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[4-8], and Roch Giorgi[1,2,9,*]

[1]INSERM, UMR912 " Sciences Économiques et Sociales de la Santé et Traitement de l'Information Médicale " (SESSTIM), F-13006, Marseille, France

[2]Université d'Aix-Marseille, UMR S912, IRD, F-13006, Marseille, France

[3]Université d'Aix-Marseille, CNRS, Centrale Marseille, I2M, UMR 7373, F-13453, Marseille, France

[4]Service de Biostatistique, Hospices Civils de Lyon, F-69003, Lyon, France

[5]Université de Lyon, F-69000, Lyon, France

[6]Université Lyon 1, F-69100, Villeurbanne, France

[7]CNRS, UMR5558, Laboratoire de Biométrie et Biologie Évolutive, Équipe Biostatistique-Santé, F-69100, Villeurbanne, France

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

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

**email:* roch.giorgi@univ-amu.fr

Summary: In population-based cancer studies, it is often interesting 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 if the disease under study 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 (as estimated by PPE) between several groups. We put the test within the counting process framework to introduce the inverse probability weighting procedure as required by the PPE. We built a stratified version to control for categorical covariates that affect the outcome. We performed simulation studies to evaluate the performance of this test and worked an application on real data.

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 with the excess mortality hazard, is the survival that would be observed in a hypothetical world where the disease of interest would be the only possible cause of death. The observed survival, which is more frequently used, is the result of two main survival components: one related to death from the disease and another to death from all other causes (Estève et al., 1990; Perme, Stare, and Estève, 2012). On the one hand, the observed survival does not distinguish between death from the disease of interest (or excess mortality) and death from other causes. On the other hand, net survival evaluates the burden of the disease independently of the general-population mortality; i.e., the mortality from other causes as given by life tables. In cancer research, the idea of cancer net survival is the study of the proportion of deaths due directly or indirectly to cancer. This epidemiological indicator is routinely estimated by cancer registries and population-based studies such as the EUROCARE program (De Angelis et al., 2014), the US SEER program (Howlader et al., 2011), or the CONCORD program (Allemani et al., 2015). It is crucial for comparisons between different populations (Perme et al., 2012; Danieli et al., 2012). For instance, in comparing the patterns of care between countries, it is essential to take into account the general-population mortality because of its impact on the observed survival.

In population-based studies, the exact causes of death are often unavailable and, when available, often difficult to link to a given disease (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). Nevertheless, in 2012, Perme et al. (2012) considered that, in most cases, these estimators do not estimate net survival. These authors proposed then a non-parametric estimator that corrects Ederer II estimator (Ederer and Heise, 1959) now known to be biased because of informative censoring. For instance, the

excess mortality and mortality from other causes are both age-dependent; this leads to an informative censoring. Perme et al. used population mortality to weight and correct for the subjects who leave the sample due to deaths from other causes. Later, Danieli et al. (2012) showed by simulation that the Pohar-Perme estimator (PPE) is a consistent non-parametric estimator of net survival that may be preferred to other non-parametric estimators. The PPE estimates a hypothetical quantity that allows comparisons between populations. However, to the best of our knowledge, comparing distributions of net survival over a given period in the non-parametric framework is not possible yet; only a comparison between two estimates at a given time t is currently possible with a classical Z-test. Besides, in the parametric framework, the comparison of net survival between two groups or more may use a likelihood ratio test in the context of a multivariate excess mortality model (see e.g., Remontet et al., 2007), but this requires a complex model building strategy.

In this paper, we propose a log-rank-type test to compare distributions of net survival as estimated by the PPE between two groups or more 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 very commonly used to compare distributions of observed survival between at least two groups. Second, 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 written with stochastic processes on the cumulative hazard scale, the log-rank test allows an easy introduction of the weights of the PPE into the corresponding counting processes.

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

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

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

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

Under these assumptions: $\lim_{n\to\infty} \min_h n_h = \infty$.

