Goodness-of-fit tests for parametric excess hazard rate models with covariates

Laurent Bordes, Olayidé Boussari, Valérie Jooste

Goodness-of-fit tests for parametric excess hazard rate models with covariates

Laurent Bordes

Univ. Pau & Pays Adour, CNRS, UMR 5142

Laboratoire de Mathématiques et de leurs Applications – IPRA, 64000 Pau, France

email: laurent.bordes@univ-pau.fr

and

Olayidé Boussari

Univ. Bourgogne Franche-Comté, INSERM, UMR1231, LabEX LipSTIC, ANR-11-LABX-0021

Registre Bourguignon des Cancers Digestifs, 21079 Dijon, France

email: olayide.boussari@u-bourgogne.fr

and

Valérie Jooste

Univ. Bourgogne Franche-Comté, INSERM, UMR1231, LabEX LipSTIC, ANR-11-LABX-0021

Registre Bourguignon des Cancers Digestifs, Centre Hospitalier Universitaire de Dijon, 21079 Dijon, France.

email: valerie.jooste@u-bourgogne.fr

SUMMARY: In this paper we propose a general methodology for testing the null hypothesis that an excess hazard rate model, with or without covariates, belongs to a parametric family. Estimating the excess hazard rate function parametrically through the maximum likelihood method and non-parametrically (or semi-parametrically) we build a discrepancy process which is shown to be asymptotically Gaussian under the null hypothesis. Based on this result we are able to build some statistical tests in order to decide wether or not the null hypothesis is acceptable. We illustrate our results by the construction of chi-square tests which the behavior is studied through a Monte-Carlo study. Then the testing procedure is applied to a population based colon cancer data.

KEYWORDS: Excess hazard model; Proportional excess hazards model; Maximum likelihood estimation; Covariates; Semiparametric estimation; Nonparametric estimation; colon cancer data.

1. Introduction

Cure models are used in population based cancer epidemiology, their application relies on the existence of statistical cure. Net survival, which is the survival that would be observed in a hypothetical world where cancer would be the only possible cause of death (see e.g. Cronin and Feuer 2000, Pohar Perme, Stare and Estève 2012), provides an objective measure of the proportion of patients dying from direct or indirect consequences of cancer without requiring a record of the cause of death. It is usually estimated by excess mortality rate modeling. In situations where some patients will never experience death due to cancer ("cured patients"), the net survival curve flattens at a non-zero value after a while, when the excess mortality rate due to cancer, denoted hereafter λ_{exc}, reaches zero. This is a population definition of cure and does not necessarily imply that patients are medically cured. In order to use cure models, the cure assumption needs to be assessed. Current methods are based on graphical assessment of a plateau in net survival, which may not be satisfactory, thus there is a need to provide an objective answer to the existence of statistical cure. This requires to provide more and more sophisticated models for the excess hazard function. These models can be parametric, semi- or non-parametric. Danieli et al. (2012) have shown the importance of using flexible nonparametric estimators of the net survival function in order to prevent bias due to several practical situations that maybe encountered in population based studies for cancer registries. The choice of the best estimation methodology is still an important matter, for recent contributions see for instance Yu et al. (2013), Lambert, Dickman and Rutherford (2015), Seppä, Hakulinen and Pokhrel (2015), Seppä et al. (2016). Because semi- or non-parametric models are less constrained than parametric ones, they should be preferred when we are not able to guarantee that a parametric model is compatible with the data. However, if the parametric model is data–compatible, it may be more interesting to use such a model

since usual indicators (e.g. risk functions, quantiles, conditional expectations, confidence domains, etc.) are generally obtained more easily under parametric assumptions.

The aim of the paper is to provide a general methodology for testing the hypothesis that the excess risk function λ_{exc} belongs to a parametric family. Our approach is based on a comparison of the maximum likelihood estimator (MLE) of the excess cumulative hazard and an adaptation of the semiparametric (or nonparametric) estimator of Sasieni (1996) that includes a large class of performant nonparametric estimators like Pohar Perme et al. (2012), and Kodre and Perme (2013). However we want to emphasize that the class of available nonparametric estimators is not reduced to the Sasieni familly of estimators. For instance Cortese and Scheike (2008) proposed to estimate the excess hazard rate function by a nonparametric additive regression model extending the well–known additive hazard model introduced by Aalen (1980). Let us note that in the domain of inference Cortese and Scheike (2008) also developed test procedures based on residuals for the proportional excess model (see also Stare, Pohar, and Henderson 2005), and Kannan et al. (2010) developed some graphical goodness–of–fit for a generalized exponential cure rate model with covariates. Recently Grafféo et al. (2016) developed some log-rank-type tests to compare net survival distributions.

In order to test that the excess risk function λ_{exc} belongs to a parametric family we build some statistical tests based on the discrepancy process, that is the difference between the MLE and the Sasieni estimator of the excess cumulative hazard rate, multiplied by the root of n (the sample size). The study of the asymptotic behavior (with respect to n) of the discrepancy process, using usual martingale methods for counting processes a la Andersen et al. (1993), allows to derive some test statistics as well as their distribution. In this paper we focus on chi-square type statistics which present the advantage of having an asymptotic

distribution free of the unknown model parameters under the null hypothesis (for more on chi-square testing see for instance Greenwood and Nikulin 1996 and Khmaladze 2013).

The paper is organized as follows. In Section 2 we describe both parametric and semiparametric models for the excess hazard rate and we recall the basic martingale properties of counting processes. Section 3 is devoted to the construction of test statistics for homogeneous data, that is data for which we consider that there is no covariate effect on the excess hazard rate. Section 4 is similar to Section 3 for data including covariate effects on the excess hazard rate function. A numerical study is conducted in Section 5 including both a simulation part giving empirical evidence that our testing procedure behaves well, and an analysis of colorectal cancer data. Some concluding remarks are given in Section 6.

2. Parametric and semiparametric models for the excess hazard rate

Let A be the age at which an individual is diagnosed, T be the time elapsed between A and the time of death of the individual, C be a right censoring time and $\boldsymbol{Z}$ be a vector of covariates in $\mathbb{R}^d$. An observation is a quadruple $(X, \Delta, A, \boldsymbol{Z})$, where $X = \min(T, C)$ and $\Delta = \mathbb{1}_{\{T \leqslant C\}}$ (here $\mathbb{1}_E$ denotes the set indicator function, equal to 1 if E is true and 0 otherwise). We denote by $\lambda_{\text{obs}}(t|a, \boldsymbol{z})$ the hazard rate function of T given $(A, \boldsymbol{Z}) = (a, \boldsymbol{z})$. When there is no covariate in the model we simply delete $\boldsymbol{Z}$ or $\boldsymbol{z}$ from our notations.

We denote by $\lambda_{\text{pop}}(a + t|\boldsymbol{z})$ the general population hazard rate for an individual with covariate $\boldsymbol{Z} = \boldsymbol{z}$ at the calendar time $a + t$. Then we consider two different models for the conditional excess hazard rate function $\lambda_{\text{exc}}(t|\boldsymbol{z})$.

Proportional hazards model. We assume that the observed hazard rate function λ_{obs} is defined by

$$\lambda_{\text{obs}}(t|a, \boldsymbol{z}) = \lambda_{\text{pop}}(a + t|\boldsymbol{z}) + \exp(\boldsymbol{\beta z})\lambda_{\text{exc}}(t), \tag{1}$$

where $\boldsymbol{\beta}$ is an unknown regression parameter in $\mathbb{R}^d$ and λ_{exc} is an unknown baseline hazard rate function. Both parameters $\boldsymbol{\beta}$ and λ_{exc} have to be estimated. Notice that considering a single vector of covariates $\boldsymbol{z}$ is a simplification of notations in the sense that (1) could be rewritten

$$\lambda_{\text{obs}}(t|a, \boldsymbol{z}) = \lambda_{\text{pop}}(a + t|\boldsymbol{z}_1) + \exp(\boldsymbol{\beta}\boldsymbol{z}_2)\lambda_{\text{exc}}(t),$$

with $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$ two subsets of covariates from $\boldsymbol{z}$.

