Supplementary Material :
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.

1. Procedure used to generate T_E, the time to death from cancer

1.1 *Simulation under the alternative hypothesis.*

To assess the power of the test, T_E must be generated from a model where *age* has a non-proportional effect. As recommended by Crowther and Lambert (2013), we simulated realistic data through a full parametric model which can be used in the general algorithm of simulation described by Crowther and Lambert (2013). T_E was thus simulated from the following model $\log\left[\lambda_E\left(t|age,\boldsymbol{\theta}\right)\right] = \lambda_0\left(t,\boldsymbol{\chi}\right) + \beta\left(t\right) \times age$ (Model 1) where the baseline excess mortality rate $\lambda_0\left(t,\boldsymbol{\chi}\right)$ is a fractional polynomial :

$$\lambda_0\left(t,\boldsymbol{\chi}\right) = FP(t) = \beta_0 + \frac{\beta_1}{t+1} + \beta_2\log\left(t+1\right) + \beta_3 t + \beta_4 t^2$$

and $\beta\left(t\right)$ is a cubic regression spline with 1 knot at one year :

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

To simulate plausible effects with the three desired levels of non-proportionality, the values of β_0 to β_9 were chosen as follows (these choices are summarized in Table 1 in this document - http ://www.biostatistics.oxfordjournals.org) : i) we selected the 20 most frequent cancer sites from the common database of French cancer registries, ii) we fitted an excess hazard model on samples of each of these 20 real datasets (up to 10 years of follow-up), the excess mortality hazard λ_E being specified as in Model 1, iii) for each cancer site, we measured the strength of the NPH effect with a particular measure defined as the area between $\beta\left(t\right)$ and the mean of $\beta\left(t\right)$ over time, and iv) we defined arbitrarily the "low NPH effect" as the 60th percentile of this measure (which corresponds to stomach cancer), the "medium NPH effect" as the 80th percentile (colon cancer), and the "high NPH effect" as the 90th percentile (thyroid cancer). We chose percentiles over the 50th so that the NPH effect could exist (i.e., simulation under H_1). Figure 1 (in this document) shows the shape of the three NPH effects $\beta\left(t\right)$ according to the time since diagnosis. Furthermore, the power of the test was deemed dependent on the number of deaths observed

in the simulated cohort, this number being dependent on the baseline hazard level and on the NPH level. So, to analyze the specific effect of the NPH levels, we controlled the baseline excess mortality hazard and fixed its shape and strength using parameters β_0 to β_4 observed in colon cancer.

1.2 *Simulation under the null hypothesis.*

Finally, to assess the size, T_E was generated from the following proportional hazard model :

$log\left[\lambda_E\left(t|age, \boldsymbol{\theta}\right)\right] = \beta_0 + \frac{\beta_1}{t+1} + \beta_2 log\left(t+1\right) + \beta_3 t + \beta_4 t^2 + \beta_5 \times age$ (Model 2) where the values of β_0 to β_5 were estimated from colon cancer data (Table 1 in this document).

After determining the parameters of the model, T_E was generated using the inverse transform approach described by Crowther and Lambert (2013) for time-dependent effects : the parametric form of the model allows obtaining the cumulative distribution which is then numerically inverted.

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

This simulation study is largely based on the elements already presented in the previous section. Thus, we present here only the elements that differ from this previous section. We assessed the size and the power of the new test designed to check the FF of a covariate (called thereafter "the new FF test"). This assessment was performed in the particular case where hypothesis H_0 is "the functional form of the covariate is linear". The performance of this test was estimated according to the sample size and the strength of the non-linearity (NLIN) effect (low, medium, high) under the alternative hypothesis H_1. To obtain plausible effects with the three levels of NLIN strength, we operated as follows (Table 1 in this document) : i) for each of the 20 most frequent cancer sites, we fit the following model $log\left[\lambda_E\left(t|age, \boldsymbol{\theta}\right)\right] = \lambda_0\left(t, \boldsymbol{\chi}\right) + g\left(age\right)$, $\lambda_0\left(t, \boldsymbol{\chi}\right)$ being a fractional polynomial and $g\left(age\right)$ a cubic spline with one knot at the mean age, $g\left(age\right) =$

