Statistica Sinica Preprint No: SS-2021-0349	
Title	Test for Informative Cluster Size with Right Censored Survival Data
Manuscript ID	SS-2021-0349
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202021.0349
Complete List of Authors	Alessandra Meddis and Aur´elien Latouche
Corresponding Authors	Alessandra Meddis
E-mails	alme@sund.ku.dk
Notice: Accepted version subject to English editing.	

Statistica Sinica

Test for informative cluster size with right censored survival data

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

[1] *Section of Biostatistics, University of Copenhagen, Copenhagen K, Denmark*

[2] *Institut Curie, PSL Research University, INSERM, U900, F-92210, Saint Cloud*

[3] *Conservatoire National des Arts et Métiers, Paris, France*

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. The assumption of non informative cluster size is commonly used even though it may be not true in some situations. We propose a test for the assumption of Informative Cluster Size (ICS) in clustered survival data with right censoring. Standard martingale results are used to obtain the asymptotic distribution of the test statistic. Simulation studies show that the proposed test works well under various scenario. 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 by immunotherapy.

Key words and phrases: Clustered data, Hypothesis testing, Informative cluster size, Survival analysis

1. Introduction

Clustered data are often encountered in several fields of biomedical research. Observations are organized within group of various sample sizes, and while cluster 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 frailty model or marginal models, but they assume that the outcome is unrelated to the clusters sample sizes. This assumption is not always satisfied; and if not, the cluster size is said to be informative. For instance, consider the time to tooth loss in one individual with periodontal disease, subjects may already have lost some teeth due to the disease. Thus, time to loss in one individual (cluster) is linked to the number of teeth (cluster sizes) of the same individual. An other 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 clearance the worm is longer in men with multiple nests than in men with one nest. Moreover, informative cluster size might be detected in a multicentric study or meta-analysis where the size of the study can be linked to the magnitude of the treatment effect.

In the last decade, there has been substantial interest on how to handle informative cluster size (Zhang et al. (2015); Chiang and Lee (2008); Pavlou and R. (2013)). Various approaches related to this issue have been introduced: Hoffman et al. (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 et al. (2003) considered a GEE method inversely weighted by clusters sample sizes. Cong et al. (2007) investigated the WCR method for clustered survival data analyzing the resampled data sets using a Cox model. They also generalized the marginal models weighting the score function by the inverse of cluster sample size. Williamson et al. (2008) explored the estimation of the marginal distribution for multivariate survival data with informative cluster size using cluster-weighted Weibull and Cox models. For all these methods, they rely on the assumption of ICS, without testing it in the application study.

In practice, it is not generally possible to know in advance whether the informative cluster size assumption is suitable for a particular data set. While appropriate methods to handle the issue may be applied, one important point is that unnecessarily allowing for ICS leads to substantial loss of efficiency (Benhin et al. (2005)). Therefore, assessing whether the

cluster size is informative is of fundamental importance for the decision on the statistical approach analysis. Even if the nature of the the link between cluster sample size and the outcome would be an interesting point, it is not the aim of this work where we introduce a test to detect informative cluster size to avoid possible bias in the analysis.

Benhin et al. (2005) employed a Wald-type test for ignorability of cluster size in the estimating equations framework for linear and logistic regression models. Nevalainen et al. (2017) introduced a test for ICS using a balanced bootstrap to estimate the null distribution. However, for survival analysis, testing procedures are limited to ad-hoc procedures comparing the marginal distributions between strata defined by cluster size (Meddis et al. (2020)). The scope of this work is to provide a more general method to test for ICS for 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 the article is organized as follows. In Section 2, we illustrate the problem of informative cluster size and we introduce some more in depth description for possible target populations in clustered data. We further introduce a new method to test for ICS in right censored survival data and we provide the asymptotic distribution of the test statistic. In Section 3 a

simulation study is conducted to assess the power of the test. In Section 4, we illustrate the usefulness of the method in several applications, and we provide some discussion in Section 5.

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 event $\Delta_{ik} = \mathbf{I}(T_{ik} \leq C_{ik})$, where $\Delta_{ik} = 1$ if the event occurred, 0 otherwise. Let $(G_1, G_2, ..., G_K)$ be a sample of K i.i.d. observations where each G_k represents a cluster consisting of $\{N_k, (\tilde{T}_{1k}, \Delta_{1k}), ..., (\tilde{T}_{N_k k}, \Delta_{N_k k})\}$. We assume that T_{ik} and C_{ik} are independent for all i, k and that T_{ik} are independent between cluster, but within cluster k $(T_{1k}, T_{2k}, ..., T_{N_k k})$ can be correlated.