Parametric proportional hazards model. We assume that the observed hazard rate function λ_{obs} is defined by

$$\lambda_{\text{obs}}(t|a, \boldsymbol{z}) = \lambda_{\text{pop}}(a + t|\boldsymbol{z}) + \exp(\boldsymbol{\beta}\boldsymbol{z})\lambda_{\text{exc}}(t|\boldsymbol{\theta}), \tag{2}$$

where $\boldsymbol{\beta}$ is an unknown regression parameter in $\mathbb{R}^d$ and $\lambda_{\text{exc}}(t; \boldsymbol{\theta})$ belongs to a parametric family $\mathcal{A} = \{\lambda_{\text{exc}}(\cdot|\boldsymbol{\theta}); \boldsymbol{\theta} \in \Theta \subset \mathbb{R}^p\}$. Assuming the absence of covariate effects on the excess risk function is equivalent to suppose that $\boldsymbol{\beta} = 0$ in the above models (1) and (2). We note that generally $\mathcal{A}$ is a parametric family of cure models. It means that the net survival function corresponding to the excess risk function $\lambda_{\text{exc}}(\cdot|\boldsymbol{\theta})$ is written $S_{\text{net}}(\cdot|\boldsymbol{\theta}) = \pi + (1 - \pi)S_0(\cdot|\boldsymbol{\gamma})$ with $\boldsymbol{\theta} = (\boldsymbol{\gamma}, \pi) \in \Gamma \times [0, 1]$ where $S_0(\cdot|\boldsymbol{\gamma})$ is a parametric survival function indexed by the Euclidean parameter $\boldsymbol{\gamma} \in \Gamma$, and $\pi \in [0, 1]$ is the cure rate. An example of such a Weibull cure model is defined in the appendix. Because π represents the fraction of cured people and because π belongs to $[0, 1]$, such a model includes the possibility of curing ($\pi \in (0, 1]$) as well as the possibility of dealing with an incurable disease ($\pi = 0$). Thus, for a parametric family $\mathcal{A}$, testing absence of cure is possible only if we are able to verify that the data are compatible with the parametric model (2), which in turn requires to test the composite null hypothesis H_0 that $\lambda_{\text{exc}} \in \mathcal{A}$.

Data and martingale properties. Let us now consider that we observe n independent and

identically distributed copies $\{(X_i, \Delta_i, A_i, \boldsymbol{Z}_i); 1 \leqslant i \leqslant n\}$ of $(X, \Delta, A, \boldsymbol{Z})$. The realization of $(X, \Delta, A, \boldsymbol{Z})$ (resp. $(X_i, \Delta_i, A_i, \boldsymbol{Z}_i)$) is denoted $(x, \delta, a, \boldsymbol{z})$ (resp. $(x_i, \delta_i, a_i, \boldsymbol{z}_i)$).

Let us introduce the usual processes N (counting process) and Y (at risk process) defined by $N(t) = \mathbb{1}_{\{X \leqslant t; \Delta = 1\}}$ and $Y(t) = \mathbb{1}_{\{X \geqslant t\}}$ and in a similar way the $2n$ processes N_i and Y_i. Then, considering the natural filtration $\mathbb{F}_n = (\mathcal{F}_t; t \geqslant 0)$ generated by the $2n$ processes N_i and Y_i, and the n diagnosis times a_i and covariates $\boldsymbol{z}_i$ we define

$$M_i(t|a_i, \boldsymbol{z}_i) = N_i(t) - \int_0^t Y_i(s)\lambda_{\text{obs}}(s|a_i, \boldsymbol{z}_i)ds \tag{3}$$

which are square integrable martingales with respect to the filtration $\mathbb{F}_n$ (see Andersen et al. 1993). Note that even if there is no covariate effect to specify in the excess hazard rate function, the population risk function generally varies with the individuals through several characteristics (sex, age, etc.). Thus we mention this fact by noticing $\lambda_{\text{pop}}^{(i)}$ the population risk of the ith individual.

3. Testing a parametric model without covariates effect

3.1 *Maximum likelihood principle*

The maximum likelihood estimator (MLE) satisfies

$$\begin{aligned} \hat{\boldsymbol{\theta}} &= \arg\max_{\boldsymbol{\theta} \in \Theta} \sum_{i=1}^{n} \left\{ \log\left(\lambda_{\text{pop}}^{(i)}(x_i + a_i) + \lambda_{\text{exc}}(x_i|\boldsymbol{\theta})\right) \delta_i - \Lambda_{\text{exc}}(x_i|\boldsymbol{\theta}) \right\} \\ &= \arg\max_{\boldsymbol{\theta} \in \Theta} \ell_n(\boldsymbol{\theta}). \end{aligned}$$

For any function $f : \boldsymbol{\theta} \mapsto f(\boldsymbol{\theta})$ we note $\dot{f}(\boldsymbol{\theta}) = \frac{\partial f}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta})$ and $\ddot{f}(\boldsymbol{\theta}) = \frac{\partial^2 f}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}(\boldsymbol{\theta})$. Because

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) = -\left[\frac{1}{n}\ddot{\ell}_n(\boldsymbol{\theta}^*)\right]^{-1} \frac{1}{\sqrt{n}}\dot{\ell}_n(\boldsymbol{\theta}_0),$$

where $\boldsymbol{\theta}^*$ lies between $\boldsymbol{\theta}_0$ and $\hat{\boldsymbol{\theta}}$, and

$$\dot{\ell}_n(\boldsymbol{\theta}_0) = \sum_{i=1}^{n} \int_0^{\tau} \frac{\dot{\lambda}_{\text{exc}}(x|\boldsymbol{\theta}_0)}{\lambda_{\text{pop}}^{(i)}(x + a_i) + \lambda_{\text{exc}}(x|\boldsymbol{\theta}_0)} dM_i(x|a_i),$$

where τ is the study duration (in practice we set $\tau = +\infty$). Then, using standard martingale methods for counting processes we obtain under standard regularity conditions (see Andersen

et al. 1993)

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\right)=I^{-1}(\boldsymbol{\theta}_0)\frac{1}{\sqrt{n}}\dot{\ell}_n(\boldsymbol{\theta}_0)+o_P(1), \tag{4}$$

where

$$\widehat{I}=\frac{1}{n}\sum_{i=1}^{n}\left[\frac{\delta_i\dot{\lambda}_{\text{exc}}(x_i|\hat{\boldsymbol{\theta}})}{\lambda^{(i)}_{\text{pop}}(x_i+a_i)+\lambda_{\text{exc}}(x_i|\hat{\boldsymbol{\theta}})}\right]^{\otimes 2}\xrightarrow{P} I(\boldsymbol{\theta}_0),$$

where for a column vector $\boldsymbol{u}$, $\boldsymbol{u}^{\otimes 2}=\boldsymbol{u}\boldsymbol{u}^T$. From the previous results we obtain

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\right)\xrightarrow{d}\mathcal{N}\left(0,I^{-1}(\boldsymbol{\theta}_0)\right). \tag{5}$$

The practical use of the last result consists in considering that $\mathcal{N}\left(\boldsymbol{\theta}_0,n^{-1}\widehat{I}^{-1}\right)$ is a good approximation of the distribution of the consistent estimator $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}_0$.

3.2 *Nonparametric estimation principle*

In Sasieni (1996) it is proved that the following nonparametric estimator of Λ_{exc} is asymptotically efficient:

$$\tilde{\Lambda}_{\text{exc}}(t)=\sum_{i=1}^{n}\int_0^t\frac{w_i(s)}{\sum_{j=1}^{n}w_j(s)Y_j(s)}\left(dN_i(s)-Y_i(s)\lambda^{(i)}_{\text{pop}}(a_i+s)ds\right)$$

whenener the weights functions w_i are defined by

$$w_i(s)=\frac{\lambda^{(i)}_{\text{pop}}(a_i+s)}{\lambda^{(i)}_{\text{pop}}(a_i+s)+\lambda_{\text{exc}}(s)}.$$

Because the weights w_i depend on the unknown quantity λ_{exc} and because our aim is to obtain a good estimator of Λ_{exc} (whether the data fit the parametric model or not) we propose to replace the unavailable quantities w_i by

$$w_i(s;\hat{\boldsymbol{\theta}})=\frac{\lambda^{(i)}_{\text{pop}}(a_i+s)}{\lambda^{(i)}_{\text{pop}}(a_i+s)+\lambda_{\text{exc}}(s|\hat{\boldsymbol{\theta}})}. \tag{6}$$

Thus $\Lambda_{\text{exc}}(t)$ is estimated by

$$\hat{\Lambda}_{\text{exc}}(t|\hat{\boldsymbol{\theta}})=\sum_{i=1}^{n}\int_0^t\frac{w_i(s;\hat{\boldsymbol{\theta}})}{\sum_{j=1}^{n}w_j(s;\hat{\boldsymbol{\theta}})Y_j(s)}\left(dN_i(s)-Y_i(s)\lambda^{(i)}_{\text{pop}}(a_i+s)ds\right). \tag{7}$$

REMARK 1: For various definitions of the weights we retrieve some well known estimators.

