Goftte: A R package for assessing goodness-of-fit in proportional (sub) distributions hazards regression models

P. Sfumato, T. Filleron, R. Giorgi, R.J. Cook, J.M. Boher

Goftte: a R package for assessing goodness-of-fit in proportional (sub) distributions hazards regression models

Sfumato P [1], Filleron T [2], Giorgi R[3,5], Cook RJ[4], Boher JM[1,5]

(1) Institut Paoli-Calmettes, Biostatistics Unit, Marseille, France.
(2) Institut Claudius Regaud-IUCT-O, Biostatistics Unit, Toulouse, France.
(3) Hopital Timone, BioSTIC, Marseille, France.
(4) Department of Statistics and Actuarial Science, University of Waterloo, 200 University Avenue West, Waterloo, ON, N2L 3G1, Canada.
(5) Aix Marseille Univ, INSERM, IRD, SESSTIM, Marseille, France.

*Corresponding Author:
Jean-Marie Boher
Institut Paoli Calmettes, Aix Marseille Univ,
232 Boulevard Sainte-Marguerite
13009 Marseille
France
boher.jm@ipc.unicancer.fr
Tél : +33 (0) 491 223 876

Disclosure: The authors declare that they have no conflict of interest.

Abstract

Background and objective: In this paper, we introduce a new R package *goftte* for goodness-of-fit assessment based on cumulative sums of model residuals useful for checking key assumptions in the Cox regression and Fine and Gray regression models.

Methods: Monte-Carlo methods are used to approximate the null distribution of cumulative sums of model residuals. To limit the computational burden, the main routines used to approximate the null distributions are implemented in a parallel C++ programming environment. Numerical studies are carried out to evaluate the empirical type I error rates of the different testing procedures. The package and the documentation are available to users from CRAN R repositories.

Results: Results from simulation studies suggested that all statistical tests implemented in *goftte* yielded excellent control of the type I error rate even with modest sample sizes with high censoring rates.

Conclusions: As compared to other R packages *goftte* provides new useful method for testing functionals, such *as* Anderson-Darling type test statistics for checking assumptions about proportional (sub-) distribution hazards. Approximations for the null distributions of test statistics have been validated through simulation experiments. Future releases will provide similar tools for checking model assumptions in multiplicative intensity models for recurrent data. The package may help to spread the use of recent advocated goodness-of-fit techniques in semiparametric regression for time-to-event data.

Keywords: goodness-of-fit; survival data, competing risks; cumulative sums of residuals; goftte.

1. Introduction

In biomedical research, Cox regression analysis [1] is the most commonly used statistical method for investigating the association between a survival time and one or more predictor variables. Other areas of application are finance [2], criminology [3], reliability [4] and industrial engineering ([5],[6]). Typically Cox regression is based on a semi-parametric proportional hazards (PH) model relying on two fundamental assumptions: (i) a time-invariant multiplicative effect of covariates on the hazard rate function, and (ii) a log-linear effect of covariates on the hazard rate function. Several authors have studied the implications of model misspecification in Cox regression (Lagakos and Schoenfeld [7]; Lagakos [8]; Struthers and Kalbfleisch [9]; Lin [10]; Gerds and Schumacher [11]). The Fine and Gray (FG) proportional sub-distribution hazards model is often viewed a weighted Cox regression model using inverse probability of censoring weighting commonly adopted for analyzing the direct effect of covariates on a specific cumulative incidence function (CIF) [12]. Latouche [13] and Grambauer [14] studied the implications of model misspecification in the FG proportional sub-distribution hazards (PSH) model.

Lin et al. [15] proposed model diagnostic tools and Kolmogorov-Smirnov (KS) goodness-of-fit (GOF) tests based on cumulative sums of martingale residuals for checking the PH, the linearity assumptions, and the link function in the Cox model, and showed how to approximate the asymptotic distribution of the different test statistics under the null using Monte-Carlo methods. Recently Li et al. [16] extended these methods to propose KS goodness-of-fit tests for checking the FG model assumptions. Empirical studies of the finite-sample properties suggested that the KS goodness-of-fit test for the PH or PSH assumptions based on cumulative sums of residuals are more sensitive to general non-PH or non-PSH alternatives than other tests assuming specific time-varying covariates effects (Kvaloy and Neef [17], Li [16]). Moreover, Kvaloy and Neef illustrated the good frequency properties of Cramer-von Mises (CvM) and Anderson-Darling (AD) type statistics against general non-PH alternatives.

Two packages in R propose KS-type GOF tests for checking the model assumptions in the Cox model and the FG model, respectively the *gof* [18] and *crskdiag* [19] packages. Each package has its own advantages and limitations. The package *gof* allows to test for PH using the maximal deviation and a CvM type statistic, but provides no test for checking the

functional form of a covariate. The package *crskdiag* implements GOF tests based on cumulative sums of model residuals that extend the class of GOF tests first pioneered by Lin for the Cox model to the FG model [16]. In both packages, the p values are estimated using a similar Monte Carlo method, but slightly different from the method originally proposed by Lin in the Cox model. Neither *gof* nor *crskdiag* implements AD test statistics to test for PH and PSH against general alternatives.

The main purpose of this manuscript is to give an overview of the functionalities of a new R package *goftte* [21] implementing KS-type goodness-of-fit tests based on cumulative sums of residuