A log-rank type test to compare net survival distributions

Nathalie Graffé o, Fabienne Castell, Aurélien Belot, Roch Giorgi

Biometrics 000, 1–20 DOI: 000
000 0000

A log-rank type test to compare net survival distributions

Nathalie Grafféo[1,2], Fabienne Castell[3], Aurélien Belot[4,5,6], and Roch Giorgi[1,2,7,*]

[1]INSERM, UMR912 " Economics and Social Sciences Applied to Health & Analysis of Medical Information " (SESSTIM), 13006 Marseille, France

[2]Aix Marseille University, UMR_S912, IRD, 13006, Marseille, France

[3]Aix Marseille University, CNRS, Centrale Marseille, I2M, UMR 7373, 13453 Marseille, France

[4]Hospices Civils de Lyon, Service de Biostatistique, Lyon, France

[5]University Lyon 1, UMR 5558 Laboratoire Biostatistique-Santé, , Villeurbanne, France

[6]Institut de Veille Sanitaire, DMCT, Saint-Maurice, France

[7]APHM, Hôpital Timone, BIOSTIC, Marseille, France

email: roch.giorgi@univ-amu.fr

Summary: In population-based cancer studies, it is often of interest to compare cancer survival between different populations. However, in such studies the exact causes of death are often unavailable or unreliable. Net survival methods were developed to overcome this difficulty. Net survival is the survival that would be observed, in a hypothetical world, if the studied disease were the only possible cause of death. The Pohar-Perme estimator (PPE) is a non-parametric consistent estimator of net survival. In this paper, we present a log-rank-type test for comparing net survival functions estimated by this estimator between several groups. We expressed our test in the counting process framework to introduce the inverse probability weighting procedure as done in the PPE. We built a stratified version to control for categorical covariates affecting the outcome. Simulation studies were performed to evaluate the performance of our test and an application on real data is provided.

Key words: Cancer; Log-rank; Net survival; Pohar-Perme estimator; Stochastic process; Test.

This paper has been submitted for consideration for publication in *Biometrics*

1. Introduction

Net survival, the survival associated to the excess mortality hazard, is the survival observed in an hypothetical world where the disease of interest would be the only possible cause of death. The observed survival, which is the most frequently used, is the result of two main survival components. One part comes from the studied disease whereas the second part comes from all other causes that we are all exposed (Estève et al., 1990; Perme, Stare, and Estève, 2012). On one hand, the observed survival do not distinguish between death from the disease of interest (or excess death) and death from other causes. On the other hand, net survival evaluates the burden of this disease independently of the differences in general population mortality given by life tables, that is to say the mortality due to other causes. In cancer research, the idea of net cancer survival is to study the proportion of cancer deaths, that is to say patients dying, directly or indirectly, from cancer. So, this epidemiological indicator, routinely estimated in cancer registries and in population-based studies (see e.g. the EUROCARE program (De Angelis et al., 2014), the US SEER program (Howlader et al., 2011) or the CONCORD programme (Allemani et al., 2015)), is crucial for comparison between different populations (Perme et al., 2012; Danieli et al., 2012). For instance, when comparing patterns of care between countries, it is essential to take into account the general population mortality because of its weight on observed survival.

In population-based studies the exact causes of death are often unavailable and, when available, it is often difficult to state whether they are disease related (Berkson and Gage, 1950). Net survival methods were developed to overcome this difficulty (Estève et al., 1990). Historically, several non-parametric estimators have been proposed to estimate net survival (Ederer and Heise, 1959; Ederer, Axtell, and Cutler, 1961; Hakulinen, 1982). But in 2012 Perme et al. (2012) argued that, in most cases, these estimators do not estimate net survival. They proposed a non-parametric estimator that corrects the Ederer II estimator (Ederer and

Heise, 1959) which is biased due to informative censoring. For instance, excess mortality and other causes mortality share the influence of age leading to informative censoring. So Perme et al. used population mortality information to weight and correct for those who left the sample due to deaths of other causes. In addition, Danieli et al. (2012) showed by a simulation study that the Pohar-Perme estimator (PPE) is a consistent non-parametric estimator of net survival, which may be preferred to the other existing non-parametric estimators. The PPE assesses a hypothetical quantity which allows comparison across populations. However, to the best of our knowledge, it is not yet possible to compare distributions of net survival over a given period. We can only compare two estimates at a given time t with a classical Z-test. Besides in the parametric framework, we could use a likelihood ratio test from the multivariate excess mortality model (see e.g. Remontet et al., 2007) to compare net survival between 2 or more groups, but it requires some complex model building strategy.

In this paper, we propose a log-rank type test to compare distributions of net survival estimated by the PPE between at least 2 groups over a defined follow-up period. This choice was made for several reasons. First, the log-rank test (Mantel, 1966; Peto and Peto, 1972) is the most commonly used test to compare distributions of observed survival between at least two groups. Secondly, the log-rank test uses the cumulative hazard function and can be represented with stochastic processes (Aalen, Borgan, and Gjessing, 2008; Fleming and Harrington, 2011; Andersen, Borgan, Gill, and Keiding, 1993). Finally, because the PPE is developed on that scale and is written with stochastic processes, the log-rank test allows to introduce easily the weights of the PPE in the corresponding counting processes.

In section 2 we present the building of our proposed log-rank type test and the stratified version of this test in Section 3. Section 4 presents a simulation study where we investigated the performance of our test and Section 5 provides an application to a colorectal cancer data set. We conclude this paper with a brief discussion.

2. A log-rank type test for $k \geqslant 2$ groups

The proposed test compares the distribution of net survival estimated by the PPE (Perme et al., 2012) between $k \geqslant 2$ groups over a defined follow-up period. Assume that observations are made on n_h patients from group h with $h \in [\![1; k]\!]$ and $k \geqslant 2$. Let $n = \sum_{h=1}^{k} n_h$ denote the total number of patients. Let's also assume (Fleming, Harrington, and O'sullivan, 1987)

$$\forall h \in [\![1; k]\!], \lim_{n\to\infty} \frac{n_h}{n} = \alpha_h; \alpha_h \in]0; 1[\,.$$

Note that under these assumptions: $\lim_{n\to\infty} \min_h n_h = \infty$.

2.1 *Notations and model*

For each patient i in the group h, we consider that the time to death, $T_{h,i}$, is the minimum of two distinct times: $T_{P_{h,i}}$ due to "population hazard" and $T_{E_{h,i}}$ due to "excess hazard". Let $C_{h,i}$ denote the time to censoring and define $U_{h,i} = \min(T_{h,i}, C_{h,i})$ the follow-up time of patient i. $\tilde{\delta}_{h,i}$ denotes the failure indicator equal to 1 if the true failure time, $T_{h,i}$, is observed and 0 if patient i is censored. Each patient i in a group h has covariates denoted by the vector $\boldsymbol{X_{h,i}}$. $\boldsymbol{D_{h,i}}$ is a sub-vector of $\boldsymbol{X_{h,i}}$ describing all the demographic covariates so that $\boldsymbol{X_{h,i}} \setminus \boldsymbol{D_{h,i}}$ and $T_{P_{h,i}}$ are independent. We take the same set of assumptions as in Perme et al. (2012) that is :

(1) $(T_{P_{h,i}}, T_{E_{h,i}}, C_{h,i}, \boldsymbol{X_{h,i}})_{h,i}$ are mutually independent;

(2) $(T_{P_{h,i}}, T_{E_{h,i}}, C_{h,i}, \boldsymbol{X_{h,i}})_i$ have the same distribution;

(3) $T_{E_{h,i}}$ and $T_{P_{h,i}}$ are conditionally independent given $\boldsymbol{X_{h,i}}$;

(4) censoring times $C_{h,i}$ are independent of the pair $(T_{h,i}, \boldsymbol{X_{h,i}})$.

