Biostatistics (2016), **0**, 0, *pp.* 1–27
doi:10.1093/biostatistics/output

Performance of two formal tests based on martingales residuals to check the proportional hazard assumption and the functional form of the prognostic factors in flexible parametric excess hazard models

CORALINE DANIELI[1−3,*], NADINE BOSSARD[1−2], LAURENT ROCHE[1−2], AURELIEN BELOT[4], ZOE UHRY[5,1−2], HADRIEN CHARVAT[6], LAURENT REMONTET[1−2], and the CENSUR Working Survival Group

[1]*Hospices Civils de Lyon, Service de Biostatistique-Bioinformatique, F-69003, Lyon, France ;*

[2]*CNRS, UMR5558, Laboratoire de Biométrie et Biologie Evolutive, Equipe Biotatistique-Santé, Université Lyon 1, F-69100, Villeurbanne, France ;*

[3]*McGill University Health Center, Department of Epidemiology, Biostatistics and Occupational Health, H3A 1A2, Montreal, QC, Canada ;*

[4]*Cancer Research UK Cancer Survival Group, Faculty of Epidemiology and Population Health, Department of NonCommunicable Disease Epidemiology, London School of Hygiene and Tropical Medicine, London WC1E 7HT, U.K ;*

[5] *Department of Chronic Diseases and Injuries, The French Public Health Agency, Saint Maurice, France ;*

[6] *Epidemiology and Prevention Group, Research Center for Cancer Prevention and Screening, National Cancer Center, Tokyo, Japan ;*

coraline.danieli@rimuhc.ca

∗. To whom correspondence should be addressed.

Summary

Net survival, the one that would be observed if the disease under study were the only cause of death, is an important, useful, and increasingly used indicator in public health, especially in population-based studies. Estimates of net survival and effects of prognostic factor can be obtained by excess hazard regression modeling. Whereas various diagnostic tools were developed for overall survival analysis, few methods are available to check the assumptions of excess hazard models. We propose here two formal tests to check the proportional hazard assumption and the validity of the functional form of the covariate effects in the context of flexible parametric excess hazard modeling. These tests were adapted from martingale-residual-based tests for parametric modeling of overall survival to allow adding to the model a necessary element for net survival analysis : the population mortality hazard. We studied the size and the power of these tests through an extensive simulation study based on complex but realistic data. The new tests showed sizes close to the nominal values and satisfactory powers. The power of the proportionality test was similar or greater than that of other tests already available in the field of net survival. We illustrate the use of these tests with real data from French cancer registries.

Key words: Checking model assumptions ; Flexible parametric excess hazard model ; Functional form ; Martingale residuals ; Net survival ; Proportional hazard assumption.

1. Introduction

Net survival is defined as the survival that would be observed if the disease under study were the only cause of death ; that is, the survival that would be observed after removing deaths from other causes. Net survival is a relevant indicator now widely used in public health studies because

it allows comparisons between countries (Allemani *and others*, 2015; De Angelis *and others*, 2015) and analyses of time trends. Indeed, once removed, deaths from other causes (often varying with time and place) disturb no more the above mentioned comparisons.

Point estimations of net survival can be obtained with the non-parametric Pohar-Perme estimator (Pohar-Perme *and others*, 2012). However, this estimator quantifies the prognostic effects of the covariates through stratification only, which is rapidly impracticable ; an efficient analysis requires an excess hazard model. Since the seminal paper of Estève (Esteve *and others*, 1990), several extensions of this model have been developed. The main improvements were the introduction of flexible parametric continuous functions (such as splines) to model the baseline excess mortality hazard, the non-proportional effects, and the functional forms of the covariates (Giorgi *and others*, 2003; Dickman *and others*, 2004; Lambert *and others*, 2005; Nelson *and others*, 2007; Remontet *and others*, 2007; Crowther and Lambert, 2014).

In comparison with non-parametric or semi-parametric models, these flexible parametric models offer numerous key advantages (Lin and Spiekerman, 1996). In particular, parametric models allow : i) simple efficient inference from full likelihood ; ii) dealing adequately with non-linear and non-proportional effects ; and iii) reporting relative but useful absolute effect measures (King *and others*, 2012) such as the shape of the dynamics of the excess rate, which is of great clinical interest (Corm *and others*, 2011). These flexible parametric excess hazard models have been regularly used in cancer epidemiology to provide estimations of net survival (Bossard *and others*, 2007; Corm *and others*, 2011) or derived indicators such as the crude probability of death (Charvat *and others*, 2013) or the number of years of life lost due to cancer (Andersson *and others*, 2013).

Though the excess hazard model is widely used, few methods are available to check its assumptions. To our knowledge, there are only two methods, both based on the analysis of the residuals.

The first method was proposed by Stare *and others* (2005) to check the assumption of pro-

portional hazards (PHs). First, they defined residuals for excess hazard models by imitating the form of the residuals as proposed by Shoenfeld (Schoenfeld, 1982). The latter residuals have two main properties : they arise directly from the score function of the partial likelihood of the Cox model and they are the differences between the observed covariate values at the time of death and their predictions according to the model. In contrast, the Shoenfeld-like residuals defined by Stare have only the second property ; however, they can be used to check the PH assumption within the framework of the excess hazard model (Stare *and others*, 2005). To this end, a test statistic is obtained by forming a cumulative sum process of standardized Shoenfeld-like residuals, which can be approximated by a Brownian bridge process. The distribution of this statistic under the null hypothesis is then obtained using standard results from the Brownian-bridge theory.

In agreement with the ideas of Lin *and others* (1993) on overall survival, the second method was proposed by Cortese and Scheike (2008) for semi-parametric excess hazard models. This method allows checking the PH assumption and the functional form (FF) of the covariates by cumulating martingale residuals over the time of follow-up and over the covariate values, respectively. The distribution under the null hypothesis of these processes can be approximated by simulating corresponding Gaussian processes.

Despite the availability of these two methods, there is a real need for diagnostic tools specific to parametric models. Indeed, the semi-parametric Cortese approach does not fit within this parametric paradigm and, contrary to the choice made for the Stare test, it is potentially interesting to privilege residuals that arise naturally from the score function and develop new tests to assess proportionality using the useful properties of the score process. Furthermore, no tool is yet available in the excess hazard parametric setting for assumptions other than the PH (e.g., FF assumption).

In this paper, we propose a test to check the PH assumption and another test to check the FF of a covariate in the framework of flexible parametric excess hazard model. To achieve this goal,

we used the general approach proposed by Lin and Spiekerman (1996) for parametric models in the overall survival setting and adapted it to the net survival setting. This led to two formal tests based on similar stochastic processes.

Herein, Section 2 includes a brief presentation of Lin's approach and its adaptation to the context of the excess hazard model. Section 3 shows a simulation study to determine the size and the power of the PH test and Section 4 determines the same for the FF test. Section 5 shows the assessment of the performances of these tests when the other assumptions are misspecified. We illustrate the use of these tests on real data in Section 6 and discuss the results in the last section.

