# Test for informative cluster size with survival data

Alessandra Meddis[1], A. Latouche[1,2]

1.Institut Curie, U900, F-92210, Saint Cloud
2.Conservatoire National des Arts et Métiers,Paris

GDR Statistique et Santé

October 11th

# Outline

- Context and motivation
- Notations and definitions
- Test statistic and its distribution
- Perspectives

# Motivation

- Clustered survival data :
  - observations contributed by the same cluster (eg individual, center) tend to be dependent, while those from different clusters are independent.

- General methodologies consider the cluster size to be a fixed design. However, in some scenarios the cluster size can be informative for inference
  $\rightarrow$ **Informative Cluster Size (ICS)**

# Motivating example

- French patients with hepatocellular carcinoma[1]:
  - 538 patients
    - ⋆ cirrhosis
    - ⋆ hepatitis B/C
  - 90 different institutions
    - ⋆ different sample sizes (5-55)
    - ⋆ patients in bigger hospitals have better prognosis
  - aim of the study: compare three scores for predicting survival
- Our goal is to investigate on **I**nformative **C**luster **S**ize (ICS): when the outcome depends on the cluster size conditionally on a set of covariates.

[1] S.Collette & all. *Prognosis of advanced hepatocelluar carcinoma:comparison of three staging system in two French clinical trials*. Annals of Oncology (2008)

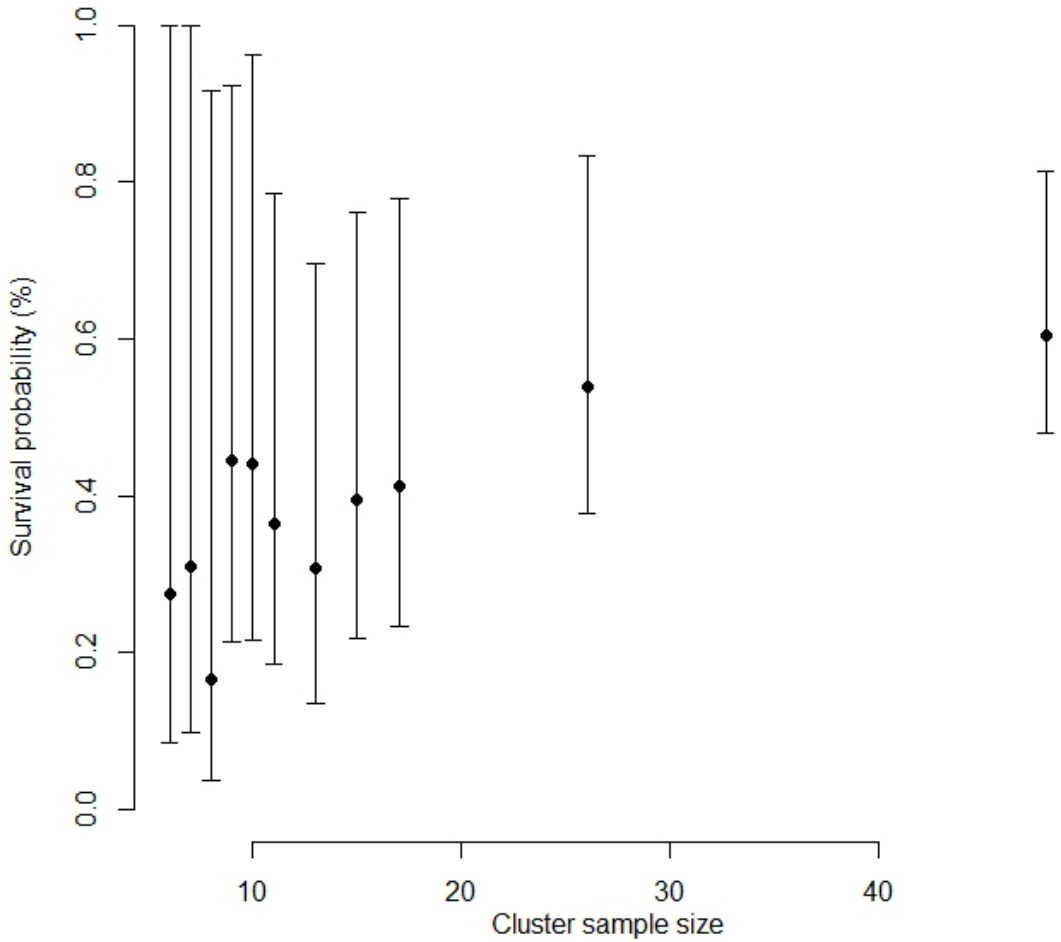
# Example data with ICS

- We can provide some typical studies where the cluster size can be informative:
  - Dental data: the probability for a teeth to fall in one individual (cluster) is linked to the number of tooth (cluster sizes) of the same.
  - Metastatic cancer data: several metastasis sites are explored . Sites from same individual are correlated and the number of metastatic site has an impact on the response to treatment.
  - Meta-analysis: pooling data from different trials with different sample sizes.