2.1 *Notations and model*

We consider that $T_{h,i}$, the time to death of each patient i in group h, is the minimum of two distinct times: $T_{P_{h,i}}$ due to the "population hazard" and $T_{E_{h,i}}$ due to the "excess hazard". Let $C_{h,i}$ be the censoring time and $U_{h,i} = \min(T_{h,i}, C_{h,i})$ the follow-up time of patient i. $\tilde{\delta}_{h,i}$ denotes the failure indicator equal to 1 when the true failure time $T_{h,i}$ is observed and 0 when patient i is censored (the use of $\sim$ avoids a possible confusion with the notation of Kronecker delta). Each patient i in a group h has covariates denoted by vector $\boldsymbol{X_{h,i}}$. $\boldsymbol{D_{h,i}}$ is a sub-vector of $\boldsymbol{X_{h,i}}$ that describes all the demographic covariates such that $\boldsymbol{X_{h,i}} \setminus \boldsymbol{D_{h,i}}$ and $T_{P_{h,i}}$ are independent. Besides, we adopt the same set of assumptions as in Perme et al. (2012):

$$\begin{aligned}
&\text{a) } (T_{P_{h,i}}, T_{E_{h,i}}, C_{h,i}, \boldsymbol{X_{h,i}})_{h,i} \text{ are mutually independent;}\\
&\text{b) } (T_{P_{h,i}}, T_{E_{h,i}}, C_{h,i}, \boldsymbol{X_{h,i}})_{i} \text{ have the same distribution;}\\
&\text{c) } T_{E_{h,i}} \text{ and } T_{P_{h,i}} \text{ are conditionally independent given } \boldsymbol{X_{h,i}};\\
&\text{d) censoring times } C_{h,i} \text{ are independent of pair } (T_{h,i}, \boldsymbol{X_{h,i}}).
\end{aligned} \tag{1}$$

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 pa-

tient i in group h. The conditional net survival function of $T_{E_{h,i}}$ that corresponds to each patient i of 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 $P(T_{P_{h,i}} > t \mid \boldsymbol{D_{h,i}})$ because $\boldsymbol{X_{h,i}} \setminus \boldsymbol{D_{h,i}}$ 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 the 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 $S_{E,h}(t) = E(\tilde{S}_{E,h,1}(t))$; 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. From assumption (1.c), it follows that the conditional observed mortality hazard is the sum of the conditional population mortality hazard plus the conditional excess mortality hazard: $\tilde{\lambda}_{P,h,i}(t) + \tilde{\lambda}_{E,h,i}(t)$.

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

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

where T is a constant that represents the maximum follow-up time. Note that these assumptions require that T be not too long vs. T_P or T_E. For instance, (2.a) is not satisfied if $T_E < T$ (a.s.) and (2.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 $[0,T]$, where $[0,T]$ denotes 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 $\Lambda_h\,(h \in [\![1;k]\!])$ 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 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 Ederer II estimator for the subjects who leave the sample due to deaths from other causes using the inverse probability weighting procedure (Robins, 1993). The weights are the survival probabilities of death from other causes that 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 from 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 each group $h \in [\![1;k]\!]$. Second, we use the same weighting procedure as in the PPE. The expected weighted number of deaths from 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

$$Z^w_h(T) = \int_0^T \mathbb{1}(Y^w_{\cdot}(s) > 0) dN^w_{E,h}(s) - \int_0^T \mathbb{1}(Y^w_{\cdot}(s) > 0) \frac{Y^w_h(s)}{Y^w_{\cdot}(s)} dN^w_{E,\cdot}(s), \tag{3}$$

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 may be considered as a "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 observed instantaneous hazard at time s, $\lambda_h(s)$.

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

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

Consistently with the idea of calculating 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

$$dM_h^w(s) \overset{def}{=} \sum_{i=1}^{n_h} \frac{dM_{h,i}(s)}{\tilde{S}_{P,h,i}(s)} = dN_{E,h}^w(s) - Y_h^w(s)\lambda_{E,h}(s)ds, \tag{4}$$

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 hypothesis $(\forall h \in [\![1;k]\!])$. Then we have

$$dN_{E,\cdot}^w(s) = \sum_{h=1}^k dN_{E,h}^w(s) = \sum_{h=1}^k dM_h^w(s) + \lambda_E(s)\sum_{h=1}^k Y_h^w(s)ds. \tag{5}$$

Introducing (4) and (5) in formula (3), we obtain under the null hypothesis

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

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 $[0,T]$.