2. Methods

2.1 *Lin's approaches to check the usual assumptions in parametric survival analysis*

Lin *and others* (1993) and Lin and Spiekerman (1996) proposed a general approach to check the assumptions of survival regression models using the martingale theoretical framework. First, we present briefly the methodology they developed, emphasizing the intuitive aspect of the reasoning, and then show the way this methodology can be adapted to check the main assumptions in the flexible parametric excess hazard framework, knowing that the intuitive aspect of the reasoning is the same. In this section, we adopted the notation used by Lin and Spiekerman (1996).

Let $(X_i, \Delta_i, \mathbf{Z_i})$ be the data available on patient i, where X_i is the time of observation (minimum between the time to death T_i and the censoring time C_i), Δ_i the death indicator, and $\mathbf{Z_i} = \{Z_{i1}, ..., Z_{ip}\}'$ a p-vector of covariates. Let further the counting process related to patient i be $N_i(u) = I(X_i \leqslant u, \Delta_i = 1)$ and the at-risk process be $Y_i(u) = I(X_i \geqslant u)$ where $I()$ is the indicator function. The martingale residuals compare the status of each subject at time t ($N_i(t) = 0$ or $N_i(t) = 1$) with its estimated expected value at time t. Let $\lambda_i(t|\mathbf{Z_i}, \boldsymbol{\theta}) = \lambda_0(t, \boldsymbol{\chi}) \exp(\boldsymbol{\beta}'\mathbf{Z_i})$

and $\Lambda_i\left(t|\mathbf{Z_i},\boldsymbol{\theta}\right)=\Lambda_0\left(t,\boldsymbol{\chi}\right)\exp\left(\boldsymbol{\beta}'\mathbf{Z_i}\right)$ denote, respectively, the instantaneous hazard function and the cumulative hazard function associated with individual i, where $\boldsymbol{\theta}=\left(\boldsymbol{\beta}',\boldsymbol{\chi}'\right)'$ is an unknown vector of parameters to be estimated and $\boldsymbol{\beta}$ and $\boldsymbol{\chi}$, respectively, the parameters for covariates $\mathbf{Z}$ and baseline hazard. The martingale residuals can be then written :

$$\widehat{M}_i\left(t\right)=N_i\left(t\right)-\int_0^t Y_i\left(\nu\right)\lambda_i\left(\nu|\mathbf{Z_i},\widehat{\boldsymbol{\theta}}\right)d\nu \tag{2.1}$$

Martingale residuals and their transforms were first developed in the non-parametric model setting and then in the Cox model setting where they were particularly used as graphical methods to detect departures from modeling assumptions. For instance, to test the PH assumption, Schoenfeld (1982) introduced residuals that allow detecting graphically whether the effect of the covariate of interest depends on time. Later, Barlow and Prentice (1988) generalized these residuals and Grambsch and Therneau (1994) introduced a standardized version of these residuals. To detect the FF of a covariate, Therneau *and others* (1990) showed that the plot of the martingale residuals over the covariate values can be used to obtain an approximation of the true functional form.

However, the interpretation of the curves may be difficult because one cannot know whether the trend shown by the residual plot is due to a real misspecification or to natural variations. One solution would be to work with formal tests based on quantities of known distributions under the null hypothesis. To this end, Lin *and others* (1993) and Lin and Spiekerman (1996) proposed a general framework to develop several techniques for examining various modeling assumptions. Each technique derives from a cumulative sum of martingale-based residuals by considering two classes of multi-parameter stochastic processes. In this article, we will focus only on the process $\mathbf{W_z}$, defined by :

$$\mathbf{W_z}\left(t,\mathbf{z}\right)=n^{-1/2}\sum_{i=1}^{n}f\left(\mathbf{Z_i}\right)I\left(\mathbf{Z_i}\leqslant\mathbf{z}\right)\widehat{M}_i\left(t\right) \tag{2.2}$$

where $f\left(.\right)$ is a known function, $\mathbf{z}=\left\{z_1,...,z_p\right\}'\in\mathbb{R}^p$, and where $I\left(\mathbf{Z_i}\leqslant\mathbf{z}\right)=I\left(Z_{1i}\leqslant z_1,...,Z_{pi}\leqslant z_p\right)$ equals to 1 when each component of $\mathbf{Z_i}$ is smaller than its corresponding component of $\mathbf{z}$.

This process is a cumulative sum of the martingale residuals "weighted" by a function f. Lin and Spiekerman (1996) showed the way the null distribution of $\mathbf{W_z}(t, \mathbf{z})$ can be approximated by simulating appropriate zero-mean Gaussian processes in the general case (that is, whatever f, t, or $\mathbf{z}$), paving the way for tests of more specific hypotheses. In particular, Lin and Spiekerman (1996) examined two special cases of $\mathbf{W_z}$ that allowed assessing the proportionality (PH) and the functional form (FF) of a single covariate.

Regarding the PH assumption, Lin and Spiekerman (1996) considered the identity f function and $\mathbf{z} = (\infty, ..., \infty)$; $\mathbf{W_z}(t, \mathbf{z})$ is then expressed as :

$$\mathbf{W_z}(t, \mathbf{z}) = n^{-1/2} \sum_{i=1}^{n} Z_i \widehat{M_i}(t) \tag{2.3}$$

which corresponds also to the score process evaluated at time t when the covariates that form $\mathbf{Z_i}$ are fixed over time, $U\left(\widehat{\boldsymbol{\theta}}, t\right) = n^{-1/2} \sum_{i=1}^{n} Z_i \widehat{M_i}(t)$. This process is informative about the PH assumption and can be used to check it (as done in the PHREG procedure of SAS software, version 9.3 (SAS Institute Inc., Cary, NC)) : the score process U_k, evaluated at $\widehat{\theta}$, is expected to be informative about the PH misspecification of the k^{th} component of $\mathbf{Z_i}$. Indeed, $U_k(\theta, t)$ has the form the usual score would have if the data were artificially censored at time t. Thus, $U_k(\theta, t)$ equals zero for value $\theta = \theta_t$ of the parameter that would be estimated on the basis of data censored at time t. However, when the assumption of PHs is met, the parameter estimate is the "same" whatever the censoring time ; i.e., $\widehat{\theta_t}$ is close to $\widehat{\theta}$ whatever t. Therefore, under the null hypothesis "the proportional hazard assumption is correct", the process $U_k\left(\widehat{\theta}, t\right)$ is expected to remain close to zero over time.