Further, we assume that the censoring process is non informative i.e. $S_{C,h}(t) := P(C_{h,i} > t)$ $(\forall i \in [\![1; n]\!], \forall h \in [\![1; k]\!])$. The observed data are given by $(U_{h,i}, \tilde{\delta}_{h,i}, \boldsymbol{X_{h,i}})_{h,i}$ for each patient i in group h. The conditional net survival function of $T_{E_{h,i}}$ corresponding to every patient i belonging to group h is denoted by $\tilde{S}_{E,h,i}(t) = P(T_{E_{h,i}} > t \mid \boldsymbol{X_{h,i}})$. The corresponding

conditional cumulative excess hazard is denoted by $\tilde{\Lambda}_{E,h,i}$. In the same way, we can define the conditional population all-cause survival as $\tilde{S}_{P,h,i}(t) = P(T_{P_{h,i}} > t \mid \boldsymbol{X_{h,i}})$ which equals $P(T_{P_{h,i}} > t \mid \boldsymbol{D_{h,i}})$ since $\boldsymbol{X_{h,i}} \setminus \boldsymbol{D_{h,i}}$ and $T_{P_{h,i}}$ are assumed to be independent. The corresponding conditional population all-cause cumulative hazard is denoted by $\tilde{\Lambda}_{P,h,i}$. We use life tables to calculate conditional population all-cause hazard functions according to individual demographic covariates such as age, sex and year of diagnosis that can be found in $\boldsymbol{D_{h,i}}$. We assume that these life tables describe adequately the all-cause death rates in the study population (Perme et al., 2012). Further, for each group h, the net survival function is defined as $S_{E,h}(t) = E(\tilde{S}_{E,h,1}(t))$ and thus we have $S_{E,h}(t) = P(T_{E_{h,1}} > t)$. Let $\Lambda_{E,h}$ denote the corresponding cumulative excess hazard. In the same way, we define the population all-cause survival by $S_{P,h}(t) = P(T_{P_{h,1}} > t)$ and the corresponding population all-cause cumulative hazard by $\Lambda_{P,h}$. Note that $\tilde{\lambda}_{E,h,i}$, $\tilde{\lambda}_{P,h,i}$, $\lambda_{E,h}$ and $\lambda_{P,h}$ denote the instantaneous hazards related to $\tilde{\Lambda}_{E,h,i}$, $\tilde{\Lambda}_{P,h,i}$, $\Lambda_{E,h}$ and $\Lambda_{P,h}$ respectively. We assumed that the conditional observed mortality hazard is the sum of the conditional population mortality hazard and the conditional excess mortality hazard:

$$\tilde{\lambda}_{P,h,i}(t) + \tilde{\lambda}_{E,h,i}(t).$$

Besides, we will also use the following additional assumptions to prove the asymptotic χ^2 distribution of our test statistic under the null:

$$\begin{aligned} \text{a)} \ & \int_0^T S_{E,h}(s)\lambda_{E,h}^2(s)ds < \infty, \\ \text{b)} \ \forall h \in [\![1;k]\!], \ & E(\frac{1}{\tilde{S}_{P,h,1}(T)^3}) < \infty, \\ \text{c)} \ \forall h \in [\![1;k]\!], \ & E(\int_0^T \frac{\tilde{\lambda}_{P,h,1}(s)^2 ds}{\tilde{S}_{P,h,1}(s)^3}) < \infty. \end{aligned} \tag{1}$$

where T is a constant denoting the maximum follow-up time. Note that these assumptions require that T is not too long compared with T_P or T_E. For instance, a) is not satisfied if $T_E < T$ (a.s.) and b) is not satisfied if $T_P < T$ (a.s.).

2.2 *The log-rank type statistic*

The usual log-rank test compares k cumulative observed hazard functions over $[0,T]$. Let $[0,T]$ denote the period of observation. The k-sample log-rank test is a test for the null hypothesis

$(H_0): \forall t \in [0,T]\,,\ \Lambda_1(t) = \ \ldots = \ \Lambda_k(t)$ where $k \geqslant 2$ is the number of groups to compare and $\Lambda_h\,(h \in [\![1;k]\!])$ is the cumulative observed hazard. Using counting process representations (see e.g. Andersen et al., 1993), the log-rank test is based on the following statistic:

$$Z_h(T) = \int_0^T \mathbb{1}(Y_{\cdot}(s) > 0) dN_h(s) - \int_0^T \mathbb{1}(Y_{\cdot}(s) > 0) \frac{Y_h(s)}{Y_{\cdot}(s)} dN_{\cdot}(s),$$

where $h \in [\![1;k]\!]$, $N_{h,i}(s) = \ \mathbb{1}(T_{h,i} \leqslant s, T_{h,i} \leqslant C_{h,i}) = \ \mathbb{1}(U_{h,i} \leqslant s, \tilde{\delta}_{h,i} = \ 1)$, $Y_{h,i}(s) = \ \mathbb{1}(T_{h,i} \geqslant s, C_{h,i} \geqslant s)$, $N_h(s) = \ \sum_{i=1}^{n_h} N_{h,i}(s)$, $Y_h(s) = \ \sum_{i=1}^{n_h} Y_{h,i}(s)$, $Y_{\cdot}(s) = \ \sum_{h=1}^{k} Y_h(s)$ and $N_{\cdot}(s) = \ \sum_{h=1}^{k} N_h(s)$ for $k \geqslant 2$. $Z_h(T)$ represents the difference between the number of observed deaths in the group h and the corresponding expected values.

Here, our goal is to test the null hypothesis

$$(H_0): \forall t \in [0,T]\,,\ \Lambda_{E,1}(t) = \ \ldots = \ \Lambda_{E,k}(t)$$

where $k \geqslant 2$. More precisely, we want to compare k cumulative excess hazard functions over this period using PPE (Perme et al., 2012). The PPE, $\hat{\Lambda}_{E,h}$, is a consistent estimator of $\Lambda_{E,h}$. It corrects the Ederer II estimator for those who left the sample due to deaths of other causes using the inverse probability weighting procedure (Robins, 1993). The weights are the survival probabilities of other causes and are applied to the counting and the at-risk processes. More precisely, we have $dN^w_{h,i}(s) = \ \frac{dN_{h,i}(s)}{\tilde{S}_{P,h,i}(s)}$, $Y^w_{h,i}(s) = \ \frac{Y_{h,i}(s)}{\tilde{S}_{P,h,i}(s)}$, $N^w_h(s) = \ \sum_{i=1}^{n_h} N^w_{h,i}(s)$, and $Y^w_h(s) = \ \sum_{i=1}^{n_h} Y^w_{h,i}(s)$ for $h \in \ [\![1;k]\!]$ and $k \geqslant 2$. The PPE is given by:

$$\forall k \geqslant 2,\ \forall h \in [\![1;k]\!],\ \hat{\Lambda}_{E,h}(t) = \int_0^t \frac{dN^w_h(s)}{Y^w_h(s)} - \int_0^t \frac{\sum_{i=1}^{n_h} Y^w_{h,i}(s) \tilde{\lambda}_{P,h,i}(s) ds}{Y^w_h(s)}.$$