2.2 Target population

When clustered data arises, cluster sample size may provide some information about the outcome. The variability of cluster sizes might be due either to the study design to collect the data, namely to an inherent feature of the data, or to missing data. In case of missing observations we

might be interested in the effect on the outcome for the complete cluster (observed + missing observations) and assumptions on the censoring are needed for inference. We assume that there is independent censoring, and we do not discuss the problem of missing data. In this context, we can distinguish two marginal analyses that might be of interest (Hoffman et al. (2001); Seaman et al. (2014b)). One makes inference for the population of *all observed members* (AOM), which refers to a random individual in the observed population. The second one, makes inference for the population of *typical observed members of a typical cluster* (TOM), which refers to a random individual belonging to a random cluster of the observed population. Whereas in the first, larger cluster contribute more to inference, since equal weights are given to all the observed member; in the second, cluster are equally weighted. For the AOM, the analysis has an interpretation for a unit randomly sampled from the overall observed populations. The TOM analysis has a cluster-based interpretation, that is 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 et al. (2014a)). However, they differ, in general, in presence of informative cluster size.

For each cluster k, let r be the index of a randomly selected member

of the observed cluster. As in Seaman et al. (2014b), we define $e_{AOM} = \mathbf{E}[N_k T_r | N_k \geq 1]/\mathbf{E}[N_k | N_k \geq 1]$ and $e_{TOM} = \mathbf{E}[T_r | N_k \geq 1]$. The cluster size is said to be non informative (NICS) when $\mathbf{E}[T_r | N_k = n] = \mathbf{E}[T_r | N_k \geq 1]$, otherwise the cluster size is informative (Hoffman et al. (2001)). More in general, informative cluster size 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 ($e_{TOM} = e_{AOM}$). This is not true, in general, when the cluster size is informative, and 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 clusters. 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:

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

$\hat{A}_{TOM}(t)$ estimates the number of events for a typical observed member and $\hat{A}_{AOM}(t)$ estimates the number of events in the sense of 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 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 the cluster they belong to. The above estimators are consistent estimators for the cumulative

hazard functions even though data are clustered and observations are dependent in each cluster (Ying and Wei (1994)).

Let τ be the follow-up time, 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 $L(t) = \frac{Y_{AOM}(t)Y_{TOM}(t)}{K}$ is a weight function defined to ensure convergence. Under the null hypothesis $Z(\tau)/\sqrt{K}$ asymptotically tends to a Gaussian with mean 0 and covariance matrix V.

2.3.1 Proof of asymptotic distribution:

By definition:

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

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

therefore

$$Z(\tau) = \int_0^\tau L(t)\Big(\frac{dM_{TOM}(t) + \alpha_{TOM}(t)Y_{TOM}(t)}{Y_{TOM}(t)}\Big) - \Big(\frac{dM_{AOM}(t) + \alpha_{AOM}(t)Y_{AOM}(t)}{Y_{AOM}(t)}\Big)$$
$$= \int_0^\tau L(t)\Big(\frac{dM_{TOM}(t)}{Y_{TOM}(t)} - \frac{dM_{AOM}(t)}{Y_{AOM}(t)}\Big) + \int_0^\tau L(t)\big(\alpha_{TOM}(t) - \alpha_{AOM}(t)\big)dt$$

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

We specify $L(t) = \frac{Y_{AOM}(t)Y_{TOM}(t)}{K}$, and we obtain:

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

We can interchange sums and 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

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

Because of the dependence between observations we cannot refer to the usual martingale theory to prove asymptotic normality. However, we assume that observations are correlated within cluster and the N_k's are finite, thus $\{T_{ik}\}$ is a m-dependent sequence (with $m = \max_k\{N_k\}$) because

$\{T_{i1}, T_{i2}, ..., T_{iN_k}\}$ and $\{T_{i1}, T_{i2}, ..., T_{iN_k^{'}}\}$ 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 $\frac{1}{\sqrt{K}} \sum_{k=1}^{K} \int_0^\tau dM_k(t)$ converges weakly to a zero-mean Gaussian process $U^Z(t)$.

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

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

and

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

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

Hence, the statistic converges to a Gaussian with mean 0 and covariance matrix V asymptotically equivalent to $V^* = \frac{1}{K} \sum_k \sum_j \sum_{j'} \epsilon_{jk}\epsilon_{j'k}$

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 $(y_{AOM}(t) - y_{TOM}(t))$ by

$\hat{\omega}_k(t) = \left(\frac{Y_{AOM}(t)}{KN_k} - \frac{Y_{TOM}(t)}{K}\right)$:

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

2.4 Extension to regression setting

One might be interested in investigating the assumption of dependence of failure times on cluster sample size given a set of covariates. Let X_{ik} denotes 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} independent of C_{ik} given X_{ik} and possible correlation in $(T_{1k}, T_{2k}, ..., T_{N_k k})$ within each cluster k given the set of covariates.