Likewise, regarding the functional form of the j^{th} covariate component, Lin and Spiekerman (1996) considered the following process :

$$W^{(j)}(x) = n^{-1/2} \sum_{i=1}^{n} I(Z_{ji} \leqslant x) \widehat{M_i} \tag{2.4}$$

This corresponds to a special case of process $\mathbf{W_z}(t, \mathbf{z})$ taking f as the constant function

$f(.) = 1$, $t = \infty$, and $\mathbf{z} = (\infty, ..., z_j, ..., \infty)$. When cumulated over z_j, (z_j being the observed values of covariate Z_j), this process is informative about the specification of the FF of the covariate and can be used to check it. Indeed, Therneau *and others* (1990) have shown that the martingale residuals at the maximum time of follow-up are informative about the functional form of the continuous covariates of the model. Therefore, their cumulative sum over the covariate values is also useful to detect a misspecification : when the functional form is correct, the martingale residuals are close to zero and their cumulative sum should not present structured trends but fluctuate around zero.

2.2 *The proposed test to check the PH assumption of a parametric excess hazard model*

We now present the way the above methodology can be adapted to the parametric excess hazard framework.

We consider the following excess hazard model (Esteve *and others*, 1990; Remontet *and others*, 2007) :

$$\lambda_{ov}(t|\mathbf{Z}, \boldsymbol{\theta}) = \lambda_E(t|\mathbf{Z}, \boldsymbol{\theta}) + \lambda_P(a+t, \widetilde{\mathbf{z}}) = \lambda_0(t, \boldsymbol{\chi}) \exp(\boldsymbol{\beta} Z) + \lambda_P(a+t, \widetilde{\mathbf{z}}) \tag{2.5}$$

where λ_{ov} represents the overall mortality hazard, λ_E the excess mortality hazard (defined by $\boldsymbol{\theta} = \left(\boldsymbol{\beta}', \boldsymbol{\chi}'\right)'$ the unknown vector of parameters to be estimated, and $\mathbf{Z}$ the set of prognostic covariates supposed fixed in time), $\widetilde{\mathbf{z}}$ the vector of the demographic covariates (other than age at time of death or censoring $(a+t)$) that define the population mortality hazard λ_P and a the age at diagnosis. In this section, we will focus on the score process of the k^{th} prognostic covariate Z_k.

2.2.1 *The score process.*

Let us first consider the score process in this setting. The log likelihood for individual i can be written (up to a constant) :

$$L_i(\boldsymbol{\theta}, X_i) = -\int_0^{X_i} \lambda_E(u|\mathbf{Z_i}, \boldsymbol{\theta})\, du + N_i(X_i) \log\left[\lambda_E(X_i|\mathbf{Z_i}, \boldsymbol{\theta}) + \lambda_P(a_i + X_i, \widetilde{\mathbf{z}_i})\right] \tag{2.6}$$

The k^{th} component $U_k(\boldsymbol{\theta}, X_i)$ of the score function can be written :

$$U_k(\boldsymbol{\theta}, X_i) = \sum_i \frac{\partial}{\partial \beta_k} L_i(\boldsymbol{\theta}, X_i) = \sum_i \left[- \int_0^{X_i} Z_{ik} \lambda_E(u|\mathbf{Z_i}, \boldsymbol{\theta})\, du + N_i(X_i) \frac{Z_{ik} \lambda_E(u|\mathbf{Z_i}, \boldsymbol{\theta})}{\lambda_E(u|\mathbf{Z_i}, \boldsymbol{\theta}) + \lambda_P(a_i + X_i, \widetilde{\mathbf{z}_i})} \right] \tag{2.7}$$

Using a counting process formulation and noting τ the maximum time of follow-up, $U_k(\boldsymbol{\theta}, \tau)$ may be written :

$$U_k(\boldsymbol{\theta}, \tau) = \sum_i Z_{ik} \left[\int_0^{\tau} -Y_i(u) \lambda_E(u|\mathbf{Z_i}, \boldsymbol{\theta})\, du + \int_0^{\tau} \frac{Y_i(u) \lambda_E(u|\mathbf{Z_i}, \boldsymbol{\theta})}{\lambda_E(u|\mathbf{Z_i}, \boldsymbol{\theta}) + \lambda_P(a_i + u, \widetilde{\mathbf{z}_i})} dN_i(u) \right] \tag{2.8}$$

which equals zero for the maximum likelihood (ML) estimate of $\boldsymbol{\theta}$. Replacing τ by t and $\boldsymbol{\theta}$ by its ML estimate, we obtain :

$$U_k\left(\widehat{\boldsymbol{\theta}}, t\right) = \sum_i \left[Z_{ik} \int_0^t \frac{\lambda_E\left(u|\mathbf{Z_i}, \widehat{\boldsymbol{\theta}}\right)}{\lambda_E\left(u|\mathbf{Z_i}, \widehat{\boldsymbol{\theta}}\right) + \lambda_P(a_i + u, \widetilde{\mathbf{z}_i})} d\widehat{M}_i(u) \right] \tag{2.9}$$

a cumulative sum of weighted martingale residuals, which is known as the score process with $\widehat{M}_i(t) = N_i(t) - \int_0^t Y_i(u) \left[\lambda_E\left(u|\mathbf{Z_i}, \widehat{\boldsymbol{\theta}}\right) + \lambda_P(a_i + u, \widetilde{\mathbf{z}_i}) \right] du$. Note that $d\widehat{M}_i$ is weighted by the product of the covariate value of patient i and the probability that death is caused by the disease under study, which is specific to the net survival setting.

2.2.2 *Expression of the score process as a special case of a new class of stochastic processes.* By analogy with the works of Lin *and others* (1993) and Lin and Spiekerman (1996) in the overall survival setting, process U_k, evaluated at $\widehat{\boldsymbol{\theta}}$, can no longer be written in terms of a particular process $\mathbf{W_z}(t, \mathbf{z})$ as in (2.3) because the integrand depends on u. This led us to consider a new class of stochastic processes :

$$\mathbf{W_z^{(2)}}(t, \mathbf{z}) = n^{-1/2} \sum_{i=1}^{n} \left[\int_0^t f(u|\mathbf{Z_i}, \boldsymbol{\theta})\, I(\mathbf{Z_i} \leqslant \mathbf{z})\, dM_i(u) \right] \tag{2.10}$$

which could be approximated by using its Taylor expansion and by approximating the unknown limiting distribution of the martingale process by a Gaussian process having the same characteristics. Note that when $f(u|\mathbf{Z_i}, \boldsymbol{\theta})$ does not depend on u and $\boldsymbol{\theta}$, $\mathbf{W_z^{(2)}}$ can be simplified to a

process of the form $\mathbf{W_z}$ as in (2.3). We can now write the score function in (2.9) evaluated at $\widehat{\boldsymbol{\theta}}$ as a special case of $\mathbf{W_z^{(2)}}$ with $f\left(u|\mathbf{x},\boldsymbol{\theta}\right) = \frac{x_k \lambda_E(u|\mathbf{x},\boldsymbol{\theta})}{\lambda_E(u|\mathbf{x},\boldsymbol{\theta})+\lambda_P(a_i+u,\widetilde{\mathbf{z}_\mathbf{i}})}$ and $\mathbf{z} = (\infty, ..., \infty)$.