$\beta_5 \times age + \beta_6 \times age^2 + \beta_7 \times age^3 + \beta_8 \times (age - mean)^3 I (age > mean)$; ii) we measured the strength of the NLIN effect with a specific measure similar to that of the new PH test but adapted for linearity; this corresponds to the area between $g(age)$ and the mean of $g(age)$ over the age range; iii) from the percentiles of this measure, we defined arbitrarily a "low", a "medium", and a "high" NLIN effect that correspond, respectively, to colon, cervix, and prostate cancers.
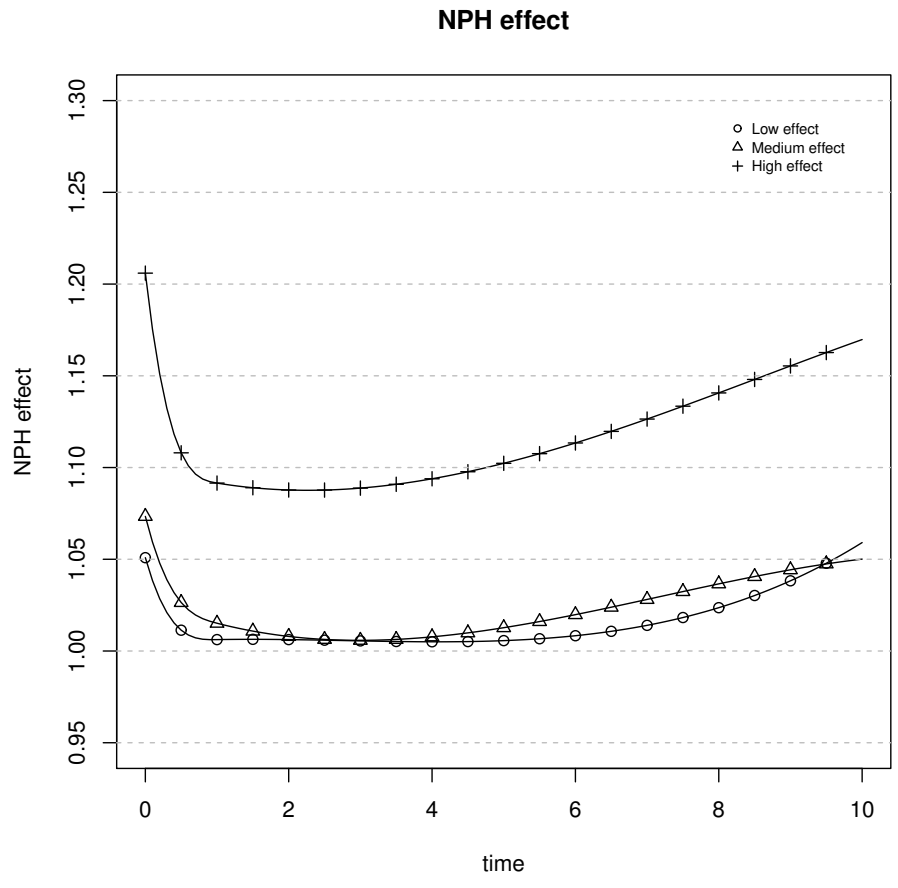
FIGURE 1. Shape of the three simulated NPH effects $\beta(t)$: circles for low NPH effect, triangles for medium NPH effect, and crosses for high NPH effect.