(i) If $w_i(s;\hat{\boldsymbol{\theta}})=1$ then (7) is nothing but the usual estimator by Andersen and Vaeth (1989), also known as Ederer II estimator (Ederer, Axtell and Cutler 1961).

(ii) If $w_i(s;\hat{\boldsymbol{\theta}}) = 1/S^{(i)}_{\text{pop}}(s)$ then (7) is the famous Pohar Perme et al. (2012) estimator.

(iii) If $w_i(s;\hat{\boldsymbol{\theta}}) = \lambda^{(i)}_{\text{pop}}(a_i+s)/(\lambda^{(i)}_{\text{pop}}(a_i+s) + \lambda_{\text{exc}}(s;\hat{\boldsymbol{\theta}}))$ then (7) is the Sasieni (1996) estimator (obtained by fixing the regression parameter to 0) that is shown to be asymptotically efficient in the nonparametric setup.

(iv) If $w_i(s;\hat{\boldsymbol{\theta}}) = 1/(\hat{S}_C(s)S^{(i)}_{\text{pop}}(s))$, where $\hat{S}_C$ is an estimator of the censoring time survival function, then (7) is the Kodre and Perme (2013) estimator.

3.3 *Testing a composite hypothesis*

Since the aim is to test the following composite null hypothesis

$$H_0 : \lambda_{\text{exc}} \in \{\lambda_{\text{exc}}(\cdot|\boldsymbol{\theta}); \boldsymbol{\theta} \in \Theta \subset \mathbb{R}^p\},$$

thus under H_0 our interest is to chose the weights defined in (6). It is important to note also that even if H_0 is not satisfied the estimator defined by (7) remains consistent.

Let us consider the discrepancy process $\mathfrak{D}_n$ defined for $t \in [0,\tau]$ by

$$\mathfrak{D}_n(t) = \sqrt{n}\left(\Lambda_{\text{exc}}(t|\hat{\boldsymbol{\theta}}) - \hat{\Lambda}_{\text{exc}}(t|\hat{\boldsymbol{\theta}})\right).$$

We show in the appendix that $\mathfrak{D}_n$ converges weakly to a centered gaussian process $\mathfrak{D}_\infty$ in $D([0,\tau])$ whose the covariance function η, defined for $(s,t) \in [0,\tau]^2$ by $\eta(s,t) = \mathbb{E}\left(\mathfrak{D}_\infty(s)\mathfrak{D}_\infty(t)\right)$, is consistently estimated by

$$\hat{\eta}(s,t) = \left(\dot{\Lambda}^T_{\text{exc}}(s|\hat{\boldsymbol{\theta}})\widehat{I}^{-1}, -1\right)\hat{\gamma}(s,t)\left(\dot{\Lambda}^T_{\text{exc}}(t|\hat{\boldsymbol{\theta}})\widehat{I}^{-1}, -1\right)^T,$$

where

$$\hat{\gamma}(s,t) = \tfrac{1}{n}\textstyle\sum_{i=1}^n \delta_i w_i^2(x_i;\hat{\boldsymbol{\theta}}) \times$$
$$\begin{pmatrix} \left(\dfrac{\dot{\lambda}_{\text{exc}}(x_i|\hat{\boldsymbol{\theta}})}{\lambda^{(i)}_{\text{pop}}(a_i+x_i)}\right)^{\otimes 2} & \dfrac{n\dot{\lambda}_{\text{exc}}(x_i|\hat{\boldsymbol{\theta}})1_{\{x_i\leqslant t\}}}{\lambda^{(i)}_{\text{pop}}(a_i+x_i)\sum_{j=1}^n w_j(x_i;\hat{\boldsymbol{\theta}})Y_j(x_i)} \\ \dfrac{n\dot{\lambda}^T_{\text{exc}}(x_i|\hat{\boldsymbol{\theta}})1_{\{x_i\leqslant s\}}}{\lambda^{(i)}_{\text{pop}}(a_i+x_i)\sum_{j=1}^n w_j(x_i;\hat{\boldsymbol{\theta}})Y_j(x_i)} & \dfrac{n^2 1_{\{x_i\leqslant s\wedge t\}}}{\left(\sum_{j=1}^n w_j(x_i;\hat{\boldsymbol{\theta}})Y_j(x_i)\right)^2} \end{pmatrix}.$$

Example 1 (Chi-square testing): The construction of chi-square goodness–of–fit tests for

right censored survival data has been studied for instance by Kim (1993) or Li and Doss (1993). Let us construct a chi-square test with d degrees of freedom with the following steps.

(a) Select a partition $0 < t_1 < \cdots < t_d < \tau \wedge \max_{1 \leqslant i \leqslant n} X_i$. Generally choosing data-driven t_i's is allowed (see Kim 1993). For instance we can set $t_i = S_{\text{net}}^{-1}(\hat{\pi} + (1-\hat{\pi})i/(d+1)|\hat{\boldsymbol{\theta}})$ for $i = 1, \ldots, d$ where for $t \geqslant 0$

$$S_{\text{net}}(t|\hat{\boldsymbol{\theta}}) = \exp\left(-\int_0^t \lambda_{\text{exc}}(s|\hat{\boldsymbol{\theta}})ds\right),$$

and $\hat{\pi}$ is the estimate of the cure rate π.

(b) Set $\hat{\mathcal{Y}} = (\mathfrak{D}_n(t_1), \ldots, \mathfrak{D}_n(t_d))^T$ be a $d \times 1$ real valued vector and $\hat{\Sigma} = (\hat{\sigma}_{ij})_{1 \leqslant i,j \leqslant d}$ the $d \times d$ real-valued matrix with entry (i,j) equal to $\hat{\sigma}_{ij} = \hat{\eta}(t_i, t_j)$ for $1 \leqslant i, j \leqslant d$. Then calculate $\hat{\mathcal{X}} = \hat{\mathcal{Y}}^T \hat{\Sigma}^{-1} \hat{\mathcal{Y}}$.

(c) Let $\alpha \in (0,1)$, if $\hat{\mathcal{X}} > \chi_d^2(1-\alpha)$ where $\chi_d^2(1-\alpha)$ is the (1-α)–quantile of a chi-square distribution with d degrees of freedom, then reject H_0 with an α–risk of type I.

4. Testing a parametric model in presence of covariates

4.1 *Maximum likelihood principle*

Let us write $\boldsymbol{\xi} = (\boldsymbol{\theta}^T, \boldsymbol{\beta}^T)^T \in \Theta \times \mathbb{R}^p = \Xi$ be the Euclidean parameter of the model (2). Defining the log–likelihood function by

$$\ell_n(\boldsymbol{\xi}) = \sum_{i=1}^n \left\{ \log\left(\lambda_{\text{pop}}^{(i)}(x_i + a_i) + e^{\boldsymbol{\beta}^T \boldsymbol{z}_i} \lambda_{\text{exc}}(x_i|\boldsymbol{\theta})\right) \delta_i - e^{\boldsymbol{\beta}^T \boldsymbol{z}_i} \Lambda_{\text{exc}}(x_i|\boldsymbol{\theta}) \right\}$$

the maximum likelihood estimator (MLE) satisfies

$$\hat{\boldsymbol{\xi}} = \arg\max_{\boldsymbol{\xi} \in \Xi} \ell_n(\boldsymbol{\xi}). \tag{8}$$

