by $\hat{V}^x=(1/K)\sum_k\sum_j\sum_{j'}\hat{\epsilon}_{jk}\hat{\epsilon}_{j'k}$, where

$$\hat{\epsilon}_{jk}=\Delta_{jk}\hat{\omega}_k^\beta(T_{jk})-\sum_i\sum_l\frac{\Delta_{li}\hat{\omega}_k^\beta(T_{li})Y_{jk}(T_{li})\exp(\hat{\beta}X_{jk})}{\sum_m\sum_f Y_{fm}(T_{li})\exp(\hat{\beta}X_{fm})},$$

$$\hat{\omega}_k^\beta(t)=\left(\frac{\overline{Y}_{AOM}(t,\hat{\beta})}{KN_k}-\frac{\overline{Y}_{TOM}(t,\hat{\beta})}{K}\right).$$

3. Simulation Study

To evaluate the power and the nominal level of the test under different scenarios, we conduct a simulation study in which we fix the type-I error to 5%. The correlated failure times were generated from a frailty model, that is, from the conditional cumulative distribution function $P(T\le t|U_k,X)=1-\exp(-U_kA_0(t)\exp(\beta X))$ with the frailty term U_k and a Weibull baseline hazard function $A_0(t)=st^\omega(s=6.31e^{-6},\omega=4.6)$. To obtain informative cluster size, we generate K clusters with sample size $N_k\sim Pois\{\lambda\exp(V_k)\}$ where V_k defines the cluster-specific sample size, and λ represents the expected number of observations in each cluster if there were no variability. To create the dependence between the sample size N_k and the failure times T_{ik}, we generate (U_k,V_k) from a multivariate gamma distribution with unit mean and covariance matrix

$$\Sigma=\begin{pmatrix}\sigma_U^2 & \rho\sigma_V\sigma_U\\ \rho\sigma_V\sigma_U & \sigma_V^2\end{pmatrix}$$

The variance $\sigma_U^2=1/\theta$ controls the variability of the time-to-event among the clusters. The variance $\sigma_V^2=1/\gamma$ represents the variability between the clusters sample sizes. The parameter $\rho\in[0,1]$ is the correlation between the two random effects, and defines the dependence between T_{ik} and N_k; when $\rho=0$, we have NICS. The strength of ICS depends on θ,ρ and γ: it decreases with larger values of θ, because the between-cluster time-to-event variability decreases. Defining

the link between γ and ICS is not straightforward. We suspect that increasing γ decreases the variability and the ICS. However, there is a trade-off between the variability of the cluster sample sizes and the magnitude of the difference in the time-to-event, which also depends on θ (see the Supplementary Material).

We simulate two main settings: a) highly clustered data, with $K = 100, \lambda = 5$ and $\gamma = 20$, and b) a few large clusters with $K = 25, \lambda = 20$, and $\gamma = 3$. For both settings, λ and γ were defined by simulation (over 10,000 replications) to reach an overall sample size around 1,500. Right censoring is generated by a uniform distribution, independent of the failure times, with the parameters set to obtain 30% and 80% censoring.

3.1. Simulation plan 1: Without covariates

We fix $\beta = 0$ and let θ and γ vary to determine the behavior of the test in different frameworks. We consider uncensored data, with 30% and 80% right censoring.

In Figure 1, we provide the empirical power of the test for increasing correlation ρ. The simulations suggest that the test performs well, reaching a power of 80% in most scenarios. The results confirm that θ is inversely proportional to the ICS, showing higher power for $\theta = 5$. Moreover, we decrease the overall sample size $N = 700, 300$, varying either the number of clusters (K) or the cluster sample sizes (λ, γ). A decrease in the sample size does not seem to degrade the performance, overall (Figure 2). With a smaller λ, a lower θ is needed to detect ICS because the cluster sample sizes are smaller and the between-cluster variability is not sufficiently strong. However, for $K = 10$, even with decreasing θ, low power is detected. Thus, a sufficient number of clusters is necessary for the test to be valid. Our simulation results also suggest that censoring does not affect the performance of the test. However, for heavy censoring, we need a stronger variability ($\theta = 1$) to reach a good power for $N = 700$, because of the low number of events (see Figure 2). Finally, the empirical type-I error is reasonably close to the nominal level of 5% for scenarios A and B (Table 1). In the Supplementary the cluster sample size distribution is provided for the simulated settings with varying of ρ.