Table 1. *Summary of the simulation design*

Test	Criterion	Simulated Data under	Model used for $\log[\lambda_E(t, age)]$ in simulation[1]	Source of β_0-β_4 values	Source of β_5-β_9 values	NPH/NLIN strength effect	Model used for $\log[\lambda_E(t, age)]$ in analysis[2]
PH	Power	H_1	$\sim FP(t) + \beta(t) \times age$	Colon cancer	Stomach cancer	Low	$\sim QS(t) + \alpha_5 \times age$
PH	Power	H_1	$\sim FP(t) + \beta(t) \times age$	Colon cancer	Colon cancer	Medium	$\sim QS(t) + \alpha_5 \times age$
PH	Power	H_1	$\sim FP(t) + \beta(t) \times age$	Colon cancer	Thyroid cancer	High	$\sim QS(t) + \alpha_5 \times age$
PH	Size	H_0	$\sim FP(t) + \beta_5 \times age$	Colon cancer	Colon cancer	None	$\sim QS(t) + \alpha_5 \times age$
FF	Power	H_1	$\sim FP(t) + g(age)$	Colon cancer	Colon cancer	Low	$\sim QS(t) + \alpha_5 \times age$
FF	Power	H_1	$\sim FP(t) + g(age)$	Colon cancer	Cervix cancer	Medium	$\sim QS(t) + \alpha_5 \times age$
FF	Power	H_1	$\sim FP(t) + g(age)$	Colon cancer	Prostate cancer	High	$\sim QS(t) + \alpha_5 \times age$
FF	Size	H_0	$\sim FP(t) + \beta_5 \times age$	Colon cancer	Colon cancer	None	$\sim QS(t) + \alpha_5 \times age$

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

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

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

[2] Quadratic regression spline : $QS(t) = \alpha_0 + \alpha_1 t + \alpha_2 t^2 + \alpha_3 (t-1)^2 I(t>1) + \alpha_4 (t-5)^2 I(t>5)$

3. Approximation of the limiting distribution of the score process under the null hypothesis.

Most of this section comes from Lin and Spiekerman (1996). Here, we present the main steps, starting from the score process, to derive the process that was simulated to approximate the limiting distribution of the score process.

Let $\boldsymbol{\theta_0}$ be the true values of the parameters. A Taylor expansion of $\mathbf{U}(.,t)$ (the score process evaluated at time t) in the neighborhood of $\boldsymbol{\theta_0}$ gives :

$$\mathbf{U}(\boldsymbol{\theta},t) = \mathbf{U}(\boldsymbol{\theta_0},t) - n\mathbf{I}(\boldsymbol{\theta_0},t)(\boldsymbol{\theta}-\boldsymbol{\theta_0}) + o_p(1) \tag{3.1}$$

where $\mathbf{I}(\boldsymbol{\theta},t)$ is a $p \times p$ matrix (p equals to the number of parameters in the model) with the element $I_{kj}(\boldsymbol{\theta},t)$ (i.e., the element in the kth row and j^{th} column) equals to :

$$-\frac{1}{n}\frac{\partial U_k}{\partial \theta_j}(\boldsymbol{\theta},t)$$

The matrix $\mathbf{I}$ evaluated at $(\boldsymbol{\theta_0},\tau)$ is related to the Fisher information matrix and is invertible. Evaluating (3.1) at $(\widehat{\boldsymbol{\theta}},\tau)$ and remarking that $\mathbf{U}(\widehat{\boldsymbol{\theta}},\tau)$ corresponds to the score of the likelihood function evaluated at the maximum likelihood estimate (and thus is equal to the zero vector), we obtain :

$$n\mathbf{I}(\boldsymbol{\theta_0},\tau)\left(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta_0}\right) = \mathbf{U}(\boldsymbol{\theta_0},\tau) + o_p(1)$$

$$\left(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta_0}\right) = \frac{1}{n}\mathbf{I}^{-1}(\boldsymbol{\theta_0},\tau)\mathbf{U}(\boldsymbol{\theta_0},\tau) + o_p(1) \tag{3.2}$$

Then, combining equation (3.1) and (3.2) and evaluating $\mathbf{U}$ at $\widehat{\boldsymbol{\theta}}$ and time t,

$$\mathbf{U}\left(\widehat{\boldsymbol{\theta}},t\right) = \mathbf{U}(\boldsymbol{\theta_0},t) - \mathbf{I}(\boldsymbol{\theta_0},t)\mathbf{I}^{-1}(\boldsymbol{\theta_0},\tau)\mathbf{U}(\boldsymbol{\theta_0},\tau) + o_p(1) \tag{3.3}$$