To build our log-rank type test, we first have to consider another stochastic process related to the expected number of deaths due to cancer $N_{E,h}(s) = \sum_{i=1}^{n_h} N_{E,h,i}(s)$ where $N_{E,h,i}(s)$ is given by $N_{h,i}(s) - \int_0^s Y_{h,i}(u)\tilde{\lambda}_{P,h,i}(u)du$ for each patient i and for each group $h \in [\![1;k]\!]$. Second, we use the same weighting procedure as in the PPE. The expected weighted number of deaths due to cancer is then defined by $N^w_{E,h}(s) = \sum_{i=1}^{n_h} N^w_{E,h,i}(s)$ with $dN^w_{E,h,i}(s) = \frac{dN_{E,h,i}(s)}{\tilde{S}_{P,h,i}(s)}$. For all $h \in [\![1;k]\!]$, we now consider the statistic

$$Z^w_h(T) = \int_0^T \mathbb{1}(Y^w_{\cdot}(s) > 0) dN^w_{E,h}(s) - \int_0^T \mathbb{1}(Y^w_{\cdot}(s) > 0) \frac{Y^w_h(s)}{Y^w_{\cdot}(s)} dN^w_{E,\cdot}(s), \tag{2}$$

where $Y^w_{\cdot}(s) = \sum_{h=1}^{k} Y^w_h(s)$ and $dN^w_{E,\cdot}(s) = \sum_{h=1}^{k} dN^w_{E,h}(s)$ for $k \geqslant 2$.
Note that when $k = 2$, $Z^w_1(T)$ is given by

$$\begin{aligned}
&\int_0^T \mathbb{1}(Y^w_{\cdot}(s) > 0) dN^w_{E,1}(s) - \int_0^T \mathbb{1}(Y^w_{\cdot}(s) > 0) \frac{Y^w_1(s)}{Y^w_1(s) + Y^w_2(s)} \left(dN^w_{E,1}(s) + dN^w_{E,2}(s)\right) \\
&\quad = \int_0^T \mathbb{1}(Y^w_{\cdot}(s) > 0) \left(\frac{Y^w_2(s)}{Y^w_1(s) + Y^w_2(s)} dN^w_{E,1}(s) - \frac{Y^w_1(s)}{Y^w_1(s) + Y^w_2(s)} dN^w_{E,2}(s) \right).
\end{aligned}$$

The proposed test will be called log-rank type test because of the similarity between the two tests. For $h \in [\![1;k]\!]$, $\frac{dN^w_{E,h}(s)}{Y^w_h(s)}$ is a consistent estimator of the instantaneous excess hazard at time s, $\lambda_{E,h}(s)$ (Perme et al., 2012). It serves the same purpose as $\frac{dN_h(s)}{Y_h(s)}$ which is a consistent estimator of the instantaneous observed hazard at time s, $\lambda_h(s)$.

2.3 *Estimate of the variance of Z^w_h under the null*

We used martingale theory to estimate the variance of the statistic $Z^w_h(T)$ under the null. We start by looking at the case where T_{E_h} and $\boldsymbol{X_h}$ are independent for each $h \in [\![1;k]\!]$ i.e. we assume homogeneity in each group. This is a strong assumption usually made when studying the usual log-rank test (see e.g. Andersen et al., 1993). This assumption is frequently violated in practice, for example when cancer death is related to sex of patients. Then T_E and $\boldsymbol{X}$ are dependent. We will deal with this general case by building a stratified test presented in the next section.

Following the idea of the calculation of the estimate of the variance of the PPE (Perme et al.,

2012), we introduce

$$\begin{aligned} M_{h,i}(s) &\overset{def}{=} N_{h,i}(s) - \int_0^s Y_{h,i}(u)\left(\tilde{\lambda}_{P,h,i}(u) + \lambda_{E,h}(u)\right)du \\ &= N_{E,h,i}(s) - \int_0^s Y_{h,i}(u)\lambda_{E,h}(u)du. \end{aligned}$$

$M_{h,i}(s)$ is a local square integrable martingale with respect to the filtration $\mathcal{F}_s = \sigma\left(\boldsymbol{X_{h,i}},\ \mathbb{1}(U_{h,i} \leqslant u, U_{h,i} = T_{h,i}) : 0 \leqslant u \leqslant s;\ h \in [\![1;k]\!];\ 1 \leqslant i \leqslant n_h\right)$. Its predictable variation process $\langle M_{h,i}\rangle$ is given by $\int_0^s Y_{h,i}(u)\left(\tilde{\lambda}_{P,h,i}(u) + \lambda_{E,h}(u)\right)du$. Note that $\tilde{S}_{P,h,i}$ is $(\mathcal{F}_0)-$measurable so that we can define

$$dM_h^w(s) \overset{def}{=} \sum_{i=1}^{n_h} \frac{dM_{h,i}(s)}{\tilde{S}_{P,h,i}(s)} = dN_{E,h}^w(s) - Y_h^w(s)\lambda_{E,h}(s)ds, \tag{3}$$

and $M_h^w(s)$ is a local square integrable martingale with respect to $(\mathcal{F}_s)_s$.

Let Λ_E and λ_E denote $\Lambda_{E,h}$ and $\lambda_{E,h}$ under the null ($\forall h \in [\![1;k]\!]$). Then we have

$$dN_{E,\cdot}^w(s) = \sum_{h=1}^k dN_{E,h}^w(s) = \sum_{h=1}^k dM_h^w(s) + \lambda_E(s)\sum_{h=1}^k Y_h^w(s)ds. \tag{4}$$

Introducing (3) and (4) in formula (2), we obtain under the null

$$Z_h^w(T) = \sum_{l=1}^k \int_0^T \mathbb{1}(Y_\cdot^w(s) > 0)\left(\delta_{hl} - \frac{Y_h^w(s)}{Y_\cdot^w(s)}\right)dM_l^w(s),$$

with δ_{hl} being the Kronecker delta. For all $h \in [\![1;k]\!]$, Z_h^w are local square integrable martingales with respect to $(\mathcal{F}_s)_s$. We have $E\langle Z_h^w\rangle(T) < \infty$ since $\forall h \in [\![1;k]\!]$ $E\langle Z_h^w\rangle(T) \leqslant \sum_{l=1}^k n_l E\left\{\int_0^T \frac{S_{C,l,1}(s)S_E(s)}{\tilde{S}_{P,l,1}}\left(\tilde{\lambda}_{P,l,1}(s) + \lambda_E(s)\right)ds\right\} < \infty$ (see Web Appendix A). So the Z_h^w are square integrable over $[0,T]$.