3.2. Simulation plan 2: Regression setting

Here, we fix $\theta = 5$ and assess the performance of the test by generating a continuous covariate with a normal distribution $N(0, 1)$. The covariate is generated independently of the cluster sample size. Thus, the ICS is not due to the introduction of X.

We simulate the data for $\beta = 0.5$ and 1.5 (Hazard ratio: 1.6 and 4.5) with no censoring, with 30% and 80% right censoring. We decrease the sample size to $N = 700, 300$, as in Simulation 1. Similar results are obtained, with the test

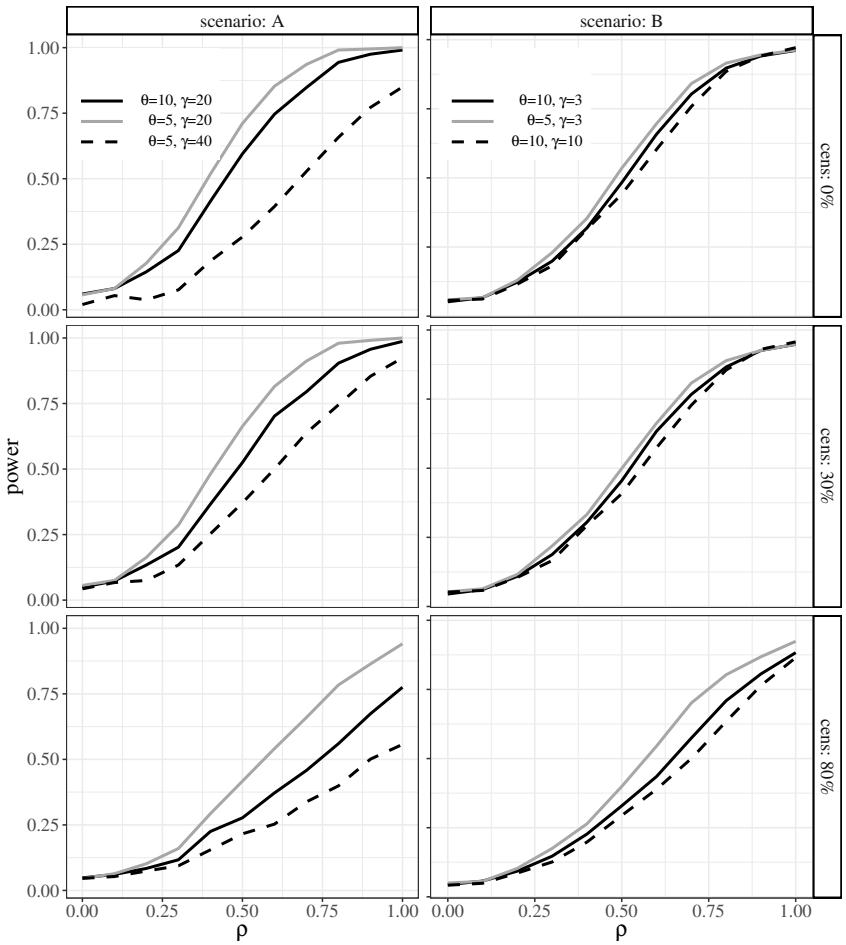

Figure 1. Power of the test for increasing correlation ρ for both scenarios considering different values of θ, γ and censoring. Each scenario is based on 1,000 replications, with $\alpha = 0.05$. Scenario A: highly clustered data ($K = 100$, $\lambda = 5$); scenario B: a few big clusters ($K = 25, \lambda = 20$).

performing well, overall (Figure 3). The low power of the test when $K = 10$ is confirmed. The nominal level is provided in Table 2.

4. Application

In this section we apply the test for ICS in different settings. Note that we are not interested in the subsequent analysis of the data, but rather in supporting our the theoretical findings and simulations.

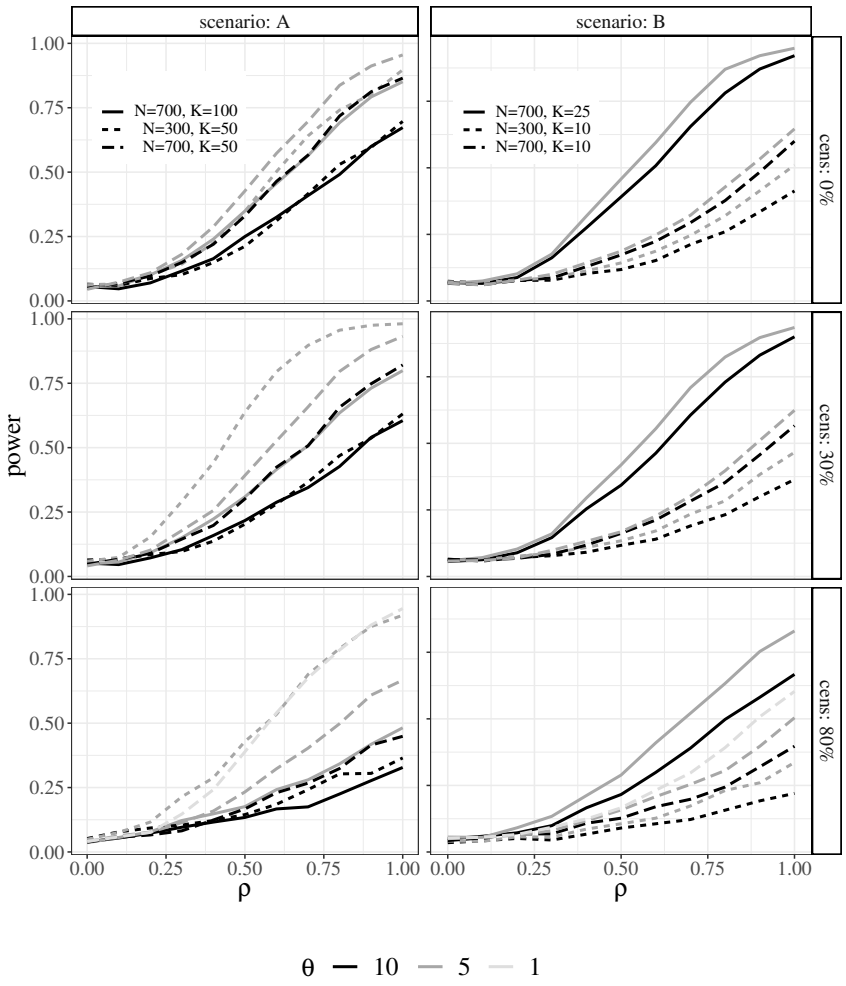

Figure 2. Power of the test for increasing correlation ρ for both scenarios with smaller sample size, varying K, λ, and censoring. Each scenario is based on 1,000 replications, with $\alpha = 0.05$.

4.1. Dental data

Here, we consider data of patients treated at the Creighton University School of Dentistry from August 2007 to March 2013. The data are available in the `MST` package in `R` as *Teeth* (Calhoun et al. (2018)). The analysis aims to construct multivariate survival trees to predict tooth loss. Data were collected for 5,336 patients with periodontal disease, yielding data on 65,228 teeth. We then excluded individuals with only one tooth, resulting in a sample size of 65,034 teeth. The average patient age is 58 years. Of the patients, 51% are women, 9% have diabetes mellitus, and 23% are smokers. The number of teeth that fell out is 4,334, with a median tooth loss time of 0.556 [0.003, 5.594] years. Several teeth