To model the hazard conditional on covariates, a Cox model $\alpha_{ik}(t) = \alpha_0(t)\exp(\beta' X_{ik})$ is considered and we define $M_{ik}(t) = \mathcal{N}_{ik}(t) - \int_0^t \alpha_0(s)Y_{ik}(s)\exp(\beta' X_{ik})ds$. The nonparametric test, proposed above, can than be extended to the regression setting replacing the Nelson-Aalen estimator by the Breslow estimator of the cumulative baseline hazard function for the two marginal analyses:

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

with $\overline{Y}_{AOM}(t,\beta) = \frac{1}{N}\sum_{k=1}^{K}\sum_{i=1}^{N_k} Y_{ik}(t)\exp(\beta' X_{ik})$ and $\overline{Y}_{TOM}(t,\beta) = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{N_k}\sum_{i=1}^{N_k} Y_{ik}(t)\exp(\beta' X_{ik})$.

The regression coefficients β are estimated by solving the score function weighted by the inverse of cluster sample size (Cong et al. (2007)):

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

The test statistic is

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

Under the null hypothesis we obtain:

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

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 $\frac{1}{\sqrt{K}}\sum_{k=1}^{K}\int_0^{\tau} dM_k(t)$ converges weakly to a Gaussian process $U^Z(t)$, and similar argument leads to the asymptotic equivalence between

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

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 $\big(\frac{\overline{Y}_{AOM}(s,\hat{\beta})}{N_k K} - \frac{\overline{Y}_{TOM}(s,\hat{\beta})}{K}\big)$.

Therefore the statistic $Z^x(\tau)\frac{1}{\sqrt{K}}$ converges to a Gaussian with mean 0 and covariance matrix asymptotically equivalent to $\frac{1}{K}\sum_k \sum_j \sum_{j'} \epsilon_{jk}\epsilon_{j'k}$ with $\epsilon_{jk} = \int_0^{\tau} \omega_k^{\beta}(t) dM_{jk}(t)$, estimated by $\hat{V}^x = \frac{1}{K}\sum_k \sum_j \sum_{j'} \hat{\epsilon}_{jk}\hat{\epsilon}_{j'k}$ where

$$\hat{\epsilon}_{jk} = \Delta_{jk}\hat{\omega}_k^{\beta}(T_{jk}) - \sum_i \sum_l \frac{\Delta_{li}\hat{\omega}_k^{\beta}(T_{li}) Y_{jk}(T_{li}) \exp(\hat{\beta}X_{jk})}{\sum_m \sum_f Y_{fm}(T_{li}) \exp(\hat{\beta}X_{fm})}, \quad \hat{\omega}_k^{\beta}(t) = \big(\frac{\overline{Y}_{AOM}(t,\hat{\beta})}{KN_k} - \frac{\overline{Y}_{TOM}(t,\hat{\beta})}{K}\big).$$

3. Simulation study

To evaluate the power and the nominal level of the test under different scenarios, we conduct a simulation study fixing the type I error to 5%. The correlated failure times were generated from a frailty model, i.e from the conditional cumulative distribution function $P(T \leq t|U_k, X) = 1 - \exp(-U_k A_0(t)\exp(\beta X))$ 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 was 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 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 time-to-event among clusters. The variance $\sigma_V^2 = 1/\gamma$ represents the variability between clusters sample sizes. The parameter $\rho \in [0, 1]$ is the correlation between the two random effects, and it defines the dependence between T_{ik} and N_k, when $\rho = 0$ there is NICS. The strength of ICS depends on θ, ρ, γ: it decreases with larger values of θ, since the between-cluster time-to-event variability decreases. Defining the link between γ and ICS is not straightforward. We would suspect that increasing γ, the variability decreases and so does ICS. However, there is a trade-off between variability of cluster sample sizes and the magnitude of difference in time-to-event, which also depends on θ (See supplementary material).

We simulate two main settings: a) highly clustered data with $K = 100, \lambda = 5, \gamma = 20$ and b) few large cluster with $K = 25, \lambda = 20, \gamma = 3$. For both settings, λ and γ were defined by simulation (over 10000 replications) to reach an overall sample size around 1500. Right censoring is generated by uniform distribution, independent on the failure times, with the parameters as to obtain 30% and 80% censoring.