Again we note $\dot{f}(\boldsymbol{\theta}) = \frac{\partial f}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta})$, $\ddot{f}(\boldsymbol{\theta}) = \frac{\partial^2 f}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}(\boldsymbol{\theta})$, $\dot{f}(\boldsymbol{\xi}) = \frac{\partial f}{\partial \boldsymbol{\xi}}(\boldsymbol{\xi})$ and $\ddot{f}(\boldsymbol{\xi}) = \frac{\partial^2 f}{\partial \boldsymbol{\xi} \partial \boldsymbol{\xi}^T}(\boldsymbol{\xi})$. Because

$$\sqrt{n}\left(\hat{\boldsymbol{\xi}} - \boldsymbol{\xi}_0\right) = -\left[\frac{1}{n}\ddot{\ell}_n(\boldsymbol{\xi}^*)\right]^{-1} \frac{1}{\sqrt{n}}\dot{\ell}_n(\boldsymbol{\xi}_0),$$

where $\boldsymbol{\xi}^*$ lies between $\boldsymbol{\xi}_0$ and $\hat{\boldsymbol{\xi}}$, and

$$\dot{\ell}_n(\boldsymbol{\xi}_0) = \sum_{i=1}^{n} \int_0^{\tau} \begin{pmatrix} \frac{e^{\boldsymbol{\beta}_0^T \boldsymbol{z}_i} \dot{\lambda}_{\text{exc}}(x|\boldsymbol{\theta}_0)}{\lambda_{\text{pop}}^{(i)}(x+a_i)+e^{\boldsymbol{\beta}_0^T \boldsymbol{z}_i}\lambda_{\text{exc}}(x|\boldsymbol{\theta}_0)} \\ \frac{\boldsymbol{z}_i e^{\boldsymbol{\beta}_0^T \boldsymbol{z}_i} \lambda_{\text{exc}}(x|\boldsymbol{\theta}_0)}{\lambda_{\text{pop}}^{(i)}(x+a_i)+e^{\boldsymbol{\beta}_0^T \boldsymbol{z}_i}\lambda_{\text{exc}}(x|\boldsymbol{\theta}_0)} \end{pmatrix} dM_i(x|a_i, \boldsymbol{z}_i),$$

where τ is the study duration (in practice we set $\tau = +\infty$). Then, using standard martingale methods for counting processes we obtain under standard regularity conditions (see Andersen et al. 1993)

$$\sqrt{n}\left(\hat{\boldsymbol{\xi}} - \boldsymbol{\xi}_0\right) = I^{-1}(\boldsymbol{\xi}_0)\frac{1}{\sqrt{n}}\dot{\ell}_n(\boldsymbol{\xi}_0) + o_P(1), \tag{9}$$

where $I(\boldsymbol{\xi}_0)$ is consistently estimated by $\widehat{I}$ be defined by

$$\widehat{I} = \frac{1}{n}\sum_{i=1}^{n} \frac{\delta_i}{(\lambda_{\text{pop}}^{(i)}(x_i + a_i|\boldsymbol{z}_i) + e^{\hat{\boldsymbol{\beta}}^T \boldsymbol{z}_i}\lambda_{\text{exc}}(x_i|\hat{\boldsymbol{\theta}}))^2} \times \begin{pmatrix} \left(e^{\hat{\boldsymbol{\beta}}^T \boldsymbol{z}_i}\dot{\lambda}_{\text{exc}}(x_i|\hat{\boldsymbol{\theta}})\right)^{\otimes 2} & e^{2\hat{\boldsymbol{\beta}}^T \boldsymbol{z}_i}\dot{\lambda}_{\text{exc}}(x_i|\hat{\boldsymbol{\theta}})\boldsymbol{z}_i^T\lambda_{\text{exc}}(x_i|\hat{\boldsymbol{\theta}}) \\ \boldsymbol{z}_i e^{2\hat{\boldsymbol{\beta}}^T \boldsymbol{z}_i}\dot{\lambda}_{\text{exc}}^T(x_i|\hat{\boldsymbol{\theta}})\lambda_{\text{exc}}(x_i|\hat{\boldsymbol{\theta}}) & \boldsymbol{z}_i^{\otimes 2} e^{2\hat{\boldsymbol{\beta}}^T \boldsymbol{z}_i}\lambda_{\text{exc}}^2(x_i|\hat{\boldsymbol{\theta}}) \end{pmatrix}.$$

4.2 *Semiparametric estimation principle*

Following Sasieni (1996) we define the following weight functions

$$w_i(s;\hat{\boldsymbol{\xi}}) = \frac{e^{\hat{\boldsymbol{\beta}}\boldsymbol{z}_i}\lambda_{\text{exc}}(x|\hat{\boldsymbol{\theta}})}{\lambda_{\text{pop}}^{(i)}(x+a_i) + e^{\hat{\boldsymbol{\beta}}^T \boldsymbol{z}_i}\lambda_{\text{exc}}(x|\hat{\boldsymbol{\theta}})},$$

and the estimator of Λ_{exc} defined by

$$\hat{\Lambda}_{\text{exc}}(t|\hat{\boldsymbol{\xi}}) = \sum_{i=1}^{n}\int_0^t \frac{w_i(s;\hat{\boldsymbol{\xi}})}{\sum_{j=1}^n w_j(s;\hat{\boldsymbol{\xi}})Y_j(s)e^{\hat{\boldsymbol{\beta}}^T \boldsymbol{z}_j}}\left(dN_i(s) - Y_i(s)\lambda_{\text{pop}}^{(i)}(s+a_i)ds\right).$$

Remark 2: Note that combining the MLE with the semiparametric estimator provided by Sasieni allows to skip the step of the semiparametric estimation of $\boldsymbol{\beta}$ since under H_0 the estimator $\hat{\boldsymbol{\xi}}$ is $\sqrt{n}$–consistent.

4.3 *Testing a composite hypothesis*

As in Section 3.3 testing the following composite null hypothesis

$$H_0 : \lambda_{\text{exc}} \in \{\lambda_{\text{exc}}(\cdot|\boldsymbol{\theta}); \boldsymbol{\theta} \in \Theta \subset \mathbb{R}^p\},$$

requires to study the discrepancy process $\mathfrak{D}_n$ defined for $t \in [0, \tau]$ by

$$\mathfrak{D}_n(t) = \sqrt{n}\left(\hat{\Lambda}_{\text{exc}}(t|\hat{\boldsymbol{\xi}}) - \Lambda_{\text{exc}}(t|\hat{\boldsymbol{\theta}})\right),$$

where $\hat{\boldsymbol{\xi}} = (\hat{\boldsymbol{\theta}}^T, \hat{\boldsymbol{\beta}}^T)^T$ is the MLE defined by (8).

We show in the appendix that $\mathfrak{D}_n$ converges weakly to a centered gaussian process $\mathfrak{D}_\infty$ in $D([0, \tau])$ whose the covariance function η, defined for $(s, t) \in [0, \tau]^2$ by $\eta(s, t) = \mathbb{E}\left(\mathfrak{D}_\infty(s)\mathfrak{D}_\infty(t)\right)$, is consistently estimated by

$$\hat{\eta}(s, t) = \hat{v}(s)^T \hat{\gamma}(s, t)\hat{v}(t),$$

where

$$\hat{\gamma}(s, t) = \frac{1}{n}\sum_{i=1}^n \delta_i w_i^2(x_i; \hat{\boldsymbol{\xi}}) \begin{pmatrix} 1/\hat{\eta}^{(0)}(x_i) \\ \dot{\lambda}_{\text{exc}}(x_i|\hat{\boldsymbol{\theta}})/\lambda_{\text{exc}}(x_i|\hat{\boldsymbol{\theta}}) \\ \boldsymbol{z}_i \end{pmatrix}^{\otimes 2} \mathbb{1}_{\{x_i \leqslant s \wedge t\}},$$

$$\hat{v}(t) = \left(1, \left(\dot{\Lambda}_{\text{exc}}^T(t|\hat{\boldsymbol{\theta}}), \int_0^t \frac{\hat{\eta}^{(1)T}(u)}{\hat{\eta}^{(0)}(u)}\lambda_{\text{exc}}(u|\hat{\boldsymbol{\theta}})du\right)\widehat{I}^{-1}\right)^T$$

and

$$\hat{\eta}^{(0)}(t) = \sum_{i=1}^n w_i(t; \hat{\boldsymbol{\xi}})Y_i(t)e^{\hat{\boldsymbol{\beta}}^T \boldsymbol{z}_i} \quad \text{and} \quad \hat{\eta}^{(1)}(t) = \sum_{i=1}^n w_i(t; \hat{\boldsymbol{\xi}})Y_i(t)\boldsymbol{z}_i e^{\hat{\boldsymbol{\beta}}^T \boldsymbol{z}_i}.$$

EXAMPLE 2: Following the methody of Example 1 it is easy to construct a chi-square statistic for testing H_0.