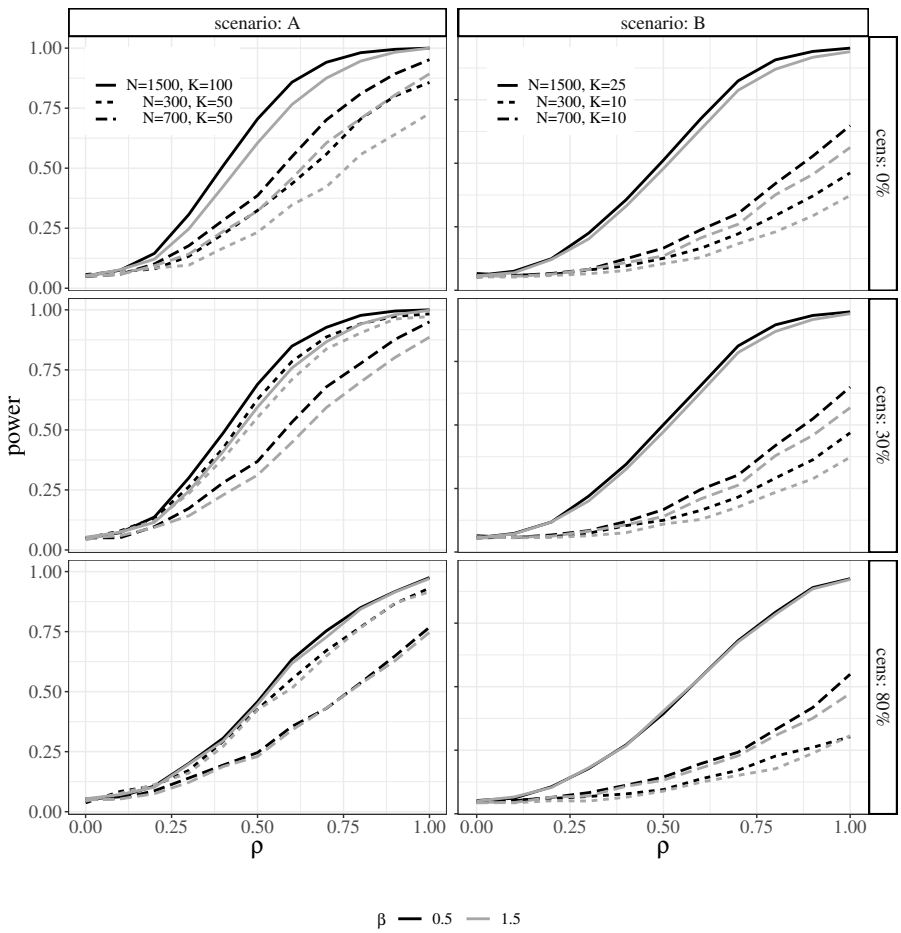

Figure 3. Power of the test for increasing correlation ρ with $X \sim N(0,1)$ for both scenarios considering different values of N, K, β and censoring. Each scenario is based on 1,000 replications, fixing $\alpha = 0.05$.

and individual characteristics are also provided in the data set, but we do not take them into consideration.

We suspect ICS exists because the number of teeth (cluster size) in each individual (cluster) is linked to the disease and, thus, a tooth is more likely to fall out in an individual with a smaller cluster size. The test shows clear evidence of ICS with a test statistic of 8.932 (p-value=0). We provide a plot of the Kaplan-Meier estimator of the survival function at each cluster sample size at the median time (Figure 4). The further suggests the existence of ICS: the tooth loss time is longer in individuals with more teeth (e.g., bigger cluster sample sizes). For instance, the probability of a tooth not falling out before the median time in individuals with 13 teeth is higher than that in individuals with

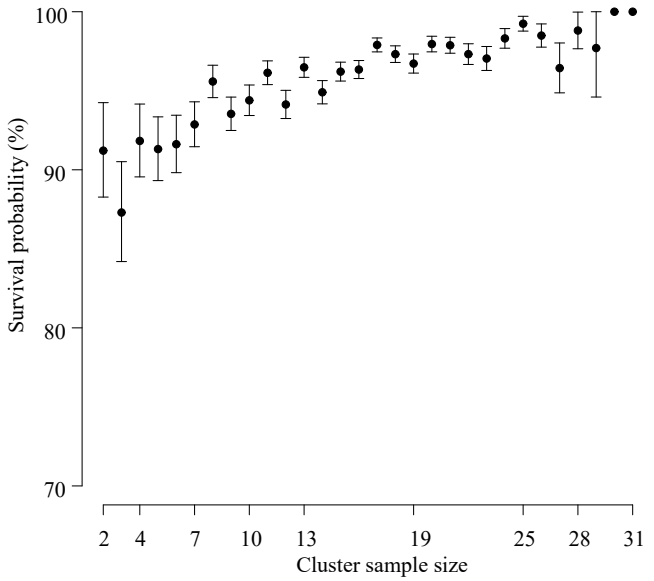

Figure 4. Estimated survival functions at the median failure time $t = 0.556$ years for an increasing cluster sample size. The confidence intervals for each probability are provided.

seven teeth. The assumption of ICS seems to be reasonable for these data. Thus, for the consequent analysis it would be appropriate to employ the WCR method (Cong, Yin and Shen (2007)) or the multivariate survival model proposed by Williamson et al. (2008).