3.1 Simulation plan 1: without covariate

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

In Figure 1 we provide the empirical power of the test for increasing correlation ρ. The simulations suggests a good performance of the test reaching a power of 80% in most scenarios. The results confirms that θ is inversely proportional to ICS, showing higher power for $\theta = 5$. Moreover, we decrease the overall sample size $N = 700, 300$ either varying the number of cluster (K) or the clusters sample sizes (λ, γ). A decrease in the sample size does not seem to degrade the performance overall (Figure 2). With smaller λ a lower θ is needed to detect ICS since the clusters sample sizes are smaller and the between-clusters variability is not strong enough. However, for $K = 10$, even with decreasing θ, low power is detected, thus a sufficient number of clusters is necessary for the validity of the test. Simulations results also suggested that censoring is not impacting the performance of the test, but 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 5% for scenario A and B (Table 1) . In the supplementary

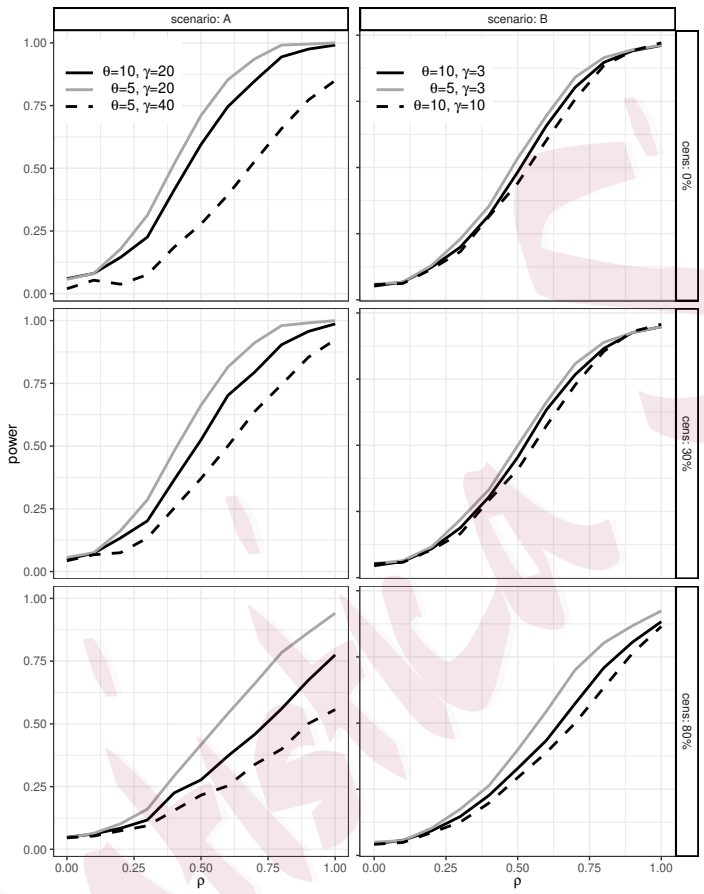

Figure 1: Power of the test for increasing correlation ρ for both scenarios considering different values of θ, γ and censoring. Each scenario is based on 1000 replications, fixing $\alpha = 0.05$. Scenario A: highly clustered data ($K = 100$, $\lambda = 5$), scenario B: few big clusters ($K = 25, \lambda = 20$).