Because 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 following statistic

$$U^w(T) = \boldsymbol{Z_0^w}(T)^t \hat{\boldsymbol{\Sigma}}_{\mathbf{0}}^{\mathbf{2,w}}(T)^{-1} \boldsymbol{Z_0^w}(T), \tag{6}$$

where $\boldsymbol{Z_0^w}(T) = \left(Z_1^w(T), \ldots, Z_{k-1}^w(T) \right)^t$ and $\hat{\boldsymbol{\Sigma}}_{\mathbf{0}}^{\mathbf{2,w}}$ is 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 assumptions (2), we can show that, under the null hypothesis, $U^w(T) \sim \chi^2(k-1)$ when $n \longrightarrow \infty$ (see proof in Web Appendix B).

3. Stratified version of the test

To estimate the variance of Z_h^w under the null hypothesis, we made the strong assumption of independence between T_E and $\boldsymbol{X}$. Now we look at the general case where T_E and $\boldsymbol{X}$ can be dependent. We define a partition of the covariate space by $(I_1, \ldots, I_m)$ and 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 group h. The $(I_s)_{1 \leqslant s \leqslant m}$ are called strata of one or more covariate. For example, when death from cancer is related to a given sex, we may consider two strata (men and women). Thus, we assume homogeneity within each stratum but allow heterogeneity between strata. We define $\Lambda_{E,h,s}$ as the cumulative excess hazard that corresponds 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 $dN^w_{E,.,s}(u) = \sum_{h=1}^{k} dN^w_{E,h,s}(u)$. According to Andersen et al. (1993), we define the statistics

$$Z^w_{h,s}(T) = \int_0^T \mathbb{1}(Y^w_{.,s}(u) > 0) dN^w_{E,h,s}(u) - \int_0^T \mathbb{1}(Y^w_{.,s}(u) > 0) \frac{Y^w_{h,s}(u)}{Y^w_{.,s}(u)} dN^w_{E,.,s}(u), \qquad (7)$$

and

$$\hat{\sigma}^{2,w}_{h,j,s}(T) = \sum_{l=1}^{k} \left\{ \int_0^T \mathbb{1}(Y^w_{.,s}(u) > 0) \left(\delta_{hl} - \frac{Y^w_{h,s}(u)}{Y^w_{.,s}(u)}\right) \left(\delta_{jl} - \frac{Y^w_{j,s}(u)}{Y^w_{.,s}(u)}\right) \times \sum_{i=1}^{n_l} \frac{dN_{l,i}(u)}{\left(\tilde{S}_{P,l,i}(u)\right)^2} \mathbb{1}(X_{l,i} \in I_s) \right\}. \qquad (8)$$

We denote for $s \in [\![1;m]\!]$ the vectors and matrices with elements given by (7) and (8) 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, under the null hypothesis, has an asymptotic χ^2 distribution with $(k-1)$ degrees of freedom. 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 cases where T_E and $\boldsymbol{X}$ are independent with $k = 2$ or 3 and where T_E and $\boldsymbol{X}$ are dependent with $k = 2$.

4.1 *Data generation and simulation design*

For each patient i, we generated independently 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 (i.e., $P(G = 0) = P(G = 1)$ with $k = 2$ or $P(G = 0) = P(G = 1) = P(G = 2)$ with $k = 3$) or unbalanced cases only (i.e., $P(G = 0) = {}^1/_4$ and $P(G = 1) = {}^3/_4$ with $k = 2$). Because T_P depends on age, we studied 3 scenarios: 1) we generated covariate age to represent approximately the empirical distribution of ages of colon cancer patients in the French registries (25 percent aged 40-64 years, 35 percent aged 65-74 years, and 40 percent aged 75 years and over); 2) we studied a young population using a uniform distribution between ages 30 and 40; and 3) we studied an old population using a uniform distribution between ages 65 and 80.