As the first and second order moments of the Z_h^w exist, we have

$$cov\left(Z_h^w(T), Z_j^w(T)\right) = E[Z_h^w, Z_j^w](T),$$

$$[Z_h^w, Z_j^w](T) = \sum_{l=1}^k \left\{\int_0^T \mathbb{1}(Y_\cdot^w(s) > 0)\left(\delta_{hl} - \frac{Y_h^w(s)}{Y_\cdot^w(s)}\right)\left(\delta_{jl} - \frac{Y_j^w(s)}{Y_\cdot^w(s)}\right)\sum_{i=1}^{n_l}\frac{dN_{l,i}(s)}{\left(\tilde{S}_{P,l,i}(s)\right)^2}\right\}.$$

Note that, when $k = 2$, we have

$$[Z_1^w, Z_1^w](T) = \int_0^T \mathbb{1}(Y_\cdot^w(s) > 0) \left\{ \left(\frac{Y_2^w(s)}{Y_1^w(s) + Y_2^w(s)} \right)^2 \sum_{i=1}^{n_1} \frac{dN_{1,i}(s)}{\left(\tilde{S}_{P,1,i}(s) \right)^2} + \left(\frac{Y_1^w(s)}{Y_1^w(s) + Y_2^w(s)} \right)^2 \sum_{i=1}^{n_2} \frac{dN_{2,i}(s)}{\left(\tilde{S}_{P,2,i}(s) \right)^2} \right\}.$$

2.4 *The test statistic*

Following closely the usual log-rank test (Andersen et al., 1993), and knowing that $\sum_{h=1}^{k} Z_h^w(T) = 0$, we propose to test the null hypothesis with the statistic

$$U^w(T) = \boldsymbol{Z_0^w}(T)^t \hat{\boldsymbol{\Sigma}}_{\mathbf{0}}^{\mathbf{2},\boldsymbol{w}}(T)^{-1} \boldsymbol{Z_0^w}(T), \tag{5}$$

with $\boldsymbol{Z_0^w}(T) = \left(Z_1^w(T), \ldots, Z_{k-1}^w(T) \right)^t$ and $\hat{\boldsymbol{\Sigma}}_{\mathbf{0}}^{\mathbf{2},\boldsymbol{w}}$ being the matrix of general term

$$\hat{\sigma}_{h,j}^{2,w}(T) = \sum_{l=1}^{k} \left\{ \int_0^T \mathbb{1}(Y_\cdot^w(s) > 0) \left(\delta_{hl} - \frac{Y_h^w(s)}{Y_\cdot^w(s)} \right) \left(\delta_{jl} - \frac{Y_j^w(s)}{Y_\cdot^w(s)} \right) \sum_{i=1}^{n_l} \frac{dN_{l,i}(s)}{\left(\tilde{S}_{P,l,i}(s) \right)^2} \right\}$$

for $(h, j) \in [\![1; k-1]\!]^2$.

Under the assumptions (1) we can show that, under the null, $U^w(T) \sim \chi^2(k-1)$ when $n \longrightarrow \infty$ (see proof in Web Appendix B).

3. Stratified version of the test

We made the strong assumption of independence between T_E and $\boldsymbol{X}$ to estimate the variance of Z_h^w under the null. Now we look at the general case where T_E and $\boldsymbol{X}$ can be dependent. We define a set partition of the covariates set by $(I_1, \ldots, I_m)$ and we assume that $P(T_{E_h} > t \mid \boldsymbol{X_h}) = \sum_{s=1}^{m} P(T_{E_h} > t \mid \boldsymbol{X_h} \in I_s).\, \mathbb{1}(\boldsymbol{X_h} \in I_s)$, where $\boldsymbol{X_h}$ denotes the set of covariates in the group h. The $(I_s)_{1 \leqslant s \leqslant m}$ are called strata of one or more covariate. When cancer death is related to sex of patients, for example, we would consider 2 strata for men and women. Thus we assume homogeneity within each stratum but we allow heterogeneity between strata. We define $\Lambda_{E,h,s}$ as the cumulative excess hazard corresponding to the net

survival function $S_{E,h,s}(t) = P(T_{E_h} > t \mid \boldsymbol{X_h} \in I_s)$.

We want to test $(H_0) : \ \forall t \in [0,T]\,,\ \forall s \in [\![1;m]\!]\ \Lambda_{E,1,s}(t) = \ldots = \Lambda_{E,k,s}(t)$.

We define $Y^w_{.,s}(u) = \sum_{h=1}^{k} Y^w_{h,s}(u)$ with $Y^w_{h,s}(u) = \sum_{i=1}^{n_h} \frac{Y_{h,i}(u)}{\tilde{S}_{P,h,i}(u)} \mathbb{1}(\boldsymbol{X_{h,i}} \in I_s)$. In the same way, we define $dN^w_{E,.,s}(u) = \sum_{h=1}^{k} dN^w_{E,h,s}(u)$. Following Andersen et al. (1993), we define the statistics

$$Z^w_{h,s}(T) = \int_0^T \mathbb{1}(Y^w_{.,s}(u) > 0) dN^w_{E,h,s}(u) - \int_0^T \mathbb{1}(Y^w_{.,s}(u) > 0) \frac{Y^w_{h,s}(u)}{Y^w_{.,s}(u)} dN^w_{E,.,s}(u), \qquad (6)$$

and

$$\hat{\sigma}^{2,w}_{h,j,s}(T) = \sum_{l=1}^{k} \left\{ \int_0^T \mathbb{1}(Y^w_{.,s}(u) > 0) \left(\delta_{hl} - \frac{Y^w_{h,s}(u)}{Y^w_{.,s}(u)}\right) \left(\delta_{jl} - \frac{Y^w_{j,s}(u)}{Y^w_{.,s}(u)}\right) \times \sum_{i=1}^{n_l} \frac{dN_{l,i}(u)}{\left(\tilde{S}_{P,l,i}(u)\right)^2} \mathbb{1}(X_{l,i} \in I_s) \right\}. \qquad (7)$$

We denote for $s \in [\![1;m]\!]$ the vectors and matrices with elements given by (6) and (7) by $\boldsymbol{Z^w_s}$ and $\hat{\boldsymbol{\Sigma}}^{2,w}_s$. Then we will test the null hypothesis with the statistic

$$\left(\sum_{s=1}^{m} \boldsymbol{Z^w_{s,0}}(T)\right)^t . \left(\sum_{s=1}^{m} \hat{\boldsymbol{\Sigma}}^{2,w}_{s,0}(T)\right)^{-1} . \left(\sum_{s=1}^{m} \boldsymbol{Z^w_{s,0}}(T)\right),$$

which has asymptotic χ^2 distribution with $(k-1)$ degrees of freedom under the null. Note that, for $s \in [\![1;m]\!]$, $\boldsymbol{Z^w_{s,0}}(T) = \left(Z^w_{1,s}(T), \ldots, Z^w_{k-1,s}(T)\right)^t$ and $\hat{\boldsymbol{\Sigma}}^{2,w}_{s,0}$ is the same matrix as $\hat{\boldsymbol{\Sigma}}^{2,w}_s$ without the last row and the last column.

4. Simulations

We evaluated the performance of the proposed log-rank type test by simulation studies in the cases where T_E and $\boldsymbol{X}$ were (1) independent when $k = 2$ and $k = 3$; and (2) dependent when $k = 2$.