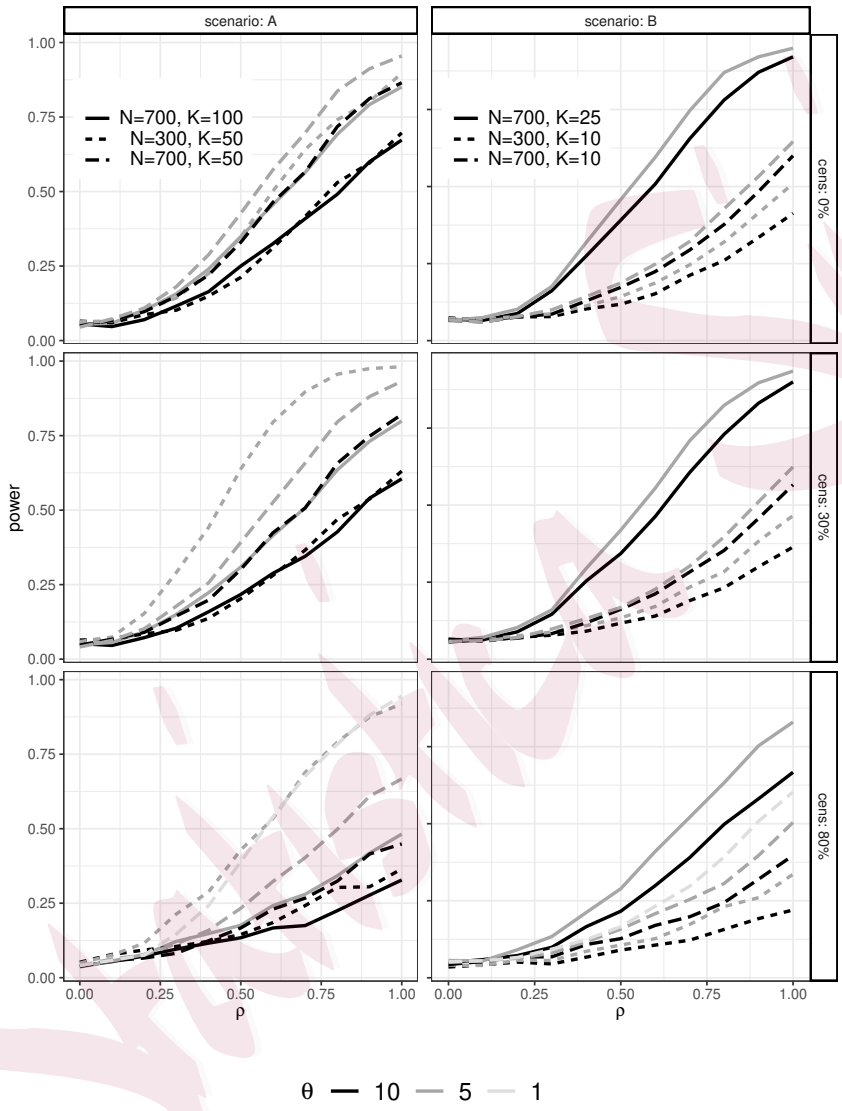

Figure 2: Power of the test for increasing correlation ρ for both scenarios for smaller sample size at varying of K, λ and censoring. Each scenario is based on 1000 replications, fixing $\alpha = 0.05$.

material the cluster sample size distribution is provided for the simulated settings at varying of ρ.

3.2 Simulation plan 2: regression setting

We fix $\theta = 5$ and we assess the performance of the test generating a continuous covariate with normal distribution $N(0, 1)$. The covariate is generated independently on the cluster sample size, with the result that 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, 30% and 80% of right censoring. We decrease the sample size to $N = 700, 300$ as in Simulation 1. Similar results are obtained, with a good performance of the test 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 to support the theoretical findings and simulations.

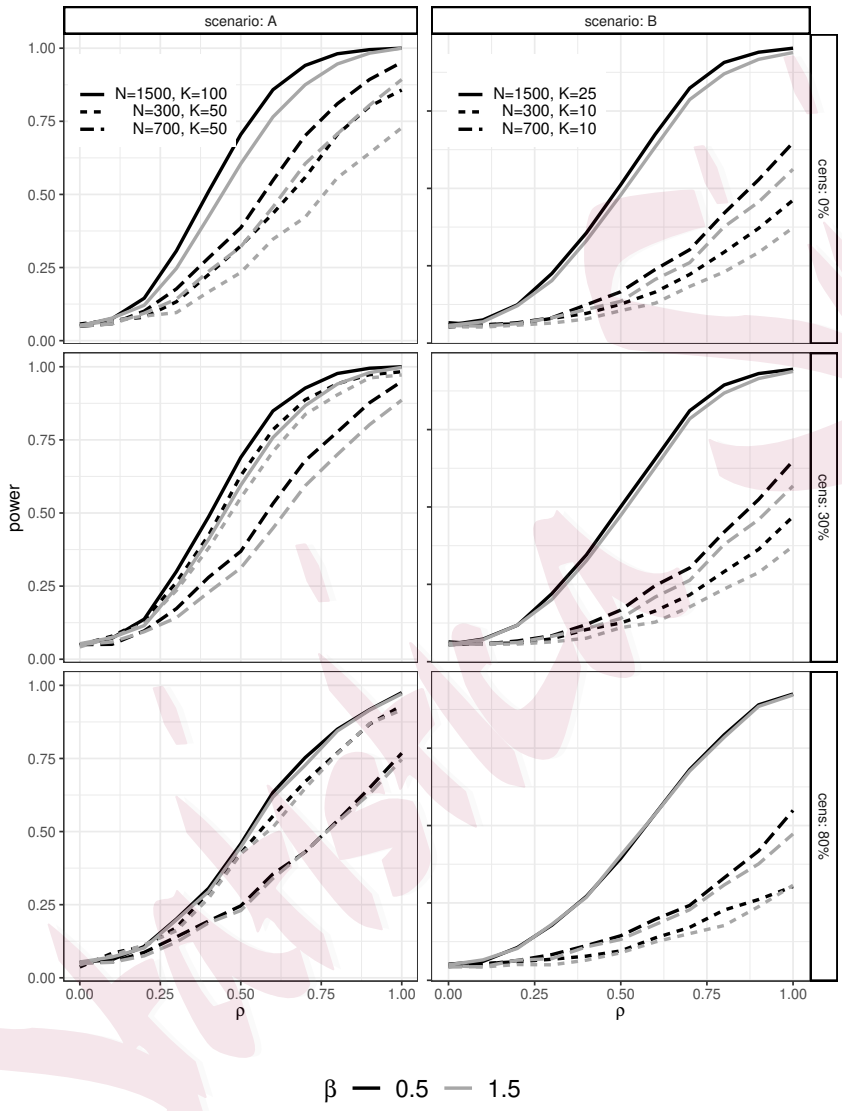

Figure 3: Power of the test for increasing correlation ρ with $X \sim N(0,1)$ for both scenarios considering different values of N, K, β and censoring. Each scenario is based on 1000 replications, fixing $\alpha = 0.05$.