Danieli et al. (2012) have shown that the multivariable modeling estimator, which is based on the multivariable additive excess hazard model, is a consistent parametric estimator of net survival after adjustment on 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 to death T_i is defined as the sum of the instantaneous conditional population all-cause death hazard and disease excess hazard, $\tilde{\lambda}_{P,i}$ and $\tilde{\lambda}_{E,i}$. For this, T_i was generated as follows: first, for each patient i, the time to death due to the population hazard, T_{P_i}, was obtained from the 2004 American life table, `survexp.us`, stratified by $\boldsymbol{D_i} = (age_i, sex_i)$, as provided by `survival` package in R software (Therneau, 2015). Second, for each patient i, the time to death from cancer, T_{E_i}, was obtained from $\tilde{\lambda}_{E,i}$ modeled 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 modeled 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 in the groups defined by the levels of G vary when the effects of G on the excess mortality vary. More precisely,

the null hypothesis is true when the HR(s) of G equals 1. Conversely, the farther the HR is from 1, the more different are the groups in terms of net survival and the farther we move away from the null hypothesis. When k was equal to 2, the HR of G belonged to $\{0.7; 0.8; 0.9; 1; 1.2; 1.4; 1.6\}$. When k was equal to 3, the HRs of G, (HR_1, HR_2), belonged to $\{(1, 0.7); (1, 1); (1, 1.2); (1, 1.4); (1, 1.6); (0.9, 1.2); (0.8, 1.4); (0.7, 1.6)\}$. In studying the case where T_E and $\boldsymbol{X}$ are independent, we did not introduce effects of *age* and *sex* on the excess mortality (assumption of homogeneity). Conversely, in studying the case where T_E and $\boldsymbol{X}$ are dependent, we set the HR of *sex* to 2 and 3 and chose to assume independence with respect to *age*. The higher the HR of *sex* is, the more different are the distributions of the time to death from cancer between men and women in a given 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 to death was $T_i = \ min(T_{P_i}, T_{E_i})$ whereas each individual's observed time to death was $U_i = \ min(T_{P_i}, T_{E_i}, C_i)$. All subjects still at risk at 5 years were censored.

Moreover, we defined an individual's hypothetical time to death as the minimum of two values: the time to death due to the excess mortality and the time to censoring. According to this time, we obtained another vital status that corresponds to a hypothetical world where cancer would be the only cause of death. Thus, we could compare our test with the usual log-rank test applied on "hypothetical data" and consider the latter test as the gold standard. This is possible only within a simulation framework. We note here that even if cause-specific data were available for our simulations, no direct gold standard for our log-rank-type test could be used in the "real world" because the real world is that of competing risks; thus, still subject to informative censoring.

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

4.2 *Simulation results*

The results obtained without censoring were roughly equivalent to those obtained with 30% censoring, except regarding power. Actually, the power decreased as the censoring level increased. Here, we show only the results related to a 30% censoring level (Tables 1, 2 and 3).

In the comparisons between two groups, the estimation of the type I error of our log-rank-type test was good. In Table 1, at a 5% significance level, the confidence intervals of type I error estimations contain the nominal level of 5% for our test and the usual log-rank test applied on hypothetical data. In comparison with the usual log-rank test, our test performed well in terms of power (Table 1). In the second scenario where the patients are young, the results of both tests were nearly the same. In the first and third scenarios where the patients are old (75% of the patients were more than 65 in the first scenario), our test showed a loss of power.

[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 both tests lost power when the cases were unbalanced (Web Table A).

In the comparisons between three groups, the estimation of the type I error was close to the nominal level of 5% (Table 2). Table 2 shows that, with the first scenario, our test was less powerful than the usual log-rank test with hypothetical data, especially when the three distributions of net survival were not far away from each other ($(HR_1, HR_2) = (1, 0.7)$ or $(0.9, 1.2)$). In the other cases, both tests were equally powerful. In addition, as seen above, our test had a power close to that of the usual log-rank test when the patients were young but showed a loss of power in scenarios 1 and 3 (Web Table B).

[Table 2 about here.]