♣ For example 1 and 2 we would expect ICS because of the structure of the data, while for example 3 we would assume non informative cluster size.

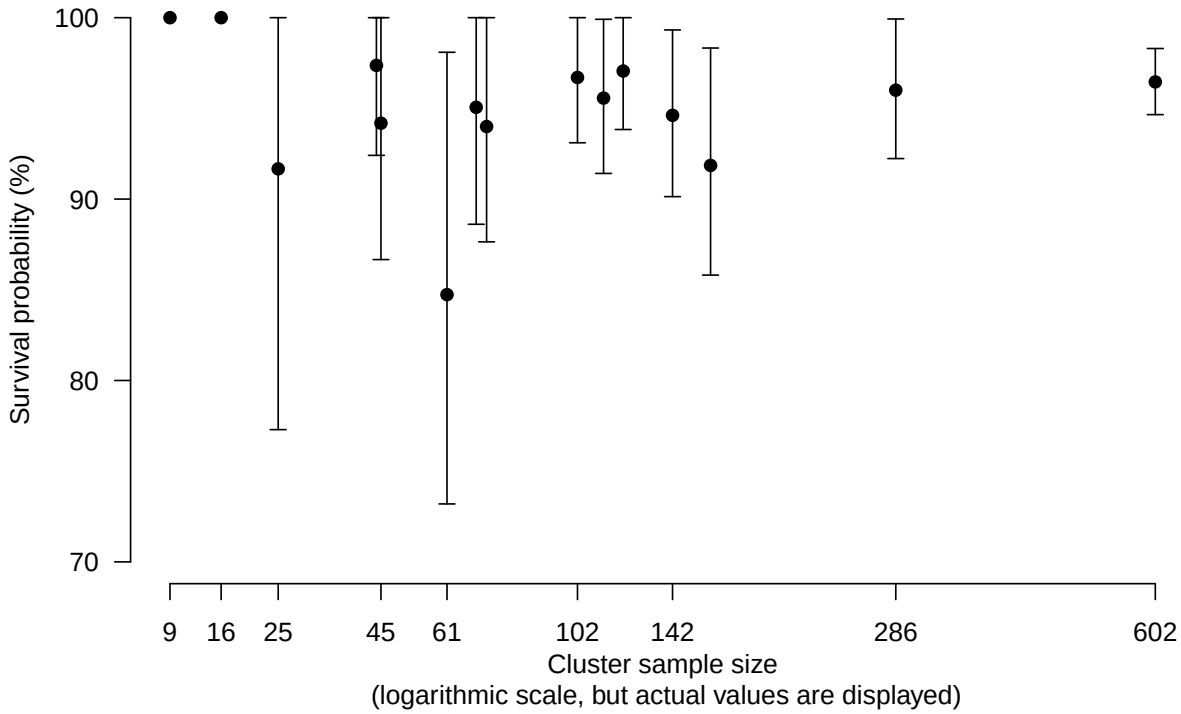
# Motivating example: ad-hoc analysis for ICS

- Kaplan-Meier estimator of the survival function at $t^* = 6$ months for each cluster in order to study the relationship between the cluster sample sizes and the outcome.