5. Numerical study

5.1 *Simulation results*

We consider the model where the age at diagnostic is uniform on $\{20, \ldots, 79\}$, the population rate is Weibull with scale (resp. shape) parameter 90 (resp. 3), and the excess hazard rate is Weibull where the unknown parameter $\boldsymbol{\theta} = (\sigma, \gamma, \pi) = (5, 2, 0.5)$ has to be estimated from a sample of size n, where σ is the scale parameter, γ the shape parameter, and π is the cure rate. An example of the three risk functions for an individual diagnosed at 40 years is

given in Figure 1. The net survival function, the excess hazard rate as well as its gradient with respect to $\boldsymbol{\theta}$ are defined in the appendix. Because our goodness–of–fit procedure is based on the MLE we start with the performance of the MLE for several sample sizes in Table 1. First, by calculating the empirical mean (*mean*) of the 1000 estimates, we note that the bias decreases as the sample size n increases. Second we can see that the standard deviations of the 1000 estimates (*st.dev*) are very close to the empirical means of the standard deviation estimates ($\widehat{st.dev}$). Third going from $n = 250$ to $n = 1000$ diminishes the standard deviations by half, which means that the asymptotic regime is quickly reached. Last we note that whatever the value of n, the coverage probabilities (*cp*) are close to 0.95 which is another indicator of the good behavior of the MLE for moderate sample sizes. In Figure 2 are provided parametric (i.e. $\Lambda_{\text{exc}}(\cdot|\hat{\boldsymbol{\theta}})$) and nonparametric (i.e. $\hat{\Lambda}_{\text{exc}}(\cdot|\hat{\boldsymbol{\theta}})$) estimates of the excess cumulative hazard function based on one sample of size 1000 under H_0. We can see that these two estimates are close to the true cumulative excess risk function.

[Table 1 about here.]

[Figure 1 about here.]

[Figure 2 about here.]

[Figure 3 about here.]

In Figure 3 we compare the empirical cumulative distribution function (cdf) of the chi-square test of Exemple 1 (for 1000 simulated samples of size $n = 1000$ under H_0 and 10 degrees of freedom according the method defined in the same exemple) with the theoretical asymptotic chi-square distribution with 10 degrees of freedom. This gives empirical evidence that for such a sample size the empirical distribution of the test statistic $\hat{\mathcal{X}}$ is close to the expected asymptotic distribution. At the contrary, we can see in Figures 4 and 5 that if the true underlying distribution doesn't belong to the Weibull cure model family (here the

underlying distributions belong to the log–normal cure model family) then the values of the test statistic increase, and thus the empirical distribution of $\hat{\mathcal{X}}$ is shifted to the right. This leads to a rejection rate (power) of 15.7% in Figure 4 and 34.8% in Figure 5. Of course the larger is the distance between the alternative distribution (that is the distribution under H_1) and the Weibull cure model family, the larger is the power of the test statistic. We can also check that the power increases with the sample size.

[Figure 4 about here.]

[Figure 5 about here.]

5.2 *Real data set analysis*

We consider now a colon cancer data set provided by population based specialized cancer registry: Registre Bourguignon des Cancers Digestifs. The 5,772 patients newly diagnosed with colon cancer in the two administrative areas (département) of Côte–d'Or and Saône et Loire between 1995 and 2009 were included in the dataset. The available information for each individual are: sex, département of residence, age at diagnosis, time elapsed between diagnosis and the terminal event, and the background mortality rate (mortality rate in the general population) for an individual having the same characteristics (sex, département, age, calendar year). To illustrate our goodness–of–fit method we fit a Weibull cure model on four data sets obtained by crossing the variables sex (male or female) and département (21 or 71). Parameters of the net survival $S_{\text{net}}(t;\boldsymbol{\theta}) = \pi + (1-\pi)\exp(-(t/\sigma)^{\gamma})$ are $\boldsymbol{\theta} = (\sigma, \gamma, \pi)$, the MLE is denoted by $\hat{\boldsymbol{\theta}} = (\hat{\sigma}, \hat{\gamma}, \hat{\pi})$ and the estimated standard deviations are given within parenthesis. We can see that the sample sizes of the four samples vary from 1,010 to 1,951 which is close to the conditions we fixed in the simulation study. The statistical test we used is that of Example 1.

5.2.1 *Males from département 21.* The sample size is 1,531 and $\hat{\sigma} = 3.409$ (0.461), $\hat{\gamma} = 0.773$ (0.038) and $\hat{\pi} = 0.491$ (0.028). On Figure 6 are given MLE and nonparametric (NP) estimates of the cumulative excess risk function while on Figure 7 are given the p–values of the chi–square test as a function of the degrees of freedom. The mean of the p–values is 0.113, thus H_0 can be accepted. Here we do not reject the Weibull cure model, however we have to mention that 44% of the calculated p–values are less than 0.05 even if they are close to 0.05. Overall the relevance of this model remains unclear here. We can see also that the variations of the p–values may be quite important as the number of the degrees of freedom varies (which corresponds to a variation of the location of the partition when building the text statistic $\hat{\mathcal{X}}$). This is why it is important to perform several tests with several degrees of freedom.

[Figure 6 about here.]

5.2.2 *Males from département 71.* The sample size is 1,951 and $\hat{\sigma} = 3.766$ (0.414), $\hat{\gamma} = 0.802$ (0.035) and $\hat{\pi} = 0.452$ (0.026). On Figure 8 are given MLE and NP estimates of the cumulative excess risk function while on Figure 9 are given the p–values of the chi–square test as a function of the degrees of freedom. The mean of the p–values is 0.040, here the situation is clearer than in the previous case. Indeed very few p–values are larger than 0.05, and if not, they are close to 0.05 anyway. As a consequence H_0 is rejected, the Weibull cure model is not acceptable for this data set.

[Figure 7 about here.]

5.2.3 *Females from département 21.* The sample size is 1,010 and $\hat{\sigma} = 4.214$ (0.663), $\hat{\gamma} = 0.800$ (0.044) and $\hat{\pi} = 0.474$ (0.035). On Figure 10 are given MLE and NP estimates of the cumulative excess risk function while on Figure 11 are given the p–values of the chi–square test as a function of the degrees of freedom. The mean of these p–values is 0.489. Here

the situation is clear since all the calculated p–values are larger than 0.05. As a consequence H_0 can be accepted, thus on the one hand the Weibull cure model is acceptable, and on the other hand, the 95%–Wald confidence interval for the cure rate being $[0.405, 0.543]$ the presence of cure is clearly accepted.

[Figure 8 about here.]

5.2.4 *Females from département 71.* The sample size is 1,280 and $\hat{\sigma} = 3.172$ (0.300), $\hat{\gamma} = 0.881$ (0.040) and $\hat{\pi} = 0.520$ (0.021). On Figure 12 are given MLE and NP estimates of the cumulative excess risk function while on Figure 13 are given the p–values of the chi–square test as a function of the degrees of freedom. The mean of these p–values is 0.212, again the situation is clear since all the calculated p–values are larger than 0.05. As a consequence H_0 can be accepted, thus on the one hand the Weibull cure model is acceptable, and on the other hand, the 95%–Wald confidence interval for the cure rate being $[0.479, 0, 561]$ the presence of cure is clearly accepted.

[Figure 9 about here.]

It is important to notice that the earlier (with respect to time t) the separation between the MLE $\Lambda_{\text{exc}}(t|\hat{\boldsymbol{\theta}})$ and the NP estimator $\hat{\Lambda}_{\text{exc}}(t|\hat{\boldsymbol{\theta}})$, the greater the risk of rejection (see Figures 6 and 8 for males versus Figures 10 and 12 for females). For large values of t, this difference is not so important since the variance of the discrepancy process increases with$t$. The previous results also show that there is a low impact of the administrative areas (*département*) with respect to the gender. Indeed the Weibull cure model is either acceptable with small p–values for males of *département* 21 or clear rejected for males of *département* 71, while for females of both *départements* the Weibull cure model is clearly accepted.