4.1 Dental data

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 aimed to construct multivariate survival trees to predict tooth loss. A total of 5336 patients with periodontal disease were collected with 65228 teeth. We excluded from the analysis individuals with only one tooth resulting in a sample size of 65034 teeth. The average age was 58 years, with 51% women, 9% had Diabetes Mellitus, and 23% were smokers. The number of teeth that fell is 4334 with a median tooth loss time of 0.556 $[0.003, 5.594]$ years. Several teeth and individual characteristics are also provided in the data set but we do not take them into consideration.

We suspect ICS 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 in one individual with smaller cluster size. The test shows clear evidence of ICS with a test statistic of 8.932 (pvalue=0). We provide the plot of the Kaplan-Meier estimator of the survival function at each cluster sample size at the median time (Figure 4). This suggests as well ICS: the tooth loss time is longer in individuals with more teeth (e.g., bigger cluster sample size). For instance, the probability for a tooth to not fall before the

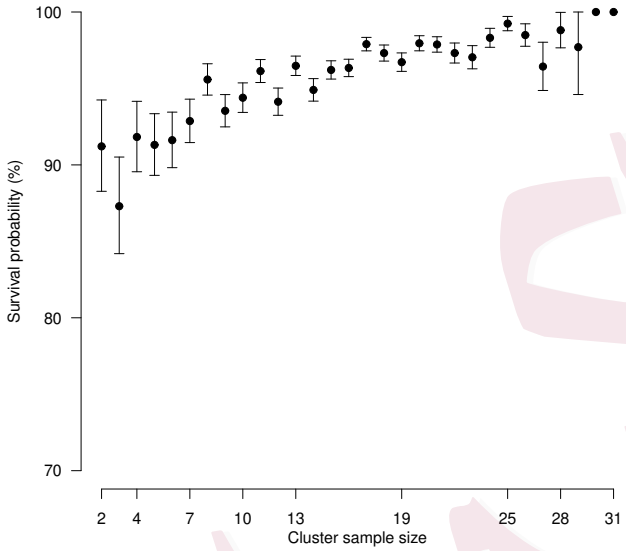

Figure 4: Estimated survival functions at the median failure time $t = 0.556$ years for increasing cluster sample size. The confidence intervals for each probability are provided.

median time in individuals with 13 teeth is higher compared to the probability for a tooth to not fall in individuals with 7 teeth. The assumption of ICS seems to be reasonable for this data, thus for the consequent analysis it would be appropriate to employ the WCR method (Cong et al. (2007)) or the multivariate survival model proposed by Williamson et al. (2008).

4.2 Multicentric data

We consider a multicentric study of patients with liver disease primary biliary cirrhosis (PBC). It is 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 either treatment with Cyclosporin A (176 patients) or placebo (173 patients). The effect of treatment on the composite outcome "failure of medical treatment" was of interest. It is defined as either death or liver transplantation. Data are characterised by 75% of censoring where 90 patients had 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, and the null hypothesis of NICS is rejected with a test statistic equal to -1.98 (pvalue=0.04). The K-M at the median time at varying of the cluster sample size are also provided in Figure 5. Because of high censoring (75%), the weighted marginal survival model would be preferred to the WCR methods for the analysis (Cong et al. (2007)).

4.3 Cancer data: Immunotherapy

We consider a data set of 100 patients with metastatic cancer treated by immunotherapy at the Institut Curie Comprehensive Cancer Center in Paris.

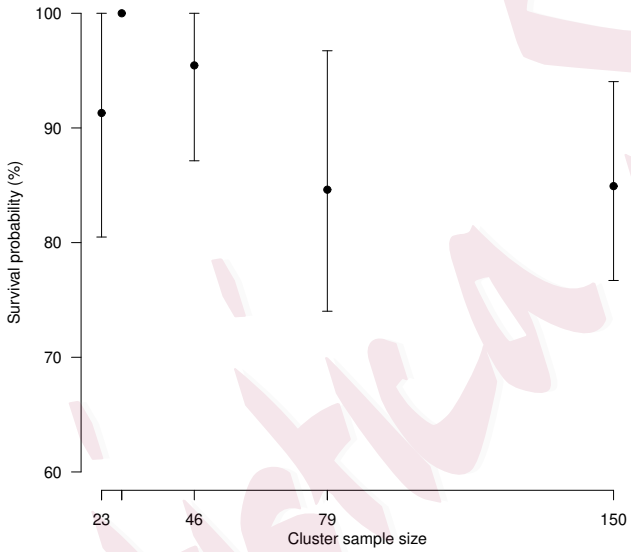

Figure 5: Estimated survival function at median failure time $t = 21$ months for increasing cluster sample size in the subgroup of patients that received the treatment. The confidence intervals for each probability are provided.