4.1 *Data generation and simulations design*

For each patient i, we independently generated covariates *sex*, *age* and G, which represents the groups (G had $k = 2$ or $k = 3$ levels). Covariate *sex* was generated from a binomial

distribution with $P(man) = P(woman) = {}^1/_2$. Covariate G was generated to study balanced cases ($P(G = 0) = P(G = 1)$ when $k = 2$ or $P(G = 0) = P(G = 1) = P(G = 2)$ when $k = 3$) or unbalanced cases only when $k = 2$ ($P(G = 0) = {}^1/_4$ and $P(G = 1) = {}^3/_4$). Because T_P depends on *age*, we studied 3 scenarios : (1) in the first scenario, we generated covariate *age* to represent approximately the empirical distribution of the ages of colon cancer patients in the French registries (25 percent of patients aged 40-64 years, 35 percent aged 65-74 years, and 40 percent aged 75 years and over); (2) in the second scenario, we studied a young population using a uniform distribution between 30 and 40; and (3) in the third scenario we studied an old population using a uniform distribution between 65 and 80.

Danieli et al. (2012) showed that the multivariable modelling estimator, which is based on the multivariable additive excess hazard model, is a consistent parametric estimator of net survival when adjusting for demographic covariates. Thus, we generated survival times from this model. In its classical additive form (Estève et al., 1990), the observed hazard related to the individual time of death, T_i, is defined as the sum of the instantaneous conditional population all-cause and excess hazards, $\tilde{\lambda}_{P,i}$ and $\tilde{\lambda}_{E,i}$. T_i was generated as follows: firstly, for each patient i, the time to death due to population hazard, T_{P_i}, was obtained from the 2004 American life table, `survexp.us`, stratified by $\boldsymbol{D_i} = (age_i, sex_i)$, and provided by the `survival` package in R software (Therneau, 2015). Secondly, for each patient i, the time to death due to cancer, T_{E_i}, was obtained from $\tilde{\lambda}_{E,i}$ modelled with the standard approach (see e.g. Giorgi et al., 2003) and using the inverse transformation method. More precisely, $\tilde{\lambda}_{E,i}(t) = f(t).\exp\left(\beta_{sex}\mathbb{1}(sex_i = man) + \sum_{l=1}^{k-1}\beta_{G,l}\mathbb{1}(G_i = l)\right)$ where β_{sex} and $\beta_{G,l}$ are the log hazard ratios (HR) of the covariates. The baseline hazard function f was modelled with a generalized Weibull distribution (Belot et al., 2010) as $t \longmapsto \dfrac{\kappa\rho^\kappa t^{\kappa-1}}{1 + \dfrac{(\rho t)^\kappa}{\alpha}}$ with $\rho = 0.5$, $\alpha = 0.2$ and $\kappa = 2$. The distributions of net survival between the groups that are defined by the levels of G vary when the effects of G on excess mortality vary. More precisely, the null is true when the

HR(s) of G equal 1. Conversely, the farther the HR(s) are from 1, the more different are the groups in terms of net survival and the farther we are from the null. When $k = 2$, the HR of G belonged to $\{0.7; 0.8; 0.9; 1; 1.2; 1.4; 1.6\}$. When $k = 3$, the HRs of G, (HR_1, HR_2), belonged to $\{(1, 0.7); (1, 1); (1, 1.2); (1, 1.4); (1, 1.6); (0.9, 1.2); (0.8, 1.4); (0.7, 1.6)\}$. When studying the case where T_E and $\boldsymbol{X}$ were independent, we did not introduce effects of age and sex on excess mortality (assumption of homogeneity). Conversely, to study the case where T_E and $\boldsymbol{X}$ were dependent, we set the HR of sex equal to 2 and 3 and we chose to assume independence with respect to age. But this could be done in the same way as done for sex. The bigger is the HR of sex, the more different are the distributions of the time to death due to cancer between men and women in the group h. Finally, individual censoring times, C_i, were generated from a uniform distribution $U[0; b]$, where the upper boundary b was selected to obtain approximately 0% or 30% overall censoring levels. Then, each individual's observable time of death was $T_i = min(T_{P_i}, T_{E_i})$ whereas each individual's observed time of death was $U_i = min(T_{P_i}, T_{E_i}, C_i)$. In addition, all subjects still at risk at 5 years were censored.

Moreover, we defined an individual's hypothetical time of death as the minimum of the excess death and censoring times. According to this time, we obtained another vital status corresponding to the hypothetical world where cancer would be the only cause of death. Thus, we could compare our test to the usual log-rank applied on data from hypothetical world. We will refer to them as "data from hypothetical world" and we will consider that the usual log-rank on these data is the gold standard. This is only possible within a simulation framework. Note that even if the cause specific data are available in our simulations, no direct gold standards for our log-rank test can be calculated in the "real world" since the real world is that of the competing risks and so still subject to informative censoring.

Each simulation run consisted of 2000 independent samples. Each of them contained 1000 patients.

4.2 *Simulation results*

Results obtained with no censoring were roughly equivalent to those obtained with 30% censoring. So we show only those related to 30% censoring level.

When compared 2 groups, the estimation of the type I error of our log-rank type test was good. In table 1, at a 5% level of significance, the confidence intervals for the estimation of the type I error contain the nominal level of 5% for our test and the usual log-rank applied on data from hypothetical world. In comparison with the usual log-rank, our test performed well in terms of power (table 1). In the second scenario, where the patients under study are young, the results were nearly the same for both tests. Nevertheless, there was a loss of power for our proposed test in the first and the third scenarios. Note that in the first scenario there were 75% of patients aged more than 65.

[Table 1 about here.]

As expected, whatever the scenario, both tests were more powerful when the number of patients increased from 500 to 2000 (results not shown) and they lost power when the cases were unbalanced (Web Table A).

When studying the comparison of 3 groups, the estimation of the type I error was close to the nominal level of 5% (table 2). In terms of power, in the first scenario, table 2 shows that our proposed test performed worse than the usual log-rank in the hypothetical world, especially when the 3 distributions of net survival were not really away from each other ($(HR_1, HR_2) = (1, 0.7)$ or $(0.9, 1.2)$). In the other cases, the results of both tests were as powerful. In addition, as previously, our test had a similar power as the usual log-rank when patients were young and we observed a loss of power in scenarios 1 and 3 (Web Table B).

[Table 2 about here.]

When studying the comparison of 2 groups when T_E and the covariate sex were dependent, we compared results from the stratified version of our test with the not-stratified version. As

expected, there was a loss of power when using the test which was not stratified (table 3). The farther β_{sex} is from 0, the bigger was this loss of power. More interestingly, as shown in table 3, when the conditional distributions of T_E were the most different ($HR_{sex} = 3$), the estimation of the type I error was equal to 2.95, 95% Confidence Interval (CI) = $[2.21; 3.69]$, when using the not stratified version of our test vs 4.60, $95\%CI = [3.68; 5.52]$, with the stratified version. However, it was equal to 4.80, $95\%CI = [3.86; 5.74]$, vs 5.45, $95\%CI = [4.46; 6.44]$, when $HR_{sex} = 2$. Thus, the stratified log-rank type test has to be used when the stratum variable has an important impact on net survival.

[Table 3 about here.]

5. Application

We applied the proposed test in one application for illustration. This analysis considered survival data on 10,108 patients with colorectal cancer diagnosed in 1998. These data came from 17 US registries obtained from the Surveillance, Epidemiology, and End Results (SEER) Program (2006) in the US. From this cohort, we excluded 816 patients who had no surgical procedure of the primary site, 2 patients in whom the use of a surgical procedure was not certain, and 167 patients with *in situ* tumors. Patient follow-up was restricted to the first five years after diagnosis and censoring set at five years in still alive patients. This left 9,123 patients for analysis. The covariates used were age at diagnosis, sex, ethnicity (black or white), and cancer stage at diagnosis (in four stages I to IV according to the stage classification of the American Joint Committee on Cancer used by SEER registries (SEER Program: comparative staging guide for cancer, 1993)). This data set is described in Web Table C.