2.2.3 *Approximation of the limiting distribution of the score process under the null hypothesis and test statistic.* Let us now consider the limiting distribution of the special case of $\mathbf{W_z^{(2)}}$ defined above. The population hazard λ_p being considered as fixed and known in the excess hazard model, f remains a predictable process. Thus, heuristically, the different steps of the demonstration made in Section 4.1 by Lin and Spiekerman (1996) in the overall parametric survival setting should remain valid in the net survival setting. Under the null hypothesis, $U_k\left(\widehat{\boldsymbol{\theta}}, t\right)$ fluctuates around zero and its limiting distribution can be approximated using a Taylor expansion and Monte-Carlo simulations of Gaussian processes (as shown in section 3 of Supplementary Material). Practically, it leads to form the following process $\widehat{\mathbf{W}}_\mathbf{z}^{(\mathbf{2})}$:

$$\widehat{\mathbf{W}}_\mathbf{z}^{(\mathbf{2})}(t) = n^{-1/2}\left(\mathbf{D_1}\left(\widehat{\boldsymbol{\theta}}, t\right) - \mathbf{I}\left(\widehat{\boldsymbol{\theta}}, t\right)\mathbf{I}\left(\widehat{\boldsymbol{\theta}}, \tau\right)^{-1}\mathbf{D_1}\left(\widehat{\boldsymbol{\theta}}, \tau\right)\right) \tag{2.11}$$

with $\mathbf{D_1}\left(\widehat{\boldsymbol{\theta}}, t\right) = \sum_i \int_0^t f\left(u|\mathbf{Z_i}, \widehat{\boldsymbol{\theta}}\right) dN_i(u)\, G_i$, $\mathbf{I}\left(\widehat{\boldsymbol{\theta}}, t\right)$ and $\mathbf{I}\left(\widehat{\boldsymbol{\theta}}, \tau\right)$ being the negative of the second derivative of the log-likelihood at time t and time τ, respectively ($\mathbf{I}\left(\widehat{\boldsymbol{\theta}}, \tau\right)$ is thus the information matrix), the limiting distribution of $n^{-1/2}U_k\left(\widehat{\boldsymbol{\theta}}, t\right)$ is obtained considering the k^{th} line of $\widehat{\mathbf{W}}_\mathbf{z}^{(\mathbf{2})}(t)$.

As indicated by Lin and Spiekerman (1996), it is natural to consider the test statistic $\sup_t \left|n^{-1/2}U_k\left(\widehat{\boldsymbol{\theta}}, t\right)\right|$: the p-value of the new PH test corresponds to the proportion of cases in which the absolute maximum of the simulated Gaussian processes is higher than the test statistic. In other words, setting the nominal value of the probability of type 1 error to 5% and simulating 1000 Gaussian processes, the PH assumption will be rejected if less than 50 (over 1000) absolute maximums of simulated Gaussian processes are higher than the absolute maximum score process value. Note that it would be also informative to plot the observed score process over time with the simulated Gaussian processes in order to see the pattern of $U_k\left(\widehat{\boldsymbol{\theta}}, t\right)$ over time. Indeed, this would show

the range of time values at which the score process moves away from the Gaussian processes ; that is, the time points at which the PH assumption is no more valid.

2.3 *The proposed test to check the FF of a covariate of a parametric excess hazard model*

Concerning the functional form, the results of Lin and Spiekerman (1996) can be directly applied by considering the following process for the j^{th} covariate component : $W^{(j)}(x) = n^{-1/2} \sum_{i=1}^{n} I(Z_{ji} \leqslant x) \widehat{M}_i$, similarly to the process considered by Lin and Spiekerman (1996) to check the functional form of covariates (see Section 2.1, formula (2.4)). Hence, Lin's framework applies and, under the null hypothesis H_0 (i.e., "the functional form is correct"), the limiting distribution of $W^{(j)}(x)$ can be approximated by simulating the following process $\widehat{W}^{(j)}(x)$:

$$n^{-1/2}\widehat{W}^{(j)}(x) \approx n^{-1/2}\left(P_1(x) - \widehat{\mathbf{J}}(x)\,\mathbf{I}\left(\widehat{\boldsymbol{\theta}},\tau\right)^{-1}\mathbf{D_1}\left(\widehat{\boldsymbol{\theta}},\tau\right)\right) \tag{2.12}$$

with $P_1(x) = \sum_i \int_0^\tau I(Z_{ji} \leqslant x)\, dN_i(u)\, G_i$, $\mathbf{D_1}\left(\widehat{\boldsymbol{\theta}},t\right) = \sum_i \int_0^t \frac{Z_i \lambda_E\left(u|\mathbf{Z_i},\widehat{\boldsymbol{\theta}}\right)}{\lambda_E\left(u|\mathbf{Z_i},\widehat{\boldsymbol{\theta}}\right)+\lambda_P(a_i+u,\widetilde{\mathbf{z_i}})} dN_i(u)\, G_i$, $\widehat{\mathbf{J}}(x) = \sum_i \int_0^\tau I(\mathbf{Z_i} \leqslant x)\, \mathbf{Z_i} \exp\left(\widehat{\boldsymbol{\beta}}'\mathbf{Z_i}\right) d\widehat{\Lambda}_0(u)$ and $\mathbf{I}\left(\widehat{\boldsymbol{\beta}},\tau\right)$ being the negative of the information matrix. As in section 2.2, the p-value can be obtained using the test statistic $\sup_y \left|n^{-1/2}W^{(j)}(x)\right|$ which may be compared with the supremum of the 1000 simulated Gaussian processes.

3. Simulation study to assess the performance of the test of the PH assumption

We performed a simulation study to determine the size and the power of the new test that checks the assumption of PHs (called thereafter "the new PH test"). The whole code to reproduce the results of the simulation study (simulating the data, fitting the model and running the test) are now available on the GitHub repository (cf the readme.pdf for explanations of the structure of the programs) `https://github.com/danielico/Biostatistics_Paper/tree/master/Review`.

3.1 *Simulation design*

The performance of this test was estimated according to the sample size (n = 500, 1000, or 2000 patients) and according to the strength of the non-proportional hazard (NPH) effect for which we defined three levels : low, medium, and high (see Supplementary Material - http ://www.biostatistics.oxfordjournals.org). The size was evaluated using 2000 simulated datasets and the power calculated using 1000 datasets. We supposed that *age* is the covariate of interest whose PH assumption is under study. For each dataset, we simulated a cohort whose distribution of covariate *age* corresponds to that of French colon cancer patients, with year of diagnosis between 1990 and 2010 and end of follow-up in 2013. The time to death from cancer, T_E, was generated from a specific procedure detailed in section 1 of Supplementary Material, whereas the time to death from other causes, T_P, and the final observation time T were obtained by the procedure described in Danieli *and others* (2012).