# Ad-hoc analysis with NICS

- Example where non informative cluster size is suggested:
  - IMENEO[2] meta-analysis for non metastatic breast cancer
  - 16 centers
  - correlation between failure times was detected




[2]Bidard F, Michiels S, Riethdorf S, et al. Circulating tumor cells in breast cancer patients treated by neoadjuvant chemotherapy: a meta-analysis *JNCI: Journal of the National Cancer Institute* 2018; 110(6): 560–567.6:

# Formalism

- $(V_1, V_2, ..., V_K)$ sample i.i.d observations where $V_i$ represents a cluster consisting of

$$(n_i, (\tilde{T}_{i1}, \Delta_{1i}, X_{i1}), ..., (\tilde{T}_{in_i}, \Delta_{in_i}, X_{in_i}))$$

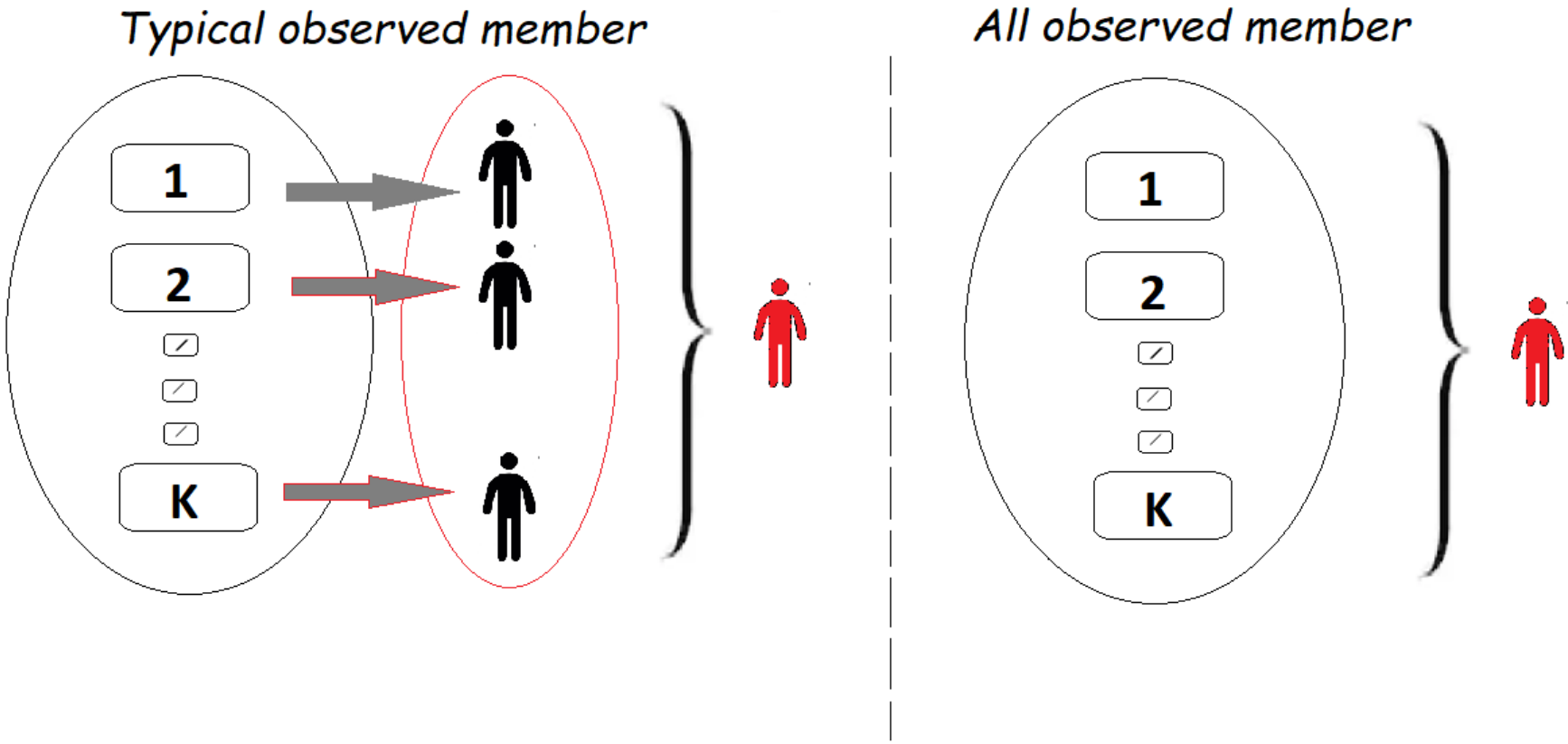
  - $n_i$: cluster sample size
  - $\tilde{T}_{ij} = \min(T_{ij}, C_{ij})$: the observed failure time
  - $\Delta_{ij} = I(T_{ij} \leq C_{ij})$ : the censoring indicator
  - $X_{ij}$ set of covariates with $i = 1..K$ and $j = 1, ..., n_i$

- we assume clustered data: in each cluster k $(T_{i1}, T_{i2}, ..., T_{in_i})$ can be correlated conditionally on $(X_{i1}, X_{i2}, ..., X_{in_i})$

# Two different marginal analyses

- When cluster data arises two marginal analyses are of interest:
  - for the population of **all observed members** (AOM)
    - ⋆ we refer to a typical individual randomly sampled by the entire population
    - ⋆ equal weight to each individual and larger clusters contribute more to inference
  - for the **typical member of a typical cluster** (TOM)
    - ⋆ we refer to a randomly selected individual from a randomly selected cluster
    - ⋆ same weight to individuals within same cluster and each cluster contribute equally to inference.

# Two marginal analyses: illustration



# (Non) Informative cluster size

- Let $r_k$ be the index of a randomly selected member of cluster $k$. Hoffman et al. [2001] define non informative cluster size (NICS)

$$\mathbf{P}(D_{r_k}(t) = 1 | X_{r_k} = x, N_k) = \mathbf{P}(D_{r_k}(t) = 1 | X_{r_k} = x)$$

otherwise the cluster size is said to be informative (ICS)

  - Given large enough sample sizes, the two marginal analyses coincides under NICS [3]
  - under ICS they differ in general $\rightarrow$ it is important to precise which quantities we are interested to.

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

# Consequences of ICS

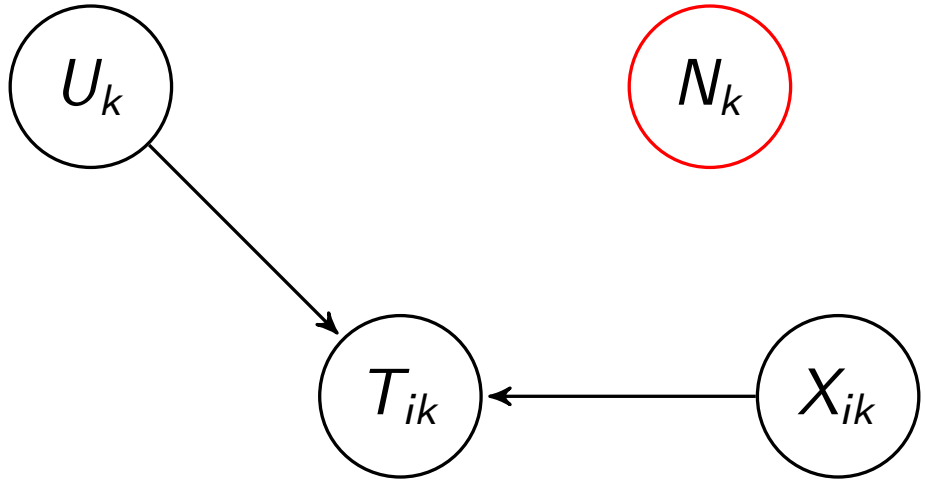
When informative cluster size is detected, more care is needed in the interpretation of results:

- the estimated quantities depend on the distribution of $N_k$ (study design to collect the data) which is specific to the population in analysis.
- it is challenging to generalize the results to other populations

  $\rightarrow$ appropriate methods to take into account the information carried by the cluster sample size are necessary.
  Several approaches have been proposed, motivated by data that rely on the assumption of ICS, but no formal test was performed.

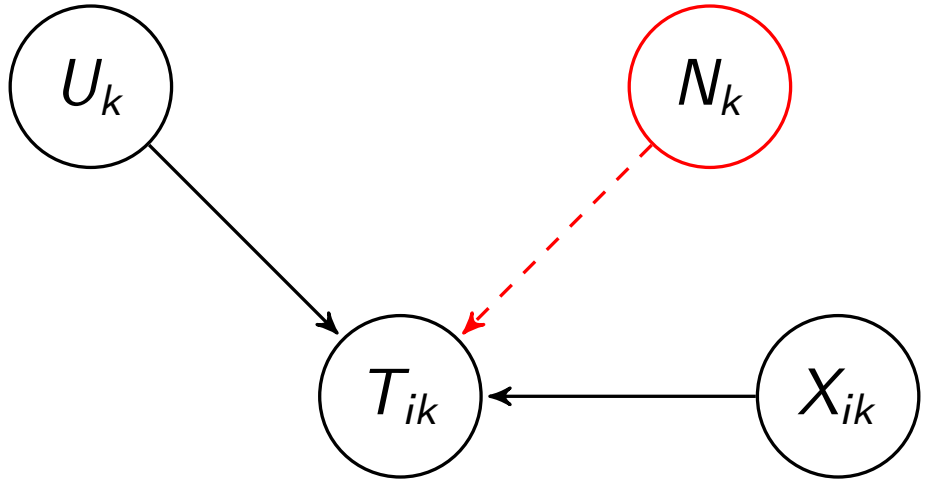
♣ **We propose a test for informative cluster size with survival data**.