We used the American life tables provided by R software `survexp.usr`, that is to say life tables stratified by age, sex, ethnicity and calendar year, from 1998 to 2003. All the analyses

were computed using R (R Core Team, 2014). The code and the .RData files are available upon request.

We used our test to compare net survival distributions between Black and White patients stratified on stage, which is known to have an important effect on net cancer survival. Moreover in net survival framework age is a strong prognostic factor in several types of cancer (Bossard et al., 2007). We built three age groups to have adults and young old patients (20-69 years), old patients (70-79 years) and very old patients ($\geqslant$ 80 years). We stratified also on these groups thereby obtaining 12 strata. Figure 1 shows the impact of age and most importantly of stages on net survival for these real data. Firstly using a test not stratified produced a test statistic equal to 19.95 (p-value $= 7.9 \times 10^{-6}$). Secondly when running our test stratified on stage, we found a test statistic equal to 5.42 (p-value $= 0.0199$). The lower proportion of Black patients in lower stages (47% in stages I-II vs. 56% for White patients) suggested later diagnosis, but even after correcting for this, the impact of ethnicity on cancer mortality remained significant and higher for Black patients. Thirdly when running our test stratified on age, we found a test statistic equal to 23.62 (p-value $= 1.2 \times 10^{-6}$). Whatever the age group, differences between net survival of Black and White patients were indeed bigger considering age strata rather than stage strata (data not shown). Finally when running our test stratified on both age and stage, we found a test statistic equal to 9.92 (p-value $= 0.0016$). Thus not stratifying on stage overestimated the differences between net survival distributions of Black and White people whereas not stratifying on age underestimated these differences. Stratifying on both provided the "true" differences which had been first distorted by heterogeneity between groups. Note that using the log-rank test on observed survival led to a test statistic equal to 19.5. So using net survival instead of observed survival allowed to remove the confounding effect of age on observed survival.

[Figure 1 about here.]

6. Discussion

Our proposed test compares distribution of net survival estimated by the Pohar-Perme estimator (Perme et al., 2012). The simulation study showed that the estimation of the type I error is correct. Our test also performs well in terms of power even if we observed a loss of power when the studied patients were old. This loss of power could be explained by the fact that elderly patients have higher expected mortality rates, that is to say there are more deaths due to other causes. Thus, there is a loss of information and higher variability in the estimates of net survival.

The stratified version is useful when dealing with covariates impacting strongly on net survival, that is to say when there is one or more covariate having different distributions in the groups to compare (see e.g. Aalen et al., 2008, p. 110-111). The decision to use the stratified version should be based on epidemiological considerations depending on studied covariates. The application on real data showed that part of difference in net cancer survival between Black and White patients is due to differences in stages.

We took the same set of assumptions as in Perme et al. (2012). T_E and T_P being two latent times defined on the same individual, they could be dependent conditionally on the covariates only via some unmeasured covariates (e.g. deprivation or smoking habits of the same individual). In addition, we made assumptions (1) in the proof of the asymptotic distribution of the statistic under the null. These are reasonable assumptions on follow-up time because they require to use small follow-up times compared with T_P given D or T_E.

A possible limitation of our work is that we only studied simulations favourable to our test. Indeed, the usual log-rank is optimal under the assumption of proportional hazard rates but performs poorly when this assumption does not hold (Qiu and Sheng, 2008). Several approaches have been proposed to deal with this problem (see e.g. Fleming et al.,

1980; Mantel and Stablein, 1988 ; Breslow, Edler, and Berger, 1984; Qiu and Sheng, 2008). Further studies are needed to adapt our proposed test starting from one of these procedures. In addition, the formula we proposed was developed with a continuous underline process (without ties). Nevertheless, event times are usually assumed to be discrete when testing (Aalen et al., 2008). In our application, there were 46% of ties between event times since only survival in months was available from the SEER. We studied the impact of the use of a non tie-corrected version of our test by simulation rounding survival times to obtain 38%, 45% and 54% of ties. Comparing the percentages of rejection of the null running the test on the same dataset with and without ties led to a maximum difference of 2% (results not shown). Thus using a non tie-corrected version of the test had hardly any impact with such percentages of ties. However a tie-corrected estimator adapted from the one presented by Andersen et al. (1993) may be of interest.

Another option to compare distributions of net survival is to use regression modelling. We compared our proposed test with the likelihood ratio test from the multivariate excess mortality model using simulations datasets, both presented in section 4.1. We assumed an excess mortality model perfectly defined, i.e. adjusted on G and sex (if needed) with proportional effect (results not shown). In terms of power, the biggest difference between the percentages of rejection of the null hypothesis at the 5% level of significance for the 2 tests was 3.15 in favour of the likelihood ratio test. However, with our proposed test, we did not have to deal with the model-building strategy (see e.g. Wynant and Abrahamowicz, 2014) within this known setting. Therefore, our non parametric test should be preferred because of its simplicity.

Since our test compares favorably with the usual log-rank on data from hypothetical world, as shown in the simulation study, it may be helpful for cancer registries to compare net cancer survival between countries or areas. In addition, it may be applied to other chronic

diseases for which net survival should be used. In the same way Schoenfeld (1981) did with the usual log-rank, it would be interesting to determine the distribution of the test statistic under the alternative hypothesis. Then deriving his formula, we could obtain the sample size providing the minimal detectable difference. Another perspective would be based on the equivalence between the usual log-rank and the score test from a Cox model. Introducing in a Cox model time dependent weights corresponding to the ones used in the Pohar-Perme estimator could be an interesting approach to investigate.

Acknowledgements

The work of the first author was funded by grants from INCa (COMPNETS project, INCa SHS-E-SP 2013). The authors thank the referees and the coeditor for many important remarks which improved the article. The authors are also very grateful to the CENSUR working survival group for their helpful comments.

Supplementary Materials

Web Appendices and Tables referenced in Sections 2.3, 2.4, 4.2 and 5 are available with this paper at the Biometrics website on Wiley Online Library.

References

Aalen, O., Borgan, O., and Gjessing, H. (2008). *Survival and event history analysis: a process point of view*. Springer.

Allemani, C., Weir, H. K., Carreira, H., Harewood, R., Spika, D., Wang, X.-S., et al. (2015). Global surveillance of cancer survival 1995–2009: analysis of individual data for 25 676 887 patients from 279 population-based registries in 67 countries (CONCORD-2). *The Lancet* **385,** 977–1010.

Andersen, P. K., Borgan, O., Gill, R. D., and Keiding, N. (1993). *Statistical models based on counting processes*. Springer.

Belot, A., Abrahamowicz, M., Remontet, L., and Giorgi, R. (2010). Flexible modeling of competing risks in survival analysis. *Statistics in medicine* **29,** 2453–2468.

Berkson, J. and Gage, R. P. (1950). Calculation of survival rates for cancer. In *Proceedings of the staff meetings. Mayo Clinic*, volume 25, pages 270–286.