Table 1: Nominal level of the test without covariates. Results are provided for 1000 replications fixing $\alpha = 0.05$.

					$\hat{\alpha}$		
N	K	λ	γ	θ	cens 0%	cens 30%	cens 80%
Scenario A							
1500	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							
1500	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 1000 replications fixing $\alpha = 0.05$.

						$\hat{\alpha}$		
β	N	K	λ	γ	θ	cens 0%	cens 30%	cens 80%
0.5	1500	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	1500	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**

For each patient, the size of each metastasis is radiologically evaluated from the treatment initiation to the date of progression of the specific metastasis. Immunotherapy may have different effect 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 2 to 4 metastases. For each subjects, a maximum of five of target metastases are considered as for the RECIST guideline (Nishino et al. (2010)). The primary cancer was of different nature: breast cancer, head neck cancer, lung cancer, urological cancer and others. The principal objective of the study was to have some insight on dissociate response that are typical of immunotherapy, notably in the same individual, the response to treatment might be of different nature 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 metastasis should affect the outcome. However, this idea was not confirmed by the test that did not reject the null hypothesis of NICS with a test statistic of -0.85 (pvalue=0.39). However, the number of metastases seems to have an impact on the survival function for metastasis disease progression

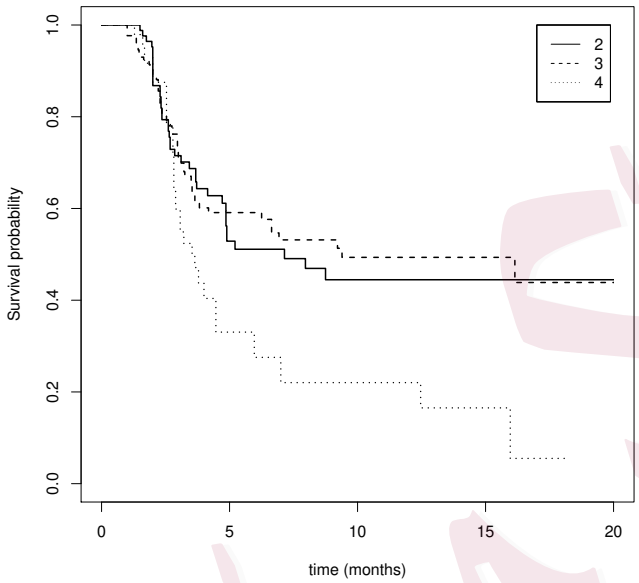

Figure 6: Estimated survival function for different number of metastases (cluster sample sizes).

(Figure 6). We computed the log-rank test for the three groups with different cluster sample sizes, which showed 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 presence of clustered data, standard statistical methods implicitly assume that the size of the clusters is unrelated to the outcome of interest. This assumption is not always verified and 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. It can be used as a pre-test to determine whether to use standard regression model for clustered survival data, valid under the NICS assumption, or methods that account for the information carried by the cluster sample size. The test statistic relies on the fact that under NICS the two 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 a result of missing data, notably when clusters would be of the same size but some members are not observed (missing observations). Hoffman et al. (2001) and Williamson et al. (2003) stated that missing completely at random (MCAR) mechanism is equivalent to non-informative cluster size. Pavlou (2012) associated NICS to missing data mechanism, of which MCAR is special case, and they proved the equality of results for the target populations in three cases (TOM, AOM, missing data). In this work, we assume that the observed clusters are complete, not considering the problem

of missing data, and the issue of informative censoring is not discussed. This is a challenging point that would require methods able to handle a possible dependence between censoring and cluster sample size.

Several applications where ICS is detected are provided. Moreover, the publication bias which characterizes some meta-analyses can also be considered as a problem of informative cluster size as the treatment effect is often linked by the study sample size. The funnel plot is often showed to investigate the presence of publication or others forms of bias in meta-analysis. It provides information on the treatment estimate against a measure of the study sample size. It is a way of examining small study effects, the tendency for the smaller studies in a meta-analysis to show larger treatment effects, that can also be interpreted as a problem of informative cluster size.

An extension of the test to the regression setting is also proposed. 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 the Breslow estimator is employed instead of 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.