3.2 *Implementation of the tests on the simulated datasets*

The new PH test relied on martingale residuals obtained from a specific model : considering each simulated dataset, we fit an excess hazard model (up to 10 years of follow-up) with a quadratic regression spline with two knots (at 1 and 5 years) for the baseline excess hazard and a linear and proportional effect of *age* (Table 1 in Supplementary Material). The maximum likelihood estimates were obtained from a homemade procedure on the basis of Cavalieri-Simpson integral approximation and a Newton-Raphson algorithm (Remontet *and others*, 2007) ; 1000 Gaussian processes were simulated to obtain the p-values of the new PH test.

We evaluated also the performance of the three tests developed by Stare *and others* (2005) to calibrate the new PH test with the only tool currently available to test formally the PH hypothesis in the net survival setting. The first test is based on the maximum value of the bridge process (called thereafter T1). The second test is a weighted version of the first one, the weights of the

residuals being set proportional to the number at risk at each event time (called thereafter T2). The third test is based on the variance of the bridge process and used the Cramér-von Mises statistic (called thereafter T3). These tests have been tailored for different departures from the null hypothesis. They are available through the function *rsadd* in R package `relsurv` 2.09 (Pohar and Stare, 2006). This function proposes several options : i) with method "max.lik", Stare tests are performed adjusting a piecewise-constant baseline excess hazard and the parameter estimates are the maximum likelihood estimates (these tests are thereafter called "Stare T1 ML", "Stare T2 ML", and "Stare T3 ML"); ii) with method "EM", Stare tests are performed adjusting an unspecified baseline excess hazard and the parameter estimates are obtained using an EM algorithm (these tests are thereafter called "Stare T1 EM", "Stare T2 EM", and "Stare T3 EM").

3.3 *Results*

Table 1 presents the sizes of the new PH test and those of Stare tests as well as their powers. When the PH assumption was met, the sizes of all tests were close to the nominal value (5%); however, the sizes of Stare tests based on ML method were slightly lower. As expected, when the PH assumption was not met, the power of each test increased together with the sample size and the strength of the NPH effect. The new PH test was found more powerful than T1 and T2 Stare tests, whatever the sample size and the NPH effect, whereas its power was similar to that of T3.

(Table 1 about here)

4. Simulation study to assess the performance of the test of the FF of a covariate

This simulation study is presented in section 2 of Supplementary Material.

Table 1 shows the size and the power of the new FF test. When the functional form of the

covariate was correctly specified, the size was close to the nominal value. When the functional form of the covariate was incorrectly specified, the power increased, as expected, with the sample size and the strength of the NLIN effect.

5. Simulation study to assess the performances of both new tests when the other assumptions than the one of interest are misspecified or over-parameterized

In this section, we assessed the amount of inflation of the size of both tests in case of misspecification of the other assumption and the loss of power in case of over-parameterization. More precisely, to evaluate the inflation of the size of the new PH test in case of misspecification concerning the FF of the covariate, we fitted a linear and proportional (LIN-PH) model on the non-linear and proportional (NLIN-PH) simulated data (first row of Table 2) : this allowed checking whether the PH test detects wrongly non-proportionality due to the presence of non-linearity. To evaluate the loss of power of the new PH test in case of over-parameterization of the FF in the fitted model, we fitted a non-linear and proportional model (NLIN-PH) on the linear and non-proportional (LIN-NPH) simulated data (second row of Table 2) : this allowed checking whether a part of the non-proportionality of the data is "captured" by the non-linear part of the fitted model leading to a loss of power of the PH test. In the same way, to evaluate the inflation of the size of the new FF test in case of misspecification concerning the PH effect of the covariate, we fitted a LIN-PH model on the LIN-NPH simulated data (third row of Table 2) : this allowed checking whether the FF test detects wrongly non-linearity due to the presence of non-proportionality. Then, to evaluate the loss of power of the new FF test in case of over-parameterization of the PH effect of the covariate in the fitted model, we fitted a LIN-NPH model on the NLIN-PH simulated data (fourth row of Table 2) : this allowed checking whether a part of the non-linearity of the data is "captured" by the non-proportional part of the fitted model leading to a loss of power of FF test.

For those scenarios summarized in Table 2, we simulated data with low effects of $g(age)$ and $\beta(t)$.

(Table 2 about here)

Table 3 presents the sizes and the powers of the new PH and FF tests in case of misspecification or over-parameterization of the other assumption than the one of interest.

(Table 3 about here)

In both tests, in case of misspecification, the size was higher than the nominal value and increased together with the sample size. In case of over-parameterization, the loss of power ranged between 9% and 16% for the new PH test and between 5% and 7% for the new FF test.

6. Application to real data

We now illustrate the new PH and FF tests using cancer cases registered between 1989 and 1997 by the French cancer registries. A previous study (Bossard *and others*, 2007) on these data suggested that the effect of age is linear and proportional in salivary gland cancer (377 cases), non-linear and non-proportional in breast cancer (30923 cases), and non-linear and proportional in non-Hodgkin lymphoma (6375 cases). The new tests were implemented using martingale residuals obtained from an excess hazard model in which the baseline hazard was modeled with a quadratic spline up to 10 years of follow-up (with two knots at 1 and 5 years) and the age effect was "linear and proportional" or "non-linear and proportional" according to cases. We simulated 1000 Gaussian processes.

Concerning salivary glands cancer, the new PH test and the new FF test provided a p-value of 0.58 and 0.83, respectively. Thus, these formal tests support previous results that a proportional and linear effect of age is adequate. The plot of the score process over time and of the cumulative martingale residuals over age are shown in Figures 1a and 1b, respectively.

Concerning breast cancer, the new PH test and the new FF test provided a p-values $\ll 0.05$. These formal tests supported also previous results suggested by the previous study. The observed processes are plotted in Figures 1c and 1d, respectively.

Concerning non-Hodgkin lymphoma, the new PH test of a model that assumes a linear effect provided a p-value $\ll 0.05$, which means that the assumption of proportional effect was rejected (Figure 1e). This result differs from the one of the previous study. It could be due to the misspecification of the functional form of age. Therefore, after assuming a non-linear effect in the model, we expected that the new PH test would provide a p-value > 0.05. However, the PH hypothesis was rejected as well (Figure 1f). This example illustrates the fact that the results can differ between our tests and the LRT results obtained in a previous work on the same data (Bossard *and others*, 2007).

(Figure 1 about here)

7. Discussion

In the present work, we propose two tests to check the main assumptions of a parametric excess hazard model ; that is, the proportional hazard assumption and the functional form of the covariates of interest. They are an extension of tests proposed by Lin and Spiekerman (1996). Our tests and those of Lin are equivalent when there is no competing risk, that is, when the population mortality hazard is set to zero. Both tests are developed in the same theoretical setting.

The performance of both tests was assessed from simulated data with realistic and complex distributions. Indeed, using a specific algorithm, we were able to simulate data in which the baseline hazard, the non-proportional, and the non-linear terms were modeled using smooth functions whose theoretical parameters were obtained by adjusting an excess hazard model on real French cancer registry data. Furthermore, because in real life the true model is not known, we used two different models to simulate the data and analyze the simulated data : a fractional

polynomial for simulation and regression splines for analysis.