In the comparisons between the results of the stratified vs. the non-stratified version of our test, in the case of two groups with dependence between T_E and sex, the latter showed less power (Table 3). The farther β_{sex} was from 0, the greater was the 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 type I error was estimated at 2.95% (95% Confidence Interval $(CI) = [2.21; 3.69]$) with the non-stratified version vs. 4.60% ($95\% CI = [3.68; 5.52]$) with the stratified version. However, when HR_{sex} was equal to 2, the corresponding type I errors were 4.80% ($95\% CI = [3.86; 5.74]$) vs. 5.45% ($95\% CI = [4.46; 6.44]$). Thus, the stratified log-rank-type test is better used when the stratum variable has an important impact on net survival.

[Table 3 about here.]

5. Application

For illustration, we used an application of our test to survival data on 10,108 patients with colorectal cancer diagnosed in 1998. These data stem from 17 cancer registries of the Surveillance, Epidemiology, and End Results (SEER) Program (2006) in the United States. 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 the censoring set at five years in still-alive patients. These exclusions left 9,123 patients for analysis. The covariates used were: age at diagnosis, sex, ethnicity (black or white), and cancer stage at diagnosis (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 dataset is described in Web Table C.

We used the American life tables provided by R software `survexp.usr`. These are stratified

by age, sex, ethnicity, and calendar year, from 1998 to 2003. All the analyses were carried out with R (R Core Team, 2014). The code and the .RData files are available upon request. We used our test to compare the net survival distributions between Black and White patients as stratified on cancer stage (I to IV), which is known to have an important effect on cancer net survival. Moreover, in the net survival framework, age is a strong prognostic factor in several types of cancer (Bossard et al., 2007). We considered three age groups: adult patients (20-69 years), old patients (70-79 years), and very old patients ($\geqslant$ 80 years). These stratifications led to 12 strata. Figure 1 shows the impact of age and, most importantly, the impact of stage on net survival. First, the use of the non-stratified version of our test gave a test statistic equal to 19.95 (p-value $= 7.9 \times 10^{-6}$). Second, running the test with stratification on cancer stage, the test statistic was 5.42 (p-value $= 0.0199$). The low proportion of Black patients with low cancer stages (47% in stages I-II vs. 56% for White patients) suggested delayed diagnoses; however, after correction for this, the impact of ethnicity on death from cancer remained significant and higher in Black than in White patients. Third, running our test with stratification on age, the test statistic was 23.62 (p-value $= 1.2 \times 10^{-6}$). Whatever the age group, the differences in net survival between Black and White patients were greater considering age strata than stage strata (data not shown). Finally, running the test stratified on both age and cancer stage, the test statistic was 9.92 (p-value $= 0.0016$). Thus, not stratifying on stage overestimated the differences between the net survival distributions of Black and White patients whereas not stratifying on age underestimated these differences. Stratifying on both age and cancer stage provided "true" differences vs. differences first distorted by group heterogeneity. Using the usual stratified log-rank test on observed survival led to a test statistic equal to 19.5. Thus, the use of net survival instead of observed survival allowed the removal of the confounding effect of age on

observed survival.

[Figure 1 about here.]

6. Discussion

The proposed test compares distributions of net survival as 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 performed well in terms of power despite some loss of power in case of elderly patient: this is because elderly patients have higher expected mortality rates than young patients (i.e., are more at risk of death from other causes). In elderly populations, the loss of information induces a higher variability in the estimates of net survival.

The stratified version is useful when dealing with covariates that exert strong impacts on net survival; that is, when one or several covariates have different distributions in the groups under comparison (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 the covariates under study. The application on real data showed that part of the difference in cancer net survival between Black and White patients is due to differences in cancer stages.

In the test design, we adopted the set of assumptions of Perme et al. (2012). In particular, the model assumed that two latent times defined on the same individual are dependent only via measured covariates (assumption (1.c)). In practice, this may fail in the case of unmeasured covariates (e.g., deprivation). In addition, we made assumptions (2) to prove the asymptotic distribution of the statistic under the null hypothesis. These are reasonable assumptions on the follow-up time because, according to these assumptions, the maximum follow-up T should be small in comparison with T_P given D or T_E.