Bossard, N., Velten, M., Remontet, L., Belot, A., Maarouf, N., Bouvier, A. M., et al. (2007). Survival of cancer patients in France: a population-based study from The Association of the French Cancer Registries (FRANCIM). *European Journal of Cancer* **43,** 149–160.

Breslow, N. E., Edler, L., and Berger, J. (1984). A two-sample censored-data rank test for acceleration. *Biometrics* **40,** 1049–1062.

Danieli, C., Remontet, L., Bossard, N., Roche, L., and Belot, A. (2012). Estimating net survival: the importance of allowing for informative censoring. *Statistics in medicine* **31,** 775–786.

De Angelis, R., Sant, M., Coleman, M. P., Francisci, S., Baili, P., Pierannunzio, D., et al. (2014). Cancer survival in Europe 1999–2007 by country and age: results of EUROCARE-5a population-based study. *The Lancet Oncology* **15,** 23–34.

Ederer, F., Axtell, L. M., and Cutler, S. J. (1961). The relative survival rate: a statistical methodology. *National Cancer Institute Monograph* **6,** 101–121.

Ederer, F. and Heise, H. (1959). The effect of eliminating deaths from cancer on general population survival rates, methodological note 11: End results evaluation section. *The effect of eliminating deaths from cancer on general population survival rates, methodological note 11: End results evaluation section* .

Estève, J., Benhamou, E., Croasdale, M., and Raymond, L. (1990). Relative survival and the estimation of net survival: elements for further discussion. *Statistics in medicine* **9,**

529–538.

Fleming, T. R. and Harrington, D. P. (2011). *Counting processes and survival analysis*. John Wiley & Sons.

Fleming, T. R., Harrington, D. P., and O'sullivan, M. (1987). Supremum versions of the log-rank and generalized Wilcoxon statistics. *Journal of the American Statistical Association* **82,** 312–320.

Fleming, T. R., O'Fallon, J. R., O'Brien, P. C., and Harrington, D. P. (1980). Modified Kolmogorov-Smirnov test procedures with application to arbitrarily right-censored data. *Biometrics* **36,** 607–625.

Giorgi, R., Abrahamowicz, M., Quantin, C., Bolard, P., Estève, J., Gouvernet, J., et al. (2003). A relative survival regression model using B-spline functions to model non-proportional hazards. *Statistics in medicine* **22,** 2767–2784.

Hakulinen, T. (1982). Cancer survival corrected for heterogeneity in patient withdrawal. *Biometrics* **38,** 933–942.

Howlader, N., Noone, A., Krapcho, M., Neyman, N., Aminou, R., Waldron, W., et al. (2011). SEER cancer statistics review, 1975–2008. *Bethesda, MD: National Cancer Institute* .

Mantel, N. (1966). Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer chemotherapy reports. Part 1* **50,** 163–170.

Mantel, N. and Stablein, D. M. (1988). The crossing hazard function problem. *The Statistician* **37,** 59–64.

Perme, M. P., Stare, J., and Estève, J. (2012). On estimation in relative survival. *Biometrics* **68,** 113–120.

Peto, R. and Peto, J. (1972). Asymptotically efficient rank invariant test procedures. *Journal of the Royal Statistical Society, Series A (General)* **135,** 185–207.

Qiu, P. and Sheng, J. (2008). A two-stage procedure for comparing hazard rate functions.

Journal of the Royal Statistical Society: Series B (Statistical Methodology) **70,** 191–208.

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Remontet, L., Bossard, N., Belot, A., and Estéve, J. (2007). An overall strategy based on regression models to estimate relative survival and model the effects of prognostic factors in cancer survival studies. *Statistics in medicine* **26,** 2214–2228.

Robins, J. M. (1993). Information recovery and bias adjustment in proportional hazards regression analysis of randomized trials using surrogate markers. In *Proceedings of the Biopharmaceutical Section, American Statistical Association*, pages 24–33. Alexandria, Virgnia, U.S.

Schoenfeld, D. (1981). The asymptotic properties of nonparametric tests for comparing survival distributions. *Biometrika* **68,** 316–319.

SEER Program: comparative staging guide for cancer (1993). *NIH Publication No. 93-3640*.

Surveillance, Epidemiology, and End Results (SEER) Program (Based on the submission November 2006). *SEER*Stat Database: Incidence - SEER 17 Regs Research Data, Nov 2006 Sub (1973-2004 varying) - Linked To County Attributes - Total U.S., 1969-2004 Counties, National Cancer Institute, DCCPS, Surveillance Research Program, Cancer Statistics Branch, released April 2007*.

Therneau, T. M. (2015). *A Package for Survival Analysis in S*. version 2.38.

Wynant, W. and Abrahamowicz, M. (2014). Impact of the model-building strategy on inference about nonlinear and time-dependent covariate effects in survival analysis. *Statistics in medicine* **33,** 3318–3337.