The results concerning the PH test were satisfactory, especially with medium and high non-proportional effects : the power was over 90% when the sample size was higher or equal to 1000 patients, size generally seen in population-based survival studies.

The power of the new PH test was higher than that of T1 or T2 Stare tests (based on the maximum value of an approximate Brownian Bridge) but similar to that of T3 test (based on the variance). This reminds us that the power of a test depends on the departure from the null hypothesis and on the sensitivity of the test in detecting this particular form of discrepancy. Indeed, as mentioned by Stare *and others* (2005), contrarily to the maximum tests, T3 test exploits the whole path of the bridge process and is more powerful than T1 or T2 test when the path of $\beta(t)$ is non-monotonic, which is the case in our analysis (Figure 1 in Supplementary Material).

However, the main goal of the present paper was not to make an exhaustive comparison of all existing tests under different alternatives, but to validate the new tests specifically designed for flexible excess hazard models.

The results concerning the new FF test are also encouraging : the size was correct and the power satisfactory. This new FF test is therefore a very useful tool to test linearity but also to test more complex forms. For example, when splines are used to model non-linearity, it allows checking the quality of the fit provided by this kind of parametric functions.

In contrast with the Stare approach, one interesting feature of the new tests is that their values do not depend on the population expected mortality of censored observations. The population expected values are essentially used to evaluate the probability that a death is caused by the disease under study.

Here, we were not able to report comparisons with tests proposed by Cortese and Scheike (2008), which are available in the R package `timereg`, because we could not obtain coherent

results from this program (the sizes were equal to zero whatever the sample size and the powers were very low compared with the ones presented in this paper).

In practice, the users of the new PH and FF tests have to be aware that these tests suppose that other assumptions are true ; i.e., the null hypothesis means that the whole model is correctly specified. This is why the results with real data have to be interpreted with caution ; indeed, the misspecification of one aspect may impact the test of another aspect. For example, the misspecification of the FF of a covariate can impact the score process behavior when checking the PH assumption. However, the results in Table 3 show that if we over-parameterize the PH effect (NPH instead of PH), the loss of power of the FF test will not be large ; likewise, if we over-parameterize the FF effect, the loss of power of the PH test will be small. Therefore, as already suggested by Abrahamowicz and Mackenzie (2007), it is desirable to carry out the PH test specifying systematically a non-linear effect for the covariate of interest ; otherwise, carry out the FF test specifying systematically a non-proportional effect for the covariate of interest.

One solution to these problems of misspecification could be the development of an approach able to test simultaneously several hypotheses such as the PH assumption and the FF of the covariates of interest. Several tools have been developed within the context of overall survival (Pohar-Perme and Andersen, 2008; Sasieni, 2003). An interesting but complex perspective is to adapt these tests to the context of net survival.

As shown above, parametric functions such as splines or fractional polynomials can be successfully used to model non-proportionality (or non-linearity). The likelihood ratio test or AIC/BIC criteria can then be used to assess the value of the added complexity. We support and extensively use this approach in our studies (Bossard *and others*, 2007; Remontet *and others*, 2007; Corm *and others*, 2011; Mounier *and others*, 2015) as it is a very useful tool. Our proposed tests have the advantage to avoid a specific alternative hypothesis. A comparison between this two way of assessing proportionality and linearity would be an interesting point to study. The new tests indi-

cate not only whether the null hypothesis should be rejected or not, but also, through the plot of the observed process with its corresponding Gaussian processes, they allow knowing over which period or range of values of the covariate of interest, the assumption is not met. In complement to this graph, one can still use smoothed residual scatterplots to investigate the behavior of the effect of the covariates over time or of their functional form.

With a sample size of 2000 patients and 1000 simulated Gaussian processes, the computational time for the model fit, the PH test and the FF test are respectively 3 secs, 13 secs and 2.1 secs with a personal computer (64 bits, 16 Go RAM, processor Intel® Core^{TM} i7-4790 3.60 GHz). With a sample size of 100,000 patients, these times are respectively 2.8 mins, 5.6 mins and 1.5 mins.

In conclusion, we have now two powerful formal tests available for flexible parametric models. To make these tests user-friendly, they will be integrated into the `mexhaz` R package (Charvat and Belot, 2016) so that checking model assumptions could be easily performed by the users of this package (which already allows to fit the flexible model used in our analysis). Staying in the same theoretical framework defined by Lin and using an adequate weighting of martingale residuals, we plan to develop, in the future, a test to check the link function (very close to the FF test) and another to check the global goodness of fit ; this would complete a toolbox that already includes the PH and FF tests.

Acknowledgments

The authors thank the Ligue Nationale Contre le Cancer which funded the work of the first author. They thank the ANR (Agence Nationale de la Recherche) for supporting this study of the CENSUR group (ANR grant number ANR-12-BSV1-0028). The authors are also grateful for Jean Iwaz for editing the latest drafts of the manuscript, Jean-Yves Dauxois whose stochastic process courses have been very helpful, Jacques Estève for his help and his relevant advice, the

two referees for their helpful comments and the French network of cancer registries (FRANCIM) which provided the data.

References

Abrahamowicz, M. and Mackenzie, T.A. (2007). Joint estimation of time-dependent and non-linear effects of continuous covariates on survival. *Statistics in Medicine* **26**, 392–408.

Allemani, C., Weir, H.K., Carreira, H. and Group, Concord Working. (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). *Lancet* **14**, 977–1010.

Andersson, T.M., Dickman, P.W., Eloranta, S., Lambe, M. and Lambert, P.C. (2013). Estimating the loss in expectation of life due to cancer using flexible parametric survival models. *Statistics in Medicine* **32**, 5286–5300.

Barlow, W.E. and Prentice, R.L. (1988). Residual for relative risk regression. *Biometrika* **75**, 65–74.

Bossard, N., Velten, M., L.Remontet, Belot, A., Maarouf, N., Bouvier, A.M., Guizard, A.V., Tretarre, B., Launoy, G., Colonna, M., Danzon, A., Molinie, F., Troussard, X., Bourdon-Raverdy, N., Carli, P.M., Jaffré, A., Bessaguet, C., Sauleau, E., Schvartz, C., Arveux, P., Maynadié, M., Grosclaude, P., Estève, J. *and others*. (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.

Charvat, H. and Belot, A. (2016). mexhaz : Mixed effect excess hazard models. R package version 1.1.

Charvat, H., Bossard, N., L.Daubisse, Binder, F., Belot, A. and Remontet, L. (2013).

Probabilities of dying from cancer and other causes in french cancer patients based on an unbiased estimator of net survival : a study of five common cancers. *Cancer Epidemiology* **37**, 857–863.