Simulation results suggest a good performance of the method overall for

both scenarios with low power of the test when the number of cluster is lower than 10 and for the highly clustered data scenario when cluster sample sizes are small ($k = 100, \lambda = 2$ in the Simulation). The proposed test detects if there is dependence between cluster sample size and outcome. We do not focus on the nature of the association nor on the several possibilities of distribution in the generating method.

The test relies on the definition of the cumulative hazard estimator, thus extension of the method to survival analysis issues such as interval censored data depends on appropriate modification of the Nelson-Aalen estimator. Moreover, in both simulation and application, we refer to unit-level covariates. In case of cluster-level covariates, the TOM and AOM definitions are still suitable.

In both simulation and application section we refer to unit-level covariates. In case of cluster-level covariate, the AOM and TOM definitions are still suitable. In the simulation study the covariate is generated independently on the cluster sample size (X is size-unbalanced). Other cases are possible where: i) the cluster sample size affects the covariate distribution but not the outcome, thus the dependence on the cluster sample size is through X; ii) covariate effect is varying with cluster sample size, that corresponds to an interaction between cluster sample size and covariate in

the survival model. We did not consider these scenarios because they are related to the issue of informative covariate structure and confounding by cluster as discussed in Pavlou (2012). Informative cluster size is mainly defined by the relationship between cluster sample size and outcome. We introduce the extension to the regression setting to show how the method would be implemented when the analysis requires adjustment for some covariates. Informative covariate structure is a related problem that can also occur without informative cluster size. This could be an interesting point for future discussion.

A test for ICS has been already introduced for clustered data for linear regression models by a balanced bootstrap method, since the distribution of the statistic under the null is analytically intractable (Nevalainen et al. (2017)). An adaptation of this method to survival data could be also employed to investigate for ICS, but it is characterized by high computational cost. Introducing the cluster sample size as covariate in the regression model might be an other option to test for ICS. However, unlike for the proposed test, a specific link between cluster sample size and outcome is assumed. Furthermore, adding the cluster sample size in the model would test for ICS but the estimated effect is conditional on the sample size. Thus, a two-step procedure with appropriates methods to handle ICS are needed to obtain

results on the marginal effect.

Supplementary Materials

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

Acknowledgements

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 by immunotherapy.

References

Benhin, E., J. Rao, and A. Scott (2005). Mean estimating equation approach to analysing cluster-correlated data with nonignorable cluster sizes. *Biometrika 92*(2), 435–450.

Calhoun, P., X. Su, M. Nunn, and J. Fan (2018). Constructing multivariate survival trees: the mst package for r. *Journal of Statistical Software 83*(12).

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

REFERENCES

Cong, X. J., G. Yin, and Y. Shen (2007). Marginal analysis of correlated failure time data with informative cluster sizes. *Biometrics 63*(3), 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., P. K. Sen, and C. R. Weinberg (2001). Within-cluster resampling. *Biometrika 88*(4), 1121–1134.

Lee, E. W., L. Wei, and Z. Ying (1993). Linear regression analysis for highly stratified failure time data. *Journal of the American Statistical Association 88*(422), 557–565.

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

Nevalainen, J., H. Oja, and S. Datta (2017). Tests for informative cluster size using a novel balanced bootstrap scheme. *Statistics in medicine 36*(16), 2630–2640.

Nishino, M., J. P. Jagannathan, N. H. Ramaiya, and A. D. Van den Abbeele (2010). Revised recist guideline version 1.1: what oncologists want to know and what radiologists need to know. *American Journal of Roentgenology 195*(2), 281–289.

Pavlou, M. (2012). *Analysis of clustered data when the cluster size is informative*. Ph. D. thesis, UCL (University College London).

Pavlou, M. and S. R. (2013). An examination of a method for marginal inference when the

cluster size is informative. *Statistica Sinica 23*, 791–808.

Seaman, S., M. Pavlou, and A. Copas (2014a). Review of methods for handling confounding by cluster and informative cluster size in clustered data. *Statistics in medicine 33*(30), 5371–5387.

Seaman, S. R., M. Pavlou, and A. J. Copas (2014b). Methods for observed-cluster inference when cluster size is informative: A review and clarifications. *Biometrics 70*(2), 449–456.

Williamson, J. M., S. Datta, and G. A. Satten (2003). Marginal analyses of clustered data when cluster size is informative. *Biometrics 59*(1), 36–42.

Williamson, J. M., H.-Y. Kim, A. Manatunga, and D. G. Addiss (2008). Modeling survival data with informative cluster size. *Statistics in medicine 27*(4), 543–555.

Ying, Z. and L. Wei (1994). The kaplan-meier estimate for dependent failure time observations. *Journal of Multivariate Analysis 50*(1), 17–29.

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

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

E-mail: (alme@sund.ku.dk)

Institut Curie, PSL Research University, INSERM, U900, F-92210, Saint Cloud

E-mail: (aurelien.latouche@curie.fr)