6. Concluding remarks

We have proposed a general goodness–of–fit procedure which may be applied to any regular parametric model. As an example of this goodness–of–fit procedure we developed some chi-square type tests which the behavior has been studied through a simulation study. These tests have been successfully applied to colon cancer data. It is important to emphasize that when the null hypothesis is not rejected it means that the assumption that the cumulative excess risk function belongs to a specified parametric cure model (here Weibull cure model) is acceptable while otherwise this the whole parametric model which is rejected. Thus when the null hypothesis is rejected it does not mean that the cure assumption is rejected, it only means that the specified parametric distribution family is not adapted to the data.

Chi-square type tests are only one example of test building. Indeed there is a large range of classical alternative statistics like for instance Kolmogorov–Smirnov, Cramér–von Mises, or Anderson-Darling statistics. However, although these statistics may be more powerful than the chi-square statistic we proposed, their asymptotic distributions are not free of the unknown parameters in general, which prevents to derive a p–value in a easy way. However, it is possible to derive such a p–value using a bootstrap approach. This will be the subject of a future work.

Acknowledgements

This work was partially funded by grant from INCa (INCa SHSESP 16-064). Olayidé Boussari was supported by the Fondation ARC pour la recherche sur le cancer and by a French Government grant managed by the French National Research Agency under the program "Investissements d'Avenir" [reference ANR-11-LABX-0021]. Data were provided by the Registre Bourguignon des Cancers Digestifs.

References

Aalen O.O. (1980). A model for non-parametric regression analysis of counting processes. In *Mathematical Statistics and Probability Theory, Klonecki W, Kozek A, Rosinski J (eds), Lecture Notes in Statistics, Springer: New York* **2**, 1–25.

Andersen, P.K., Borgan, O., Gill, R.D., and Keiding, N. (1993). *Statistical Models Based on Counting Processes*, Berlin and New York: Springer.

Andersen, P.K. and Vaeth, M. (1989). Simple parametric and nonparametric models for excess and relative mortality, *Biometrics* **45**, 523–535.

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

Cronin, K. A. and Feuer, E.J. (2000). Cumulative cause?specific mortality for cancer patients in the presence of other causes: a crude analogue of relative survival, *Statistics in medicine* **19(13)**, 1729–1740.

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(8)**, 775–786.

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

Grafféo, N., Castell, F., Belot, A., Giorgi, R. (2016). A log-rank-type test to compare net survival distributions, *Biometrics* **72(3)**, 760–769.

Greenwood, P.E. and Nikulin, M.S., E. (1996). *A Guide to Chi-Squared Testing*, John Wiley & Sons, New York.

Kannan, N., Kundu, D., Nair, P., and Tripathi, R.C. (2010). The generalized exponential cure rate model with covariates, *Journal of Applied Statistics* **37(10)**, 1625–1636.

Khmaladze, E. (2013). Note on distribution free testing for discrete distributions, *Annals of*

Statistics **41(6)**, 2979–2993.

Kim, J.H. (1993). Chi-square goodness-of-fit tests for randomly censored data, *Annals of Statistics* **21(3)**, 1621–1639.

Li, G. Doss, H. (1993). Generalized Pearson-Fisher chi-square goodness-of-fit tests, with applications to models with life history data, *Annals of Statistics* **21(2)**, 772–797.

Kodre, A.R. and Perme, M.P. (2013). Informative censoring in relative survival. *Statistics in Medicine* **32**, 4791–4802.

Lambert, P.C., Dickman, P.W., and Rutherford, M.J. (2015). Comparison of different approaches to estimating age standardized net survival, *BMC Medical Research Methodology* **15(1)**, art. no. 64.

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

Sasieni, P.D. (1996). Proportional excess hazards, *Biometrika* **83**, 127–141.

Seppä, K., Hakulinen, T., and Pokhrel, A. (2016). Choosing the net survival method for cancer survival estimation, *European Journal of Cancer* **51(9)**, 1123–1129.

Seppä, K., Hakulinen, T., Läärä, E., and Pitkäniemi, J. (2016). Comparing net survival estimators of cancer patients, *Statistics in Medicine* **35(11)**, 1866–1879.

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

Yu, X.Q., De Angelis, R., Andersson, T.M.L., Lambert, P.C., O'Connell D.L., and Dickman, P.W. (2013). Estimating the proportion cured of cancer: Some practical advice for users, *Cancer Epidemiology* **37**, 836–842.

Appendix

Asymptotics of the discrepancy process of Section 3.3

First, if $\boldsymbol{\theta}_0$ is the true value of $\boldsymbol{\theta}$ under H_0 note that for $t \in [0,\tau]$ we have

$$\mathfrak{D}_n(t) = \sqrt{n}\left(\Lambda_{\text{exc}}(t|\hat{\boldsymbol{\theta}}) - \Lambda_{\text{exc}}(t|\boldsymbol{\theta}_0)\right) - \sqrt{n}\left(\hat{\Lambda}_{\text{exc}}(t|\hat{\boldsymbol{\theta}}) - \Lambda_{\text{exc}}(t|\boldsymbol{\theta}_0)\right),$$

and in addition from (4) we can write

$$\sqrt{n}\left(\Lambda_{\text{exc}}(t|\hat{\boldsymbol{\theta}}) - \Lambda_{\text{exc}}(t|\boldsymbol{\theta}_0)\right) = \dot{\Lambda}^T_{\text{exc}}(t|\boldsymbol{\theta}_0) I^{-1}(\boldsymbol{\theta}_0) n^{-1/2} \dot{\ell}_n(\boldsymbol{\theta}_0) + o_P(1),$$

and using the martingale property in (3) and Lenglart's inequality (see Andersen et al. 1993) we derive

$$\begin{aligned}&\sqrt{n}\left(\hat{\Lambda}_{\text{exc}}(t|\hat{\boldsymbol{\theta}}) - \Lambda_{\text{exc}}(t|\boldsymbol{\theta}_0)\right)\\ &= n^{1/2}\textstyle\sum_{i=1}^n \int_0^t \frac{w_i(s;\boldsymbol{\theta}_0)}{\sum_{j=1}^n w_j(s;\boldsymbol{\theta}_0) Y_j(s)} dM_i(s|a_i) + o_P(1).\end{aligned}$$

Thus the asymptotic behavior of $\mathfrak{D}_n$ is obtained by studying the asymptotic behavior of $n^{-1/2}\dot{\ell}_n(\boldsymbol{\theta}_0)$ and $t \mapsto n^{1/2}\sum_{i=1}^n \int_0^t \frac{w_i(s;\boldsymbol{\theta}_0)}{\sum_{j=1}^n w_j(s;\boldsymbol{\theta}_0) Y_j(s)} dM_i(s|a_i)$ on $[0,\tau]$. First by the Rebolledo theorem (see Andersen et al. 1993) we show that

$$\begin{aligned}\mathfrak{U}_n(t) &= \begin{pmatrix} n^{-1/2}\dot{\ell}_n(\boldsymbol{\theta}_0) \\ n^{1/2}\sum_{i=1}^n \int_0^t \frac{w_i(s;\boldsymbol{\theta}_0)}{\sum_{j=1}^n w_j(s;\boldsymbol{\theta}_0) Y_j(s)} dM_i(s|a_i) \end{pmatrix}\\ &= n^{-1/2}\sum_{i=1}^n \int_0^\tau w_i(s) \begin{pmatrix} \frac{\dot{\lambda}_{\text{exc}}(s|\boldsymbol{\theta}_0)}{\lambda^{(i)}_{\text{pop}}(a_i+s)} \\ \frac{n 1_{\{s\leqslant t\}}}{\sum_{j=1}^n w_j(s;\boldsymbol{\theta}_0) Y_j(s)} \end{pmatrix} dM_i(s|a_i),\end{aligned}$$

converges weakly in $D([0,\tau])$ to a centered Gaussian process $\mathfrak{U}_\infty$ whose the covariance function $\gamma(s,t)$ is consistently approximated by $\gamma_n(s,t)$ defined in Section 3.3. Since

$$\mathfrak{D}_n(t) = \left(\dot{\Lambda}^T_{\text{exc}}(t|\boldsymbol{\theta}_0) I^{-1}(\boldsymbol{\theta}_0), -1\right) \mathfrak{U}_n(t) + o_P(1)$$

we conclude that $\mathfrak{D}_n$ converges weakly to the centered Gaussian $\mathfrak{D}_\infty$ in $D([0,\tau])$.