Corm, S., Roche, L., Micol, JB., Coiteux, V., Bossard, N., Nicolini, FE., Iwaz, J., Preudhomme, C., Roche-Lestienne, C., Facon, T. *and others.* (2011). Changes in the dynamics of the excess mortality rate in chronic phase-chronic myeloid leukemia over 1990-2007 : a population study. *Blood* **18**, 4331–4337.

Cortese, G. and Scheike, T.H. (2008). Dynamic regression hazards models for relative survival. *Statistics in Medicine* **27**, 3563–3584.

Crowther, M.J. and Lambert, P.C. (2014). A general framework for parametric survival analysis. *Statistics in Medicine* **33**, 5280–5297.

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., Minicozzi, P., Sant, M., Dal-Maso, L., Brewster, DH., Osca-Gelis, G., Visse, O., Maynadié, M., Marcos-Gragera, R., Troussard, X., Agius, D., Roazzi, P., Meneghini, E., Monnereau, A. *and others.* (2015). Survival variations by country and age for lymphoid and myeloid malignancies in europe 2000-2007 : Results of eurocare-5 population-based study. *European Journal of Cancer* **51**, 2254–2268.

Dickman, PW., Sloggett, A., Hills, M. and Hakulinen, T. (2004). Regression models for relative survival. *Statistics in Medicine* **23**, 51–64.

Esteve, 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.

Giorgi, R., Abrahamowicz, M., Bolard, P., Esteve, J., Gouvernet, J. and Faivre, J. (2003). A relative survival regression model using b-spline functions to model non-proportional hazards. *Statistics in Medicine* **22**, 2767–2784.

Grambsch, P.M. and Therneau, T.M. (1994). A relativeproportional hazard tests and diagnostics based on weighted residuals. *Biometrika* **81**, 515–526.

King, N.B., Harper, S. and Young, M.E. (2012). Use of relative and absolute effect measures in reporting health inequalities : structured review. *British Medical Journal* **345**, 1–8.

Lambert, PC., Smith, LK., Jones, DR. and Botha, JL. (2005). Additive and multiplicative covariate regression models for relative survival incorporating fractional polynomials for time-dependent effects. *Statistics in Medicine* **345**, 3871–3885.

Lin, DY. and Spiekerman, CF. (1996). Model checking techniques for parametric regression with censored data. *Scandinavian Journal of Statistics* **23**, 157–177.

Lin, DY., Wei, LJ. and Ying, ZL. (1993). Checking the cox model with cumulative sums of martingale-based residuals. *Biometrika* **80**, 557–572.

Mounier, M., Bossard, N., Remontet, L., Belot, A., Minicozzi, P., Angelis, R. De, Capoccacia, R., Monnereau, A., Troussard, X., Sant, M., Maynadié, M., Giorgi, R., Group, Eurocare-5 Working *and others*. (2015). Changes in dynamics of excess mortality rates and net survival after diagnosis of follicular lymphoma or diffuse large b-cell lymphoma : comparison between european population-based data (eurocare-5). *The Lancet Haematology* **11**, 481–491.

Nelson, CP., Lambert, PC., Squire, IB. and Jones, DR. (2007). Flexible parametric models for relative survival, with application in coronary heart disease. *Statistics in Medicine* **26**, 5486–5498.

Pohar, M. and Stare, J. (2006). Relative survival analysis in r. *Computer Methods and Programs in Biomedicine* **81**, 272–278.

Pohar-Perme, M. and Andersen, PK. (2008). Checking hazard regression models using pseudo-observations. *Statistics in Medicine* **27**, 5309–5328.

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

Remontet, L., Bossard, N., Belot, A., Esteve, J. and Francim, French Network of Cancer Registries FRANCIM. (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.

Sasieni, PD. (2003). Martingale difference residuals as a diagnostic tool for the cox model. *Biometrika* **90**, 899–912.

Schoenfeld, D. (1982). Partial residuals for the proportional hazards regression model. *Biometrika* **69**, 239–241.

Stare, J., Pohar, M. and Henderson, R. (2005). Goodness of fit of relative survival models. *Statistics in Medicine* **24**, 3911–3925.

Therneau, TM., Grambsch, PM. and Fleming, TR. (1990). Martingale-based residuals for survival models. *Biometrika* **77**, 147–160.