4.2. Multicentric data

We consider data from a multicentric study of patients with the liver disease primary biliary cirrhosis (PBC). The original study was a randomized clinical trial conducted in six European hospitals between 1983 and 1987. The data are provided in the `pec` package in `R` as *Pbc3* (Gerds (2009)). A total of 349 patients were randomized to treatment with either Cyclosporin A (176 patients) or a placebo (173 patients) to study the effect of treatment on the composite outcome failure of medical treatment, defined as either death or liver transplantation. The data are characterized by 75% censoring, where 90 patients experienced the event, with a median time of 21 months $[0.8, 62]$.

We applied the proposed test for informative cluster size conditional on the treatment value. We reject the null hypothesis of NICS, with a test statistic equal to -1.98 (p-value=0.04). The K-M at the median time, varying the cluster sample size, is also provided in Figure 5. Because of the high censoring (75%), the weighted marginal survival model is preferable to the WCR methods for the analysis (Cong, Yin and Shen (2007)).

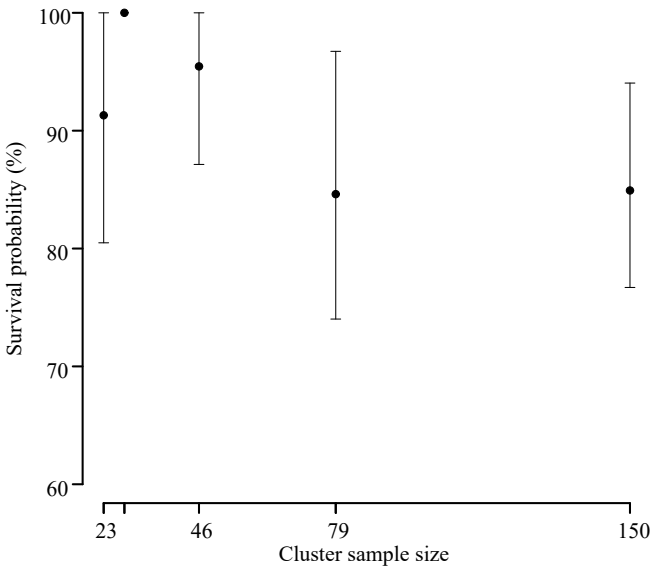

Figure 5. Estimated survival function at median failure time $t = 21$ months for an increasing cluster sample size in the subgroup of patients that received the treatment. The confidence intervals for each probability are provided.

4.3. Cancer data: Immunotherapy

We consider a data set of 100 patients with metastatic cancer treated using immunotherapy at the Institut Curie Comprehensive Cancer Center in Paris. For each patient, the size of each metastasis is evaluated radiologically from the treatment initiation to the date of progression of the specific metastasis. Immunotherapy may have different effects depending on the metastatic site. Furthermore, the treatment effect may depend on the number of metastases in the individual, which reflects the burden of the disease. A total of 272 metastases are examined, and each individual has from two to four metastases. For each subject, a maximum of five target metastases are considered, as per the Response Evaluation Criteria in Solid Tumors (RECIST) guideline (Nishino et al. (2010)). The primary cancer differed in nature, including breast cancer, head-neck cancer, lung cancer, urological cancer, and others. The principal objective of the study was gain insight into dissociate responses that are typical of immunotherapy, notably in the same individual, where the response to the treatment might vary among metastases.

The individual represents the cluster, and the number of metastases is the cluster sample size. The outcome of interest is the time to progression, which depends on the tumor growth. Intuitively, the number of metastases should affect the outcome. However, this was not confirmed by the proposed test, which did not reject the null hypothesis of NICS, with a test statistic of -0.85 (p-

Table 1. Nominal level of the test without covariates. Results are provided for 1,000 replications, with $\alpha = 0.05$.