Asymptotics of the discrepancy process of Section 4.3

We have

$$\begin{aligned}
\mathfrak{D}_n(t) &= \sqrt{n}\sum_{i=1}^{n}\int_0^t \frac{w_i(s;\hat{\boldsymbol{\xi}})}{\sum_{j=1}^n w_j(s;\hat{\boldsymbol{\xi}})Y_j(s)e^{\boldsymbol{\beta}_n^T\boldsymbol{z}_j}} dM_i(s|a_i,\boldsymbol{z}_i) \\
&\quad +\sqrt{n}\sum_{i=1}^{n}\int_0^t \frac{w_i(s;\hat{\boldsymbol{\xi}})Y_i(s)}{\sum_{j=1}^n w_j(s;\hat{\boldsymbol{\xi}})Y_j(s)e^{\boldsymbol{\beta}_n^T\boldsymbol{z}_j}} \left(e^{\boldsymbol{\beta}_0^T\boldsymbol{z}_i}\lambda_{\text{exc}}(s|\boldsymbol{\theta}_0) - e^{\boldsymbol{\beta}_n^T\boldsymbol{z}_i}\lambda_{\text{exc}}(s|\hat{\boldsymbol{\theta}})\right) \\
&= \sqrt{n}\sum_{i=1}^{n}\int_0^t \frac{w_i(s;\boldsymbol{\xi}_0)}{\sum_{j=1}^n w_j(s;\boldsymbol{\xi}_0)Y_j(s)e^{\boldsymbol{\beta}_0^T\boldsymbol{z}_j}} dM_i(s|a_i,\boldsymbol{z}_i) \\
&\quad + \begin{pmatrix} \dot{\Lambda}_{\text{exc}}(t|\boldsymbol{\theta}_0) \\ \int_0^t \frac{\eta_n^{(1)}(s)}{\eta_n^{(0)}(s)}ds \end{pmatrix}^T I_n^{-1}\frac{\ell_n(\boldsymbol{\xi}_0)}{\sqrt{n}} + o_P(1) \\
&= \left(1, \left(\dot{\Lambda}_{\text{exc}}^T(t|\boldsymbol{\theta}_0), \int_0^t \frac{\eta_n^{(1)T}(s)}{\eta_n^{(0)}(s)}ds\right) I_n^{-1}\right) \times \\
&\quad \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\int_0^t w_i(s;\boldsymbol{\xi}_0) \begin{pmatrix} 1/\eta_n^{(0)}(s) \\ \dot{\lambda}_{\text{exc}}(s|\boldsymbol{\theta}_0)/\lambda_{\text{exc}}(s|\boldsymbol{\theta}_0) \\ \boldsymbol{z}_i \end{pmatrix} dM_i(s|a_i,\boldsymbol{z}_i) + o_P(1),
\end{aligned}$$

where the second equality is obtained by a Taylor expansion of the right most term of the right hand side of the first equality. The convergence of $\mathfrak{D}_n$ to the centered Gaussian process $\mathfrak{D}_\infty$ in $D([0,\tau])$ results from the Rebolledo central limit theorem.

Weibull excess hazard model

For $\boldsymbol{\theta} = (\sigma,\gamma,\pi) \in \Theta = (0,+\infty)^2 \times [0,1]$, the excess (or net) survival function is defined for $x \in \mathbb{R}^+$ by

$$S_{\text{net}}(x|\boldsymbol{\theta}) = \pi + (1-\pi)\exp(-(x/\sigma)^\gamma).$$

The cumulative excess hazard is defined by $\Lambda_{\text{exc}}(x|\boldsymbol{\theta}) = -\log\left(S_{\text{net}}(x|\boldsymbol{\theta})\right)$ and the excess hazard is therefore equal to

$$\lambda_{\text{exc}}(x|\boldsymbol{\theta}) = \frac{(1-\pi)\frac{\gamma x^{\gamma-1}}{\sigma^\gamma}\exp(-(x/\sigma)^\gamma)}{\pi + (1-\pi)\exp(-(x/\sigma)^\gamma)}.$$

Thus we have

$$\dot{\lambda}_{\text{exc}}(x|\boldsymbol{\theta}) = \lambda(x|\boldsymbol{\theta}) \begin{pmatrix} -\frac{\gamma}{\sigma} + (\frac{\gamma}{\sigma}) \left(\frac{x}{\sigma}\right)^{\gamma} - \lambda(x|\boldsymbol{\theta}) \left(\frac{x}{\sigma}\right) \\ \frac{1}{\gamma} + \log\left(\frac{x}{\sigma}\right) \left(1 - \left(\frac{x}{\sigma}\right)^{\gamma} + \lambda(x|\boldsymbol{\theta}) \frac{x}{\gamma}\right) \\ -1/(1-\pi) - \left(1 - \exp\left(-(x/\sigma)^{\gamma}\right)\right) / S_{\text{net}}(x|\boldsymbol{\theta}) \end{pmatrix}.$$

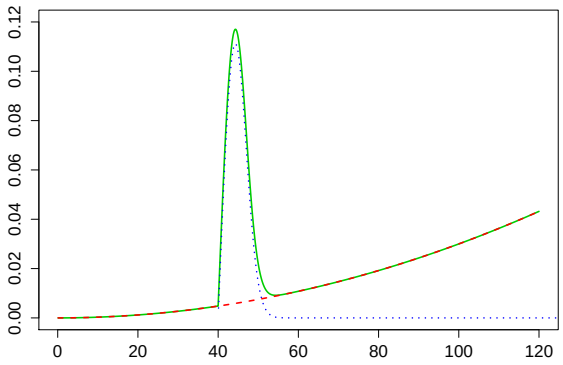

Figure 1. Observed risk function of an individual diagnosed at 40 years (green solid line), population risk function (red dashed line), and excess risk function (blue dotted line).

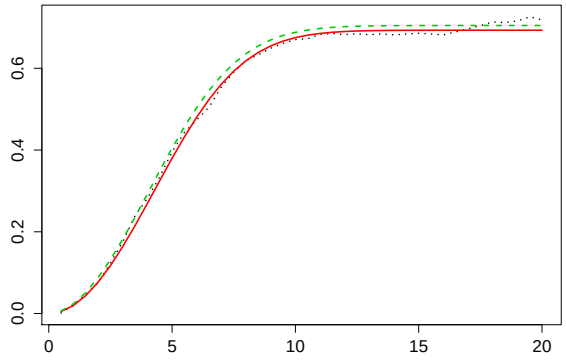

Figure 2. True cumulative excess hazard rate (solid red line) with parametric (dashed green line) and nonparametric (dotted black line) estimates based on a sample of size $n = 1000$.

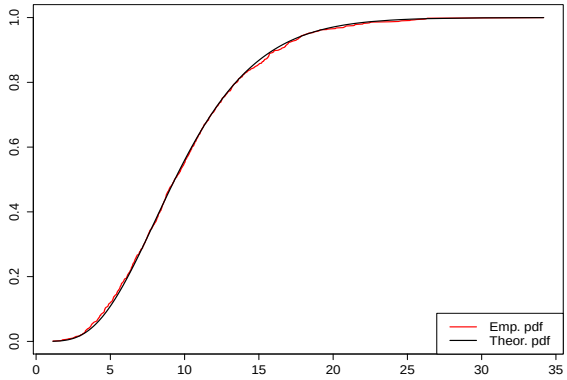

Figure 3. Empirical cdf of the chi-square statistic in Exemple 1 for 10 degrees of freedom and 1000 simulated samples of size $n = 1000$ in black versus the true pdf of the chi-square distribution with 10 degrees of freedom. The H_0 rejection rate is of 5,0%.