# Illustration: Non informative cluster size




- $U_k$ is the random effect for the unmeasured covariates which are common to all members of the same cluster k (correlated failure times)
- $N_k$ does not affect $T_{ik}$ $\rightarrow$ **non informative cluster size**

# Illustration: Informative cluster size




- $U_k$ is the random effect for the unmeasured covariates which are common to all members of the same cluster k (correlated failure times)
- $N_k$ affects $T_{ik}$ $\rightarrow$ **informative cluster size**

# Notations

- Let $i = 1, 2, ..., K$ index the cluster and $j = 1, 2, ..., n_i$ the individuals within cluster $i$ with $N = \sum_i n_i$. We define:
  - $N_{ij}(t) = I(\tilde{T}_{ij} \leq t, \Delta_{ij} = 1)$ : the counting process
  - $\alpha_{ij}(t) Y_{ij}(t)$ : the intensity
  - $Y_{ij}(t) = I(\tilde{T}_{ij} \geq t)$ : the at-risk process

- $M_{ij}(t) = N_{ij}(t) - \Lambda_{ij}(t)$ is a martingale with respect to the filtration $\mathcal{F}_{ij}(t) = \sigma\{N_{ij}(u), Y_{ij}(u) : 0 \leq u \leq t\}$.

# Nelson-Aalen estimator

We define the Nelson-Aalen estimator of the cumulative risk for the two marginal analyses:

$$\hat{\Lambda}_{tom}(t) = \int_0^t \frac{dN_{tom}(s)}{Y_{tom}(s)} ds \quad with \quad N_{tom}(t) = \frac{1}{K} \sum_i \frac{1}{n_i} \sum_j N_{ij}(t)$$

$$\hat{\Lambda}_{aom}(t) = \int_0^t \frac{dN_{aom}(s)}{Y_{aom}(s)} ds \quad with \quad N_{aom}(t) = \frac{1}{N} \sum_i \sum_j N_{ij}(t)$$

# Test statistic

- Test for Informative Cluster Size:
  - $H_0$ : equality of the intensity of the process $N_{ij}(t)$ obtained by the two analysis (tom/aom) at each time t:

    $$H_0 : \frac{1}{K}\sum_i \frac{1}{n_i}\sum_j \frac{\alpha_{ij}(t)Y_{ij}(t)}{Y_{tom}} = \frac{1}{N}\sum_i\sum_j \frac{\alpha_{ij}(t)Y_{ij}(t)}{Y_{aom}} = \frac{\alpha_k(t)}{Y_k(t)} \quad \forall t$$

  - test statistic:

    $$Z(\tau) = \int_0^\tau L(t)(d\hat{\Lambda}_{tom} - d\hat{\Lambda}_{aom})$$

    $L(\cdot)$ is a weight function

# Under NICS

Under the null hypothesis :

- we define $L(t) = \frac{Y_{aom}(t)Y_{tom}(t)}{K}$
- with some algebra we can rewrite

$$Z(\tau)\frac{1}{\sqrt{K}} = \frac{1}{\sqrt{K}}\sum_{i=1}^{K}\int_0^\tau W_i(t)dM_i(t)$$

$$W_i(t) = \frac{Y_{aom}(t)}{n_i K} - \frac{Y_{tom}(t)}{K}$$

- $\frac{1}{\sqrt{K}}\sum_i\sum_j\int_0^\tau dM_{ij}$ converges to a Gaussian process [4]

[4] Z.Ying and L.J.Wei. The Kaplan-Meier estimate for dependent failure time observations. *Journal of Multivariate Analysis* vol.50 pp 17-29,1994