					$\hat{\alpha}$		
N	K	λ	γ	θ	cens 0%	cens 30%	cens 80%
Scenario A							
1,500	100	5	20	10	**0.060**	**0.049**	**0.045**
	100	5	20	5	**0.057**	**0.056**	**0.048**
	100	5	40	10	**0.020**	**0.043**	**0.046**
700	100	2	20	10	**0.056**	**0.052**	**0.040**
	50	5	20	10	**0.049**	**0.047**	**0.038**
	100	2	20	5	**0.055**	**0.048**	**0.039**
	50	5	20	5	**0.045**	**0.041**	**0.036**
300	50	5	3	5	**0.063**	**0.059**	**0.049**
	50	5	3	10	**0.065**	**0.064**	**0.052**
Scenario B							
1,500	25	20	3	10	**0.052**	**0.045**	**0.050**
	25	20	10	10	**0.057**	**0.052**	**0.044**
	25	20	3	5	**0.059**	**0.053**	**0.042**
700	25	8	3	10	**0.069**	**0.056**	**0.050**
	10	20	3	10	**0.066**	**0.064**	**0.043**
	25	8	3	5	**0.065**	**0.059**	**0.049**
	10	20	3	5	**0.067**	**0.058**	**0.056**
300	10	8	3	5	**0.072**	**0.062**	**0.041**
	10	8	3	10	**0.074**	**0.065**	**0.035**

Table 2. Nominal level of the test with $X \sim N(0, 1)$. Results are provided for 1,000 replications, with $\alpha = 0.05$.

						$\hat{\alpha}$		
β	N	K	λ	γ	θ	cens 0%	cens 30%	cens 80%
0.5	1,500	25	20	3	5	**0.058**	**0.056**	**0.052**
		100	5	20	5	**0.051**	**0.049**	**0.053**
	700	10	20	3	5	**0.066**	**0.064**	**0.048**
		50	5	20	5	**0.049**	**0.050**	**0.052**
	300	50	5	3	5	**0.057**	**0.049**	**0.037**
		10	8	3	5	**0.052**	**0.055**	**0.044**
1.5	1,500	25	20	3	5	**0.057**	**0.056**	**0.050**
		100	5	20	5	**0.051**	**0.053**	**0.054**
	700	10	20	3	5	**0.057**	**0.055**	**0.043**
		50	5	20	5	**0.050**	**0.050**	**0.049**
	300	50	5	3	5	**0.053**	**0.042**	**0.044**
		10	8	3	5	**0.053**	**0.060**	**0.047**

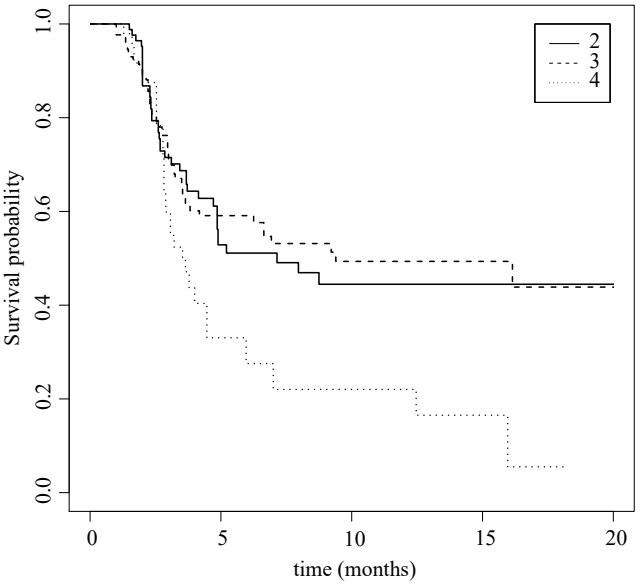

Figure 6. Estimated survival function for different number of metastases (cluster sample sizes).

value=0.39). However, the number of metastases seems to affect the survival function for metastasis disease progression (Figure 6). We computed the log-rank test for the three groups with different cluster sample sizes, finding a significant difference on survival (pvalue=0.008). This example illustrates a limitation of the proposed test when the time-to-event variability is not sufficient to detect ICS (simulation results for $K = 100, \lambda = 2$).