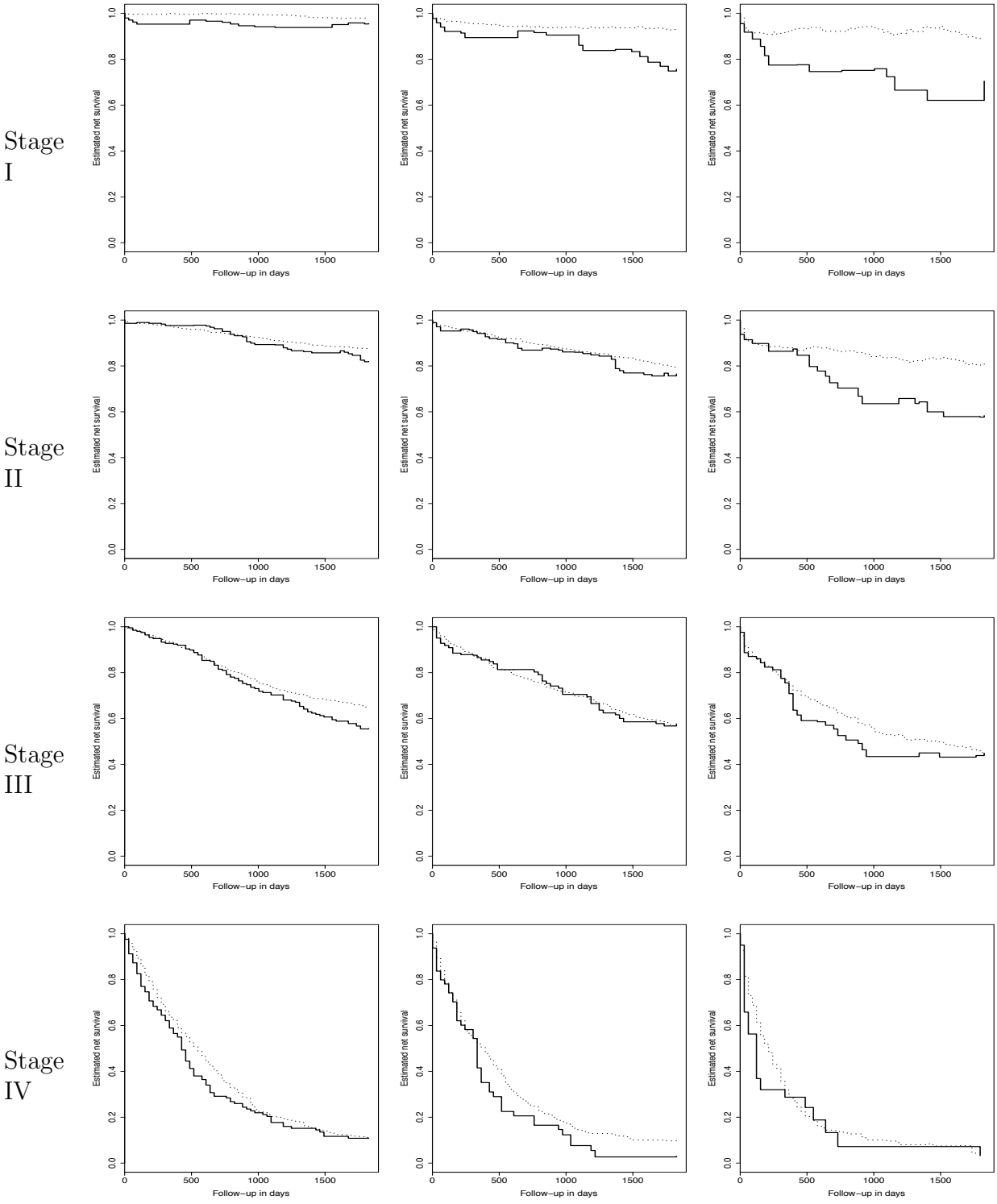

Figure 1. Net survival estimated by the Pohar-Perme estimator according to stage and age groups for: . . . . . , White patients; _____, Black patients. When the excess hazard is close to 0 and/or when the number at risk is low, the variability of the Pohar-Perme estimator can cause non-monotonic net survival curves.

Table 1

Comparison of 2 groups: percentage of rejection of the null hypothesis at the 5% level of significance for 2000 simulations of 1000 patients. Distribution of age specific to each scenario: Scenario 1: 25% aged $[40-64]$*, 35% aged* $[65-74]$*, and 40% aged* $[75-85]$*; Scenario 2:* $30 \leqslant age \leqslant 40$ *(uniform); Scenario 3:* $65 \leqslant age \leqslant 80$ *(uniform).*

HR^a	Percentage of rejection of the following tests ($95\%CI$)			
	Proposed test		Usual log-rank on data from hypothetical world	
	Scenario 1: balanced case[b]			
0.7	81.50	(79.80;83.20)	93.05	(91.94;94.16)
0.8	44.85	(42.67;47.03)	59.85	(57.70;62.00)
0.9	15.55	(13.96;17.14)	20.35	(18.59;22.11)
1	5.20	(4.23;6.17)	5.30	(4.32;6.28)
1.2	35.95	(33.85;38.05)	46.70	(44.51;48.89)
1.4	88.30	(86.89;89.71)	95.05	(94.10;96.00)
1.6	99.50	(99.19;99.81)	100	(99.81;100)
	Scenario 2: balanced case[b]			
0.7	91.80	(90.60;93.00)	92.20	(91.02;93.38)
0.8	56.90	(54.73;59.07)	57.60	(55.43;59.77)
0.9	18.15	(16.46;19.84)	18.25	(16.56;19.94)
1	4.15	(3.28;5.02)	4.35	(3.46;5.24)
1.2	47.80	(45.61;49.99)	48.45	(46.26;50.64)
1.4	94.90	(93.94;95.86)	95.30	(94.37;96.23)
1.6	99.90	(99.64;99.97)	99.90	(99.64;99.97)
	Scenario 3: balanced case[b]			
0.7	82.20	(80.52;83.88)	92.00	(90.81;93.19)
0.8	47.85	(45.66;50.04)	58.75	(56.59;60.91)
0.9	13.85	(12.34;15.36)	17.10	(15.45;18.75)
1	5.35	(4.36;6.34)	4.30	(3.41;5.19)
1.2	39.20	(37.06;41.34)	48.75	(46.56;50.94)
1.4	88.20	(86.79;89.61)	95.25	(94.32;96.18)
1.6	99.10	(98.69;99.51)	99.85	(99.56;99.95)

[a]: Hazard Ratio of the level of G on excess mortality used in data generation, where G is the covariate representing the groups;

[b]: Balanced cases correspond to the cases where groups are similar in size with $P(G = 0) = P(G = 1)$.

Table 2

Comparison of 3 groups: percentage of rejection of the null hypothesis at the 5% level of significance for 2000 simulations of 1000 patients. Distribution of age (scenario 1): 25% aged $[40-64]$*, 35% aged* $[65-74]$*, and 40% aged* $[75-85]$*.*

$(HR_1, HR_2)^a$	Percentage of rejection of the following tests $(95\%CI)$			
	Proposed test		Usual log-rank on data from hypothetical world	
	Scenario 1: balanced case[b]			
(1, 0.7)	66.75	(64.69;68.81)	82.90	(81.25;84.55)
(1, 1)	5.10	(4.14;6.06)	4.95	(4.00;5.90)
(1, 1.2)	26.20	(24.27;28.13)	35.80	(33.70;37.90)
(1, 1.4)	74.65	(72.74;76.56)	87.35	(85.89;88.81)
(1, 1.6)	97.20	(96.48;97.92)	99.70	(99.46;99.94)
(0.9, 1.2)	42.40	(40.23;44.57)	58.20	(56.04;60.36)
(0.8, 1.4)	96.10	(95.25;96.95)	98.90	(98.44;99.36)
(0.7, 1.6)	100	(99.81;100)	100	(99.81;100)

[a]: Hazard Ratios of the levels of G on excess mortality used in data generation, where G is the covariate representing the groups;

[b]: Balanced cases correspond to the cases where groups are similar in size with $P(G=0) = P(G=1) = P(G=2)$.

Table 3

Comparison of 2 groups: percentage of rejection of the null hypothesis at the 5% level of significance for 2000 simulations of 1000 patients when sex has an impact on excess mortality in the data generation. Distribution of age specific to scenario 1: 25% aged $[40-64]$*, 35% aged* $[65-74]$*, and 40% aged* $[75-85]$*.*

HR^a	Percentage of rejection of the following tests $(95\%CI)$			
	Proposed stratified test		Proposed test (not stratified)	
	Scenario 1: $HR_{sex}=2$			
0.7	90.60	(89.32;91.88)	88.55	(87.15;89.95)
0.8	57.90	(55.74;60.06)	53.25	(51.06;55.44)
0.9	18.00	(16.32;19.68)	16.40	(14.78;18.02)
1	5.45	(4.46;6.44)	4.80	(3.86;5.74)
1.2	46.50	(44.31;48.69)	43.50	(41.33;45.67)
1.4	95.00	(94.04;95.96)	93.35	(92.26;94.44)
1.6	99.90	(99.64;99.97)	99.85	(99.56;99.95)
	Scenario 1: $HR_{sex}=3$			
0.7	93.70	(92.74;94.76)	88.30	(86.89;89.71)
0.8	61.80	(59.67;63.93)	51.25	(49.06;53.44)
0.9	18.25	(16.56;19.94)	14.15	(12.62;15.68)
1	4.60	(3.68;5.52)	2.95	(2.21;3.69)
1.2	50.30	(48.11;52.49)	40.90	(38.75;43.05)
1.4	95.35	(94.43;96.27)	91.40	(90.17;92.63)
1.6	100	(99.81;100)	99.90	(99.64;99.97)

[a]: Hazard Ratios of the levels of G on excess mortality used in data generation, where G is the covariate representing the groups.