One possible limitation of our work is that we studied only simulations favorable to our

test. Indeed, the usual log-rank test is optimal under the assumption of proportional hazard rates but performs poorly when this assumption is not met (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). Actually, further studies are needed to adapt our test to one of these procedures. In addition, the formula we propose was developed with a continuous underlying process (without ties). In our application, there were 46% of ties between event times because only survivals in months were available from SEER. We studied the impact of the use of a non-tie-corrected version of our test by using simulations that rounded the survival times to obtain 38%, 45%, and 54% of ties. To compare the percentages of rejection of the null hypothesis, the test was run on the same dataset with and without ties: this led to a maximum difference of 2% (results not shown). Thus, with such percentages of ties in the simulated settings, using a version of the test not corrected for ties had hardly any impact. 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 modeling. We compared our proposed test with the likelihood ratio test from the multivariate excess mortality model using simulated datasets (both presented in section 4.1). We assumed a perfectly defined excess mortality model; i.e., with adjustment on G and sex (whenever necessary) with proportional effects (results not shown). In terms of power, the greatest difference between the percentages of rejection of the null hypothesis by the two tests at the 5% level of significance was 3.15% in favor of the likelihood ratio test. However, with a generated and well-mastered dataset, and with the proposed test, we did not have to deal with a model-building strategy (see e.g., Wynant and Abrahamowicz, 2014). Thus, our non-parametric test should be preferred for its simplicity.

Because our test compares favorably with the usual log-rank test on hypothetical data (as

shown in the simulation study), it may help cancer registries in comparing net survival from cancer between countries or areas. In addition, it may be applied to other chronic diseases in which the concept of net survival may be used. As already performed by Schoenfeld (1981) on the usual log-rank test, it would be interesting to determine the distribution of the proposed test statistic under the alternative hypothesis. By deriving Schoenfeld's formula, we may obtain the sample size that provides the minimal detectable difference. The equivalence between the usual log-rank test and the score test from a Cox model is well known. Another perspective would be to investigate whether such a relationship could exist in the net survival setting. Introducing in a Cox model time-dependent weights that correspond to the ones used in the Pohar-Perme estimator could be an interesting approach.

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.

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 Associate Editor for many important remarks that improved the article. They also thank the CENSUR Working Survival Group for several helpful comments and Jean Iwaz (Hospices Civils de Lyon, France) for the revision of the final version of the manuscript.

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.