5. Discussion

In the presence of clustered data, standard statistical methods implicitly assume that the size of the clusters is unrelated to the outcome of interest. However, this assumption is not always true, in which case, the cluster size is defined to be informative. In this work, we propose a test for the assumption of ICS with right censored survival data that can be used as a pre-test to determine whether to use a standard regression model for clustered survival data, valid under the NICS assumption, or a method that accounts for the information carried by the cluster sample size. The test statistic relies on the fact that under NICS, the marginal analyses for typical observed member and all observed member coincide.

In Section 2, we mention that the variability in cluster sample sizes can be the result of missing data, notably when clusters have the same size, but some members are not observed (missing observations). Hoffman, Sen and Weinberg (2001) and Williamson, Datta and Satten (2003) state that the missing

completely at random (MCAR) mechanism is equivalent to NICS. Pavlou (2012) associated NICS with a missing data mechanism, of which MCAR is a special case, proving the equality of the results for the target populations in three cases (TOM, AOM, missing data). In this work, we assume that the observed clusters are complete. Therefore, we do not consider the problem of missing data, and do not discuss the issue of informative censoring. This is a challenging point that requires methods able to handle possible dependence between the censoring and the cluster sample size.

We have provided several applications where ICS is detected. Moreover, the publication bias that characterizes some meta-analyses also applies to the ICS problem, because the treatment effect is often linked to the study sample size. A funnel plot is often used to investigate the presence of publication bias or others forms of bias in a meta-analysis. It provides information on the treatment estimate against a measure of the study sample size. It provides a way of examining the tendency for smaller studies in a meta-analysis to show larger treatment effects, which is also a problem with ICS.

We also propose an extension of the test to a regression setting. In this case, the definition of NICS is extended to $\mathbf{P}(T_{ik} \leq t|X_{ik}, N_k = n) = \mathbf{P}(T_{ik} \leq t|X_{ik})\ \forall n$, and we use the Breslow estimator rather than the Neslon-Aalen estimator. A simulation study for a continuous covariate is conducted. We do not consider the regression setting with a binary covariate, because in this scenario, the nonparametric approach by stratification would be a better option, avoiding the problem of misspecification.

Our simulation results suggest that the proposed method performs well overall for both scenarios, with the test exhibiting low power when there are fewer than 10 clusters and for highly clustered data with small cluster sample sizes ($k = 100, \lambda = 2$ in the simulation). The proposed test detects whether there is dependence between the cluster sample size and the outcome. We do not focus on the nature of the association or on the several possible distributions in the generating method.

The test relies on the definition of the cumulative hazard estimator. Thus, extending the method to others survival analysis issues depends on appropriate modification of the Nelson-Aalen estimator. Moreover, in our simulation and applications, we refer to unit-level covariates. In the case of cluster-level covariates, the TOM and AOM definitions are still suitable.

In the simulation study, the covariate is generated independently of the cluster sample size (X is size-unbalanced). Other cases are possible in which: i) the cluster sample size affects the covariate distribution, but not the outcome, and thus the dependence on the cluster sample size is through X, and ii) the covariate effect varies with the cluster sample size,corresponding to an interaction between the cluster sample size and the covariate in the survival model. We do not consider these scenarios because they are related to the issue of informative

covariate structure and confounding by cluster as discussed in Pavlou (2012). ICS is defined mainly by the relationship between the cluster sample size and the outcome. We introduce an extension to a regression setting to show how to implement he method when the analysis requires adjustment for some covariates. Informative covariate structure is a related problem that can also occur without ICS, and it is left to future research.

A test for ICS has been introduced for clustered data for linear regression models using a balanced bootstrap method, because the distribution of the statistic under the null is analytically intractable (Nevalainen, Oja and Datta (2017)). An adaptation of this method to survival data could be employed to investigate the presence of ICS, but it is characterized by a high computational cost. Introducing the cluster sample size as a covariate in the regression model might offer another way to test for ICS. However, unlike for the proposed test, this would require assuming a specific link between the cluster sample size and the outcome. Furthermore, adding the cluster sample size to the model would test for ICS, but the estimated effect would be conditional on the sample size. Thus, we require a two-step procedure with appropriate methods for handling ICS are needed to obtain results on the marginal effect.