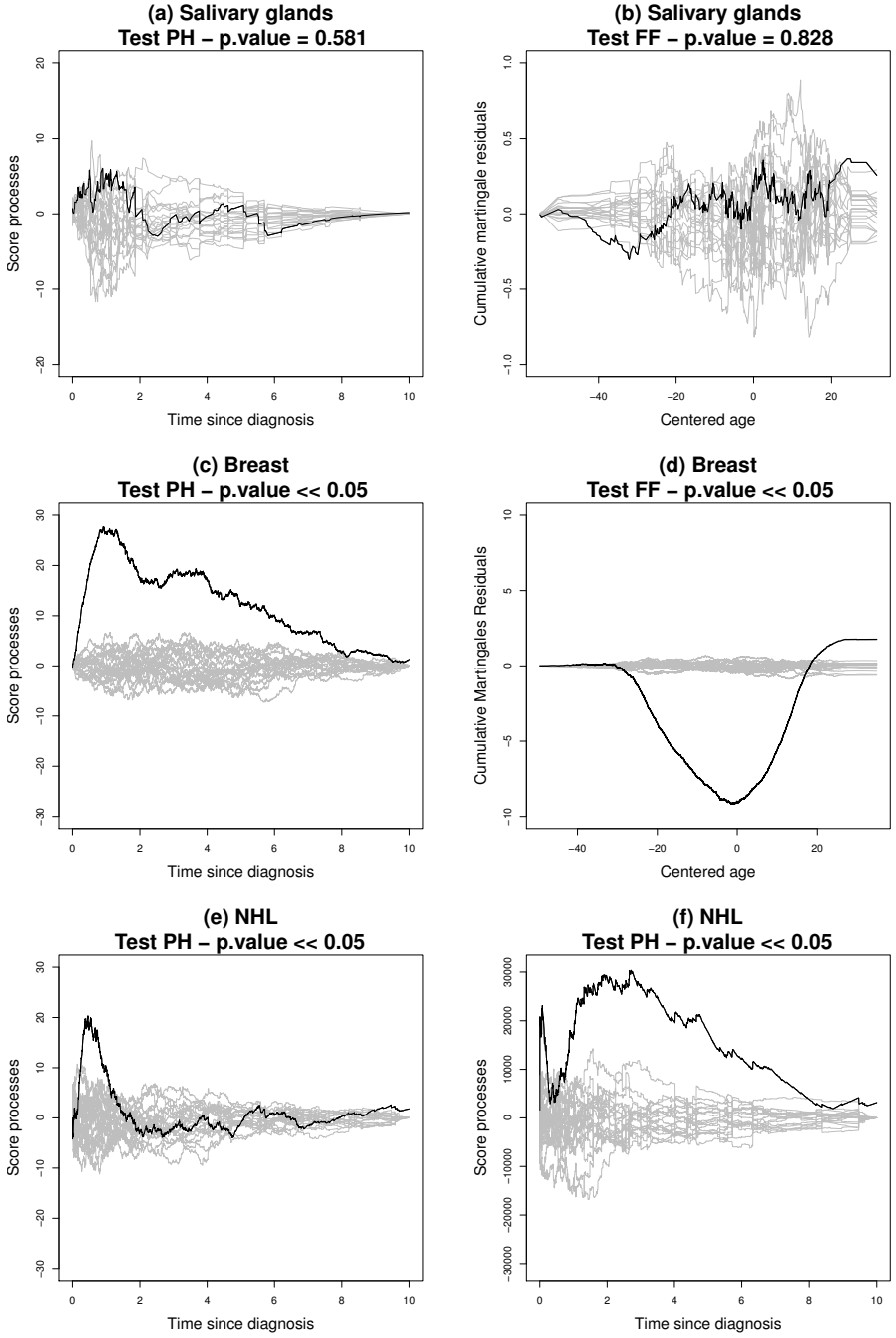

FIGURE 1. Panels a, c, and e show the plots of the score process over time (black) on the basis of the fitting model $\log\left[\lambda_E\left(t, age\right)\right] = \lambda_0\left(t\right) + \beta \times age$ with 20 corresponding Gaussian processes (grey) for salivary gland, breast, and non-Hodgkin lymphoma cancers, respectively. Panels b and d show the plots of the cumulative martingale residuals over covariate age values (black) on the basis of fitting model $\log\left[\lambda_E\left(t, age\right)\right] = \lambda_0\left(t\right) + \beta \times age$ with 20 corresponding Gaussian processes (grey) for salivary gland and breast cancer, respectively. Panel f shows the plot of the score process over time (black) on the basis of fitting model $\log\left[\lambda_E\left(t, age\right)\right] = \lambda_0\left(t\right) + g\left(age\right)$ where g is a cubic spline with 20 corresponding Gaussian processes (grey) for non-Hodgkin lymphoma. The estimated p-value stems from the simulation of 1000 Gaussian processes.

Table 1. *Size and power of the new PH test and of Stare tests in checking the proportional hazard assumption and of the new FF test in checking the functional form of a covariate.*

				Simulated L-PH scenario : size $log\left[\lambda_E\left(t, age\right)\right] = \lambda_0\left(t\right) + 0.02 \times age$				
n	New PH test (%)	T1 ML (%)	T1 EM (%)	T2 ML (%)	T2 EM (%)	T3 ML (%)	T3 EM (%)	New FF test (%)
500	4.4	2.9	5.6	3.6	5.5	4.2	5.7	4.6
1000	6.2	3.8	6.4	4.2	6.4	5.1	6.3	4.7
2000	4.9	4.0	5.9	4.0	5.9	4.9	4.6	5.7
				Simulated NPH/NLIN scenario : power				
n	New PH test (%)	T1 ML (%)	T1 EM (%)	T2 ML (%)	T2 EM (%)	T3 ML (%)	T3 EM (%)	New FF test (%)
Low NPH/NLIN effect								
500	34.7	22.4	15.4	34.7	25.8	37.1	34.9	26.2
1000	67.3	50.0	40.4	65.9	56.6	67.4	63.5	54.4
2000	95.2	89.9	81.9	95.5	91.2	95.8	93.0	81.3
Medium NPH/NLIN effect								
500	70.3	55.8	46.6	70.0	60.2	72.8	69.9	53.0
1000	96.7	90.9	84.6	96.9	92.7	96.9	96.0	91.0
2000	100.0	99.9	99.5	100.0	99.9	99.9	99.9	99.4
High NPH/NLIN effect								
500	80.7	72.4	76.0	80.4	81.0	76.9	76.1	100.0
1000	98.6	96.3	98.1	98.6	99.0	97.7	97.6	100.0
2000	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0

Table 2. *Summary of the simulation design in case of misspecification or over-parameterization of other assumptions than the one of interest*

To measure	**Simulated data under**	**Model used for $log\left[\lambda_E\left(t,age\right)\right]$ in simulation**$^{(1)}$	**Source of β_0-β_4 values**	**Source of β_5-β_9 values**	**NPH/NLIN strength effect**	**Model used for $log\left[\lambda_E\left(t,age\right)\right]$ in analysis**$^{(2)}$
Inflation of size of PH test	H_0	$\sim FP+g\left(age\right)$	Colon cancer	Colon cancer	Low	$\sim QS+\alpha_5\times age$
Loss of power PH test	H_1	$\sim FP+\beta\left(t\right)\times age$	Colon cancer	Stomach cancer	Low	$\sim QS+g\left(age\right)$
Inflation of size of FF test	H_0	$\sim FP+\beta\left(t\right)\times age$	Colon cancer	stomach cancer	Low	$\sim QS+\alpha_5\times age$
Loss of power FF test	H_1	$\sim FP+g\left(age\right)$	Colon cancer	Colon cancer	Low	$\sim QS+\beta\left(t\right)\times age$

$^{(1)}$FP=Fractional Polynomial $=\beta_0+\frac{\beta_1}{t+1}+\beta_2 log\left(t+1\right)+\beta_3 t+\beta_4 t^2$

$\beta\left(t\right)=\left(\beta_5+\beta_6 t+\beta_7 t^2+\beta_8 t^3+\beta_9\left(t-1\right)^3 I(t>1)\right)$

$g\left(age\right)=\beta_5\times age+\beta_6\times age^2+\beta_7\times age^3+\beta_8\times\left(age-mean\right)^3 I\left(age>mean\right)$

$^{(2)}$QS=Quadratic regression spline $=\alpha_0+\alpha_1 t+\alpha_2 t^2+\alpha_3\left(t-1\right)^2 I(t>1)+\alpha_4\left(t-5\right)^2 I(t>5)$

$\beta\left(t\right)=\left(\beta_5+\beta_6 t+\beta_7 t^2+\beta_8 t^3+\beta_9\left(t-1\right)^3 I(t>1)\right)$

$g\left(age\right)=\beta_5\times age+\beta_6\times age^2+\beta_7\times age^3+\beta_8\times\left(age-mean\right)^3 I\left(age>mean\right)$

Table 3. *Size and power of the new tests in case of a misspecification or over-parameterization of other assumptions than the one of interest.*

	Size (%) in case of misspecification				**Power (%) in case of over-parameterization**			
n	New PH test (Reference*)	New PH test	New FF test (Reference*)	New FF test	New PH test (Reference*)	New PH test	New FF (Reference*)	New FF test
500	4.4	5.3	4.6	5.6	34.7	22.9	26.2	20.9
1000	6.2	6.8	4.7	6.3	67.3	51.3	54.4	47.7
2000	4.9	9.3	5.7	8.8	95.2	86.5	81.3	76.9

* Reminder of the results shown in Tables 1