# Asymptotic distribution

Assume that exists $y_{aom}(t), y_{tom}(t)$ such that for $N \to \infty$

$$Y_{aom}/n_i K \to y_{aom}(t)$$
$$Y_{tom}/K \to y_{tom}(t)$$

$\Rightarrow Z(\tau)\frac{1}{\sqrt{K}}$ is asymptotically equivalent to a Gaussian with mean 0 and covariance: $V = \frac{1}{N}\sum_i \sum_j \sum_{j'} \epsilon_{ij}\epsilon_{ij'}$

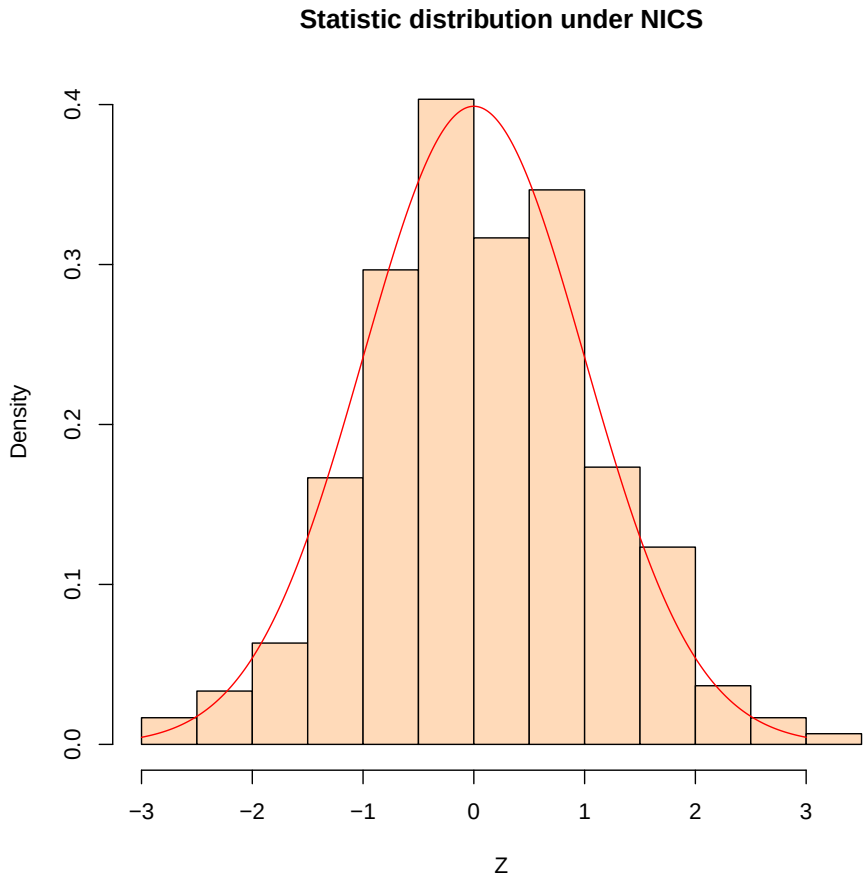
with $\epsilon_{ij} = \int_0^\tau \omega_i(t) dM_{ij}(t)$ estimated by

$\hat{\epsilon}_{ij} = \Delta_{ij}\omega_i(T_{ij}) - \sum_k \sum_l \frac{\Delta_{kl}\omega_i(T_{kl})Y_{ij}(T_{kl})}{\sum_m \sum_f Y_{mf}(T_{kl})}$

# Simulation design

We conduct a simulation to check for the asymptotic distribution of the test statistic

- Correlated survival data with NICS:
  - shared frailty model
  - frailty $U_k \sim Gamma(1.4)$ $\rightarrow var(U_k) = 0.7$
  - no covariates
- K=40 clusters with sample sizes $N_k \in [20, 70]$
- M=1000 replications



# On going work

- Simulation study
  - assess the power of the test at different number of clusters and cluster sample sizes
  - introduce covariates
- Apply the test of ICS in the example on hepatocellular carcinoma.

# References I


Hoffman, E. B., Sen, P. K., and Weinberg, C. R. (2001).

Within-cluster resampling.

*Biometrika*, 88(4):1121–1134.

Seaman, S. R., Pavlou, M., and Copas, A. J. (2014).

Methods for observed-cluster inference when cluster size is informative: A review and clarifications.

*Biometrics*, 70(2):449–456.

Williamson, J. M., Kim, H.-Y., Manatunga, A., and Addiss, D. G. (2008).

Modeling survival data with informative cluster size.

*Statistics in medicine*, 27(4):543–555.

Thank you for your attention

# Two marginal analyses: Illustration 2