Supplementary Material

The implementation in `R` of the proposed method is provided at github-AMeddis (`https://github.com/AMeddis/Informative-Cluster-Size`), together with supplementary material on the simulation results referenced in Section 3.

Acknowledgments

The authors thank Christophe Le Tourneau, Pauline Vaflard, and Xavier Paoletti (Institut Curie, Paris, France) for providing the example data of patients with metastatic cancer treated using immunotherapy.

References

Benhin, E., Rao, J. and Scott, A. (2005). Mean estimating equation approach to analysing cluster-correlated data with nonignorable cluster sizes. *Biometrika* **92**, 435–450.

Calhoun, P., Su, X., Nunn, M. and Fan, J. (2018). Constructing multivariate survival trees: The MST package for R. *Journal of Statistical Software* **83**. Web: `https://doi.org/10.18637/jss.v083.i12`.

Chiang, C.-T. and Lee, K.-Y. (2008). Efficient estimation methods for informative cluster size data. *Statistica Sinica* **18**, 121–133.

Cong, X. J., Yin, G. and Shen, Y. (2007). Marginal analysis of correlated failure time data with informative cluster sizes. *Biometrics* **63**, 663–672.

Gerds, T. A. (2009). *Prediction error curves for survival models.* (R package pec. version 1.1.5). R Foundation for Statistical Computing.

Hoffman, E. B., Sen, P. K. and Weinberg, C. R. (2001). Within-cluster resampling. *Biometrika* **88**, 1121–1134.

Lee, E. W., Wei, L. and Ying, Z. (1993). Linear regression analysis for highly stratified failure time data. *Journal of the American Statistical Association* **88**, 557–565.

Meddis, A., Blanche, P., Bidard, F. C. and Latouche, A. (2020). A covariate-specific time-dependent receiver operating characteristic curve for correlated survival data. *Statistics in Medicine* **39**, 2477–2489.

Nevalainen, J., Oja, H. and Datta, S. (2017). Tests for informative cluster size using a novel balanced bootstrap scheme. *Statistics in Medicine* **36**, 2630–2640.

Nishino, M., Jagannathan, J. P., Ramaiya, N. H. and Van den Abbeele, A. D. (2010). Revised recist guideline version 1.1: What oncologists want to know and what radiologists need to know. *American Journal of Roentgenology* **195**, 281–289.

Pavlou, M. (2012). *Analysis of Clustered Data When the Cluster Size is Informative*. Ph.D. Thesis. University College London, London.

Pavlou, M. and R., S. (2013). An examination of a method for marginal inference when the cluster size is informative. *Statistica Sinica* **23**, 791–808.

Seaman, S., Pavlou, M. and Copas, A. (2014a). Review of methods for handling confounding by cluster and informative cluster size in clustered data. *Statistics in Medicine* **33**, 5371–5387.

Seaman, S. R., Pavlou, M. and Copas, A. J. (2014b). Methods for observed-cluster inference when cluster size is informative: A review and clarifications. *Biometrics* **70**, 449–456.

Williamson, J. M., Datta, S. and Satten, G. A. (2003). Marginal analyses of clustered data when cluster size is informative. *Biometrics* **59**, 36–42.

Williamson, J. M., Kim, H.-Y., Manatunga, A. and Addiss, D. G. (2008). Modeling survival data with informative cluster size. *Statistics in Medicine* **27**, 543–555.

Ying, Z. and Wei, L. (1994). The Kaplan-Meier estimate for dependent failure time observations. *Journal of Multivariate Analysis* **50**, 17–29.

Zhang, B., Liu, W., Zhang, Z., Qu, Y., Chen, Z., and Albert, P. S. (2015). Modeling of correlated data with informative cluster sizes: An evaluation of joint modeling and within-cluster resampling approaches. *SAGE Journal* **26**, 1881–1895.

Alessandra Meddis

Section of Biostatistics, University of Copenhagen, 1353 Copenhagen K, Denmark.

E-mail: alme@sund.ku.dk

Aurélien Latouche

Institut Curie, INSERM, U900, F-92210, Saint Cloud, France.

E-mail: aurelien.latouche@curie.fr

(Received September 2021; accepted May 2022)