In equation (3.3), the distribution of the components of the score process $U_k(\boldsymbol{\theta_0},t)$ is unknown. Indeed, the kth component of the score process at $\boldsymbol{\theta_0}$ is :

$$U_k(\boldsymbol{\theta_0},t) = \sum_i \left[Z_{ki} \int_0^t \frac{\lambda_E(u|\mathbf{Z_i},\boldsymbol{\theta_0})}{\lambda_E(u|\mathbf{Z_i},\boldsymbol{\theta_0}) + \lambda_P(a_i+u,\widetilde{\mathbf{z_i}})} dM_i(u) \right]$$

with

$$M_i\left(t\right) = N_i\left(t\right) - \int_0^t Y_i\left(u\right)\lambda_E\left(u|\mathbf{Z_i},\boldsymbol{\theta_0}\right)du - \int_0^t Y_i\left(u\right)\lambda_P\left(a_i + u, \widetilde{\mathbf{z_i}}\right)du$$

The distribution of M_i is unknown and, similarly to Lin *and others* (1993) and Lin and Spiekerman (1996), we replace M_i by a similar process M_iG_i, with G_i following a normal distribution. This process shares characteristics of the process M_i, including mean and variance (Fleming and Harrington, 1991; Lin *and others*, 1993).

Thus, $U_k\left(\boldsymbol{\theta_0},.\right)$ will be approximated by the kth component of $\mathbf{D_1}\left(\boldsymbol{\theta},.\right)$ defined as follows :

$$\sum_i\left[Z_{ki}\int_0^t \frac{\lambda_E\left(u|\mathbf{Z_i},\boldsymbol{\theta}\right)}{\lambda_E\left(u|\mathbf{Z_i},\boldsymbol{\theta}\right) + \lambda_P\left(a_i + u, \widetilde{\mathbf{z_i}}\right)}dN_i\left(u\right)G_i\right] \tag{3.4}$$

Here, we focus on the score process. In this context, $n^{-1/2}U$ is a special case of the process $\mathbf{W_z^{(2)}}$ (see the end of Section 2.2.2 of the article). Finally, combining equations (3.3) and (3.4), and approximating $\boldsymbol{\theta_0}$ by $\widehat{\boldsymbol{\theta}}$, the approximation of the distribution of $\left\{n^{-1/2}\widehat{U}\right\}$ is performed by generating a 'large' number of realizations of the following process :

$$n^{-1/2}\left(\mathbf{D_1}\left(\widehat{\boldsymbol{\theta}},t\right) - \mathbf{I}\left(\widehat{\boldsymbol{\theta}},t\right)\mathbf{I}\left(\widehat{\boldsymbol{\theta}},\tau\right)^{-1}\mathbf{D_1}\left(\widehat{\boldsymbol{\theta}},\tau\right)\right) \tag{3.5}$$

To check PH assumption of the kth covariate, the "observed" test statistic value of the corresponding dataset is :

$$\widehat{T}_S = \sup\left|n^{-1/2}U_k\left(\widehat{\boldsymbol{\theta}},t\right)\right| \tag{3.6}$$

The p-value is then approximated by the proportion of cases in which the absolute maximum of the simulated Gaussian Processes (supremum over time t of the absolute value of the kth component of process (3.5), i.e., replacing U_k in (3.6) with the kth component of (3.5)) is higher than $\widehat{T}_S$.

REFERENCES

CROWTHER, M.J. AND LAMBERT, P.C. (2013). Simulating biologically plausible complex survival data. *Statistics in Medicine* **32**, 4118–4134.

FLEMING, T. R. AND HARRINGTON, D. P. (1991). *Counting processes and survival analysis*, 2th edition. New York : Wiley.

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.