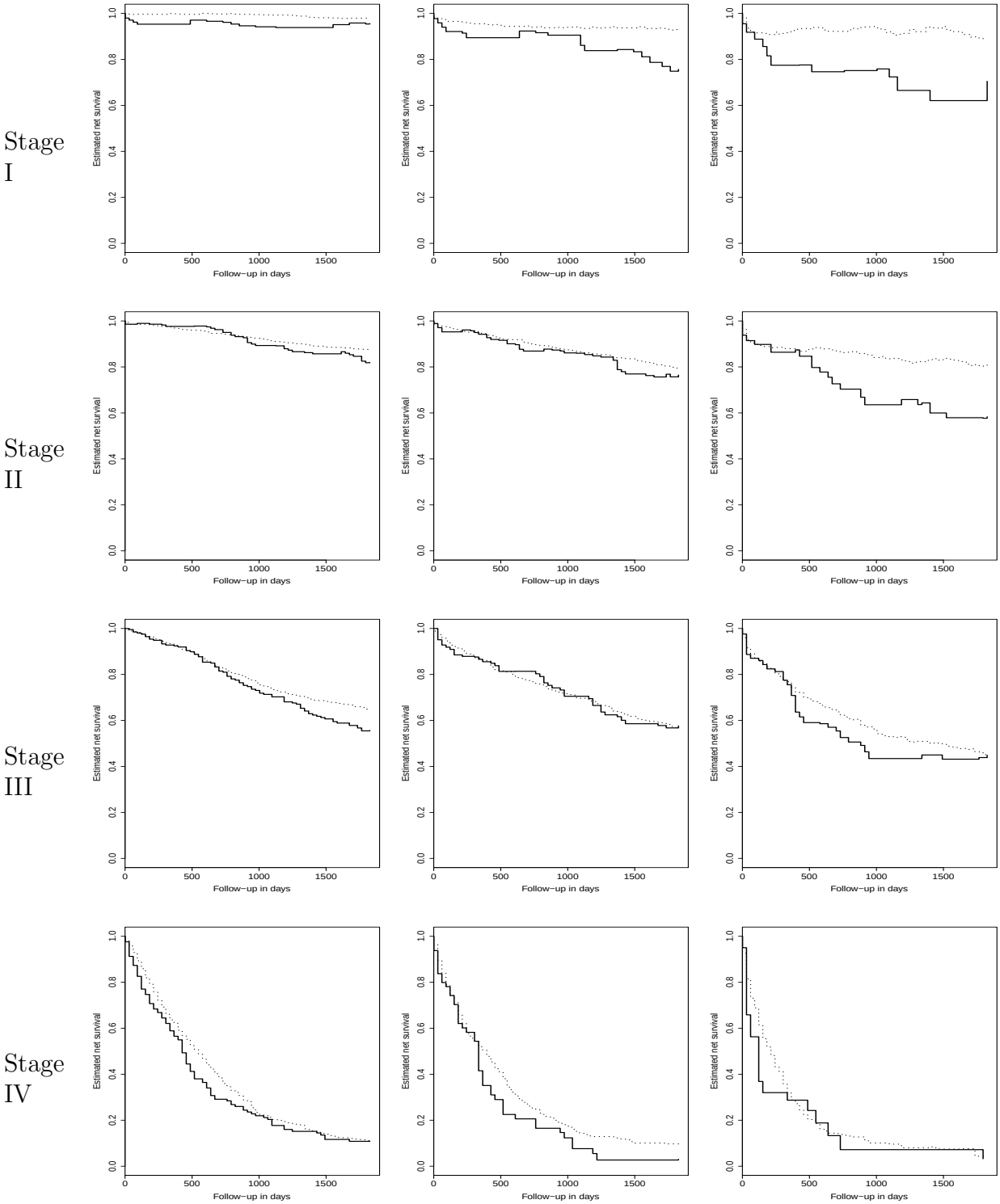

Figure 1. Net survival estimated by the Pohar-Perme estimator according to cancer stage and age groups in White (dotted line) and Black (solid line) patients. When the excess hazard is close to 0 and/or when the number of subjects at risk is low, the variability of the Pohar-Perme estimator can cause non-monotonic net survival curves.

Table 1
Comparison between two groups: percentage of rejection of the null hypothesis at the 5% level of significance with 2000 simulations of 1000 patients. The age distribution is 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 $(95\%\, CI)$			
	Proposed test		Usual log-rank test on hypothetical data	
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 to the excess mortality used in data generation, where G is the covariate that represents the group.
[b]: Balanced cases correspond to cases where the groups are similar in size with $P(G=0) = P(G=1)$.

Table 2

Comparison between three groups: percentage of rejection of the null hypothesis at the 5% level of significance with 2000 simulations of 1000 patients. The age distribution is that of Scenario 1 (25% aged $[40-64]$*, 35% aged* $[65-74]$*, and 40% aged* $[75-85]$*).*

$(HR_1, HR_2)^a$	Percentage of rejection $(95\%\, CI)$			
	Proposed test		Usual log-rank test on hypothetical data	
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 to the excess mortality used in data generation, where G is the covariate that represents the group.

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

Table 3

Comparison of between two groups: percentage of rejection of the null hypothesis at the 5% level of significance with 2000 simulations of 1000 patients when sex has an impact on the excess mortality in the data generation. The age distribution is that of Scenario 1 (25% aged $[40-64]$*, 35% aged* $[65-74]$*, and 40% aged* $[75-85]$*).*

HR^a	Percentage of rejection $(95\% CI)$			
	Proposed test, stratified		Proposed test, non-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 to the excess mortality used in data generation, where G is the covariate that represents the group.