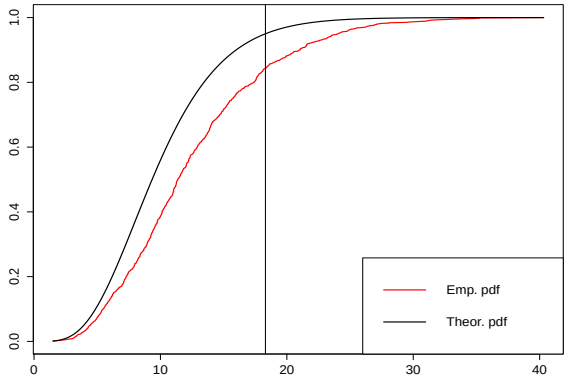

Figure 4. Empirical cdf of the chi-square statistic in Exemple 1 for 10 degrees of freedom and 1000 simulated samples of size $n = 1000$ in black versus the pdf of the chi-square distribution with 10 degrees of freedom. The true underlying distribution is log-normal with mean 2 and standard deviation 1.5. The H_0 rejection rate is of 15,7%.

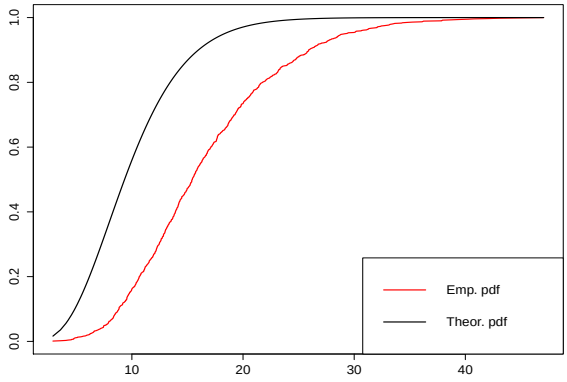

Figure 5. Empirical cdf of the chi-square statistic in Exemple 1 for 10 degrees of freedom and 1000 simulated samples of size $n = 1000$ in black versus the pdf of the chi-square distribution with 10 degrees of freedom. The true underlying distribution is log-normal with mean 2 and standard deviation 0.5. The H_0 rejection rate is of 34,8%.

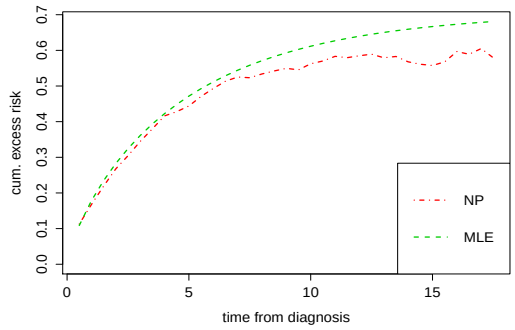

Figure 6. Parametric cumulative excess risk estimation (MLE) vs nonparametric estimation (NP).

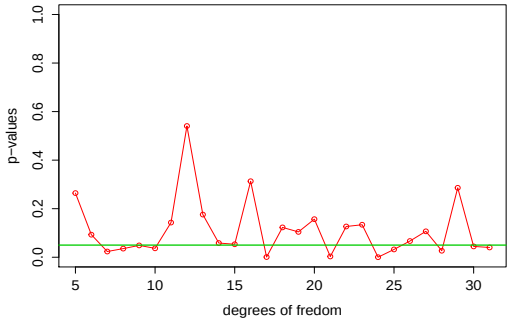

Figure 7. p–values of the chi-square tests as a function of the degrees of freedom.

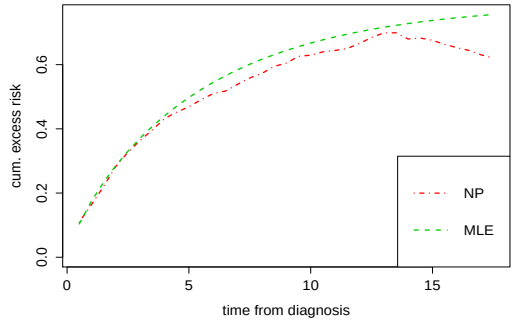

Figure 8. Parametric cumulative excess risk estimation (MLE) vs nonparametric estimation (NP).

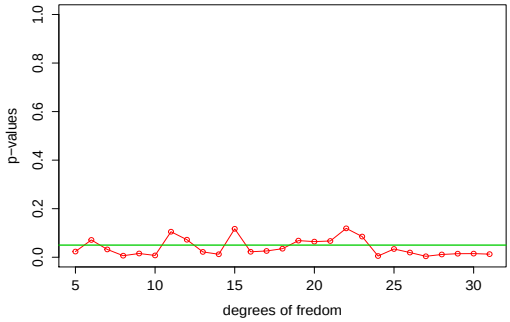

Figure 9. p–values of the chi-square tests as a function of the degrees of freedom.

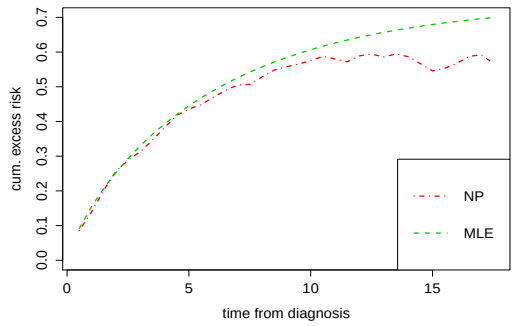

Figure 10. Parametric cumulative excess risk estimation (MLE) vs nonparametric estimation (NP).

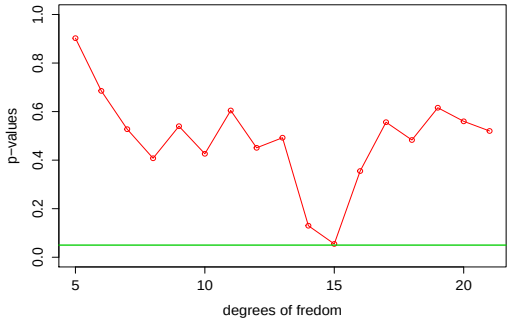

Figure 11. p–values of the chi-square tests as a function of the degrees of freedom.

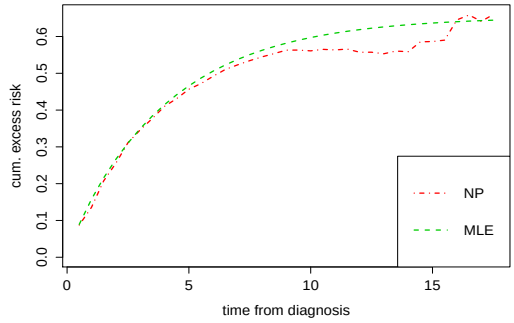

Figure 12. Parametric cumulative excess risk estimation (MLE) vs nonparametric estimation (NP).

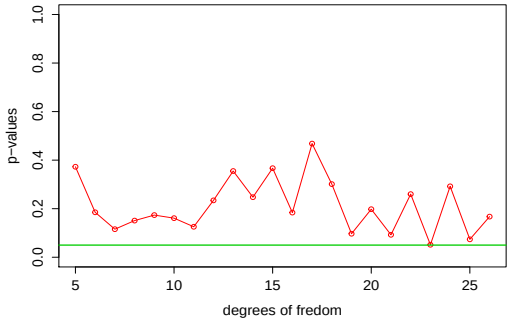

Figure 13. p–values of the chi-square tests as a function of the degrees of freedom.

Table 1

MLE performance for various sample sizes based on 1000 simulated samples: mean is the empirical mean, st.dev is the empirical standard deviation, $\widehat{st.dev}$ is the mean of the estimated standard deviations and cp is the 95% coverage probabilities. The censoring rate is about 55%.

n	indicator	$\sigma = 5$	$\gamma = 2$	$\pi = 0.5$
250	*mean*	5.016	2.027	0.499
	st.dev	0.352	0.209	0.041
	$\widehat{st.dev}$	0.340	0.206	0.041
	cp	0.931	0.957	0.944
500	*mean*	4.989	2.019	0.500
	st.dev	0.231	0.141	0.030
	$\widehat{st.dev}$	0.236	0.144	0.029
	cp	0.954	0.940	0.936
1000	*mean*	5.004	2.007	0.500
	st.dev	0.169	0.102	0.020
	$\widehat{st.dev}$	0.168	0.101	0.020
	cp	0.946	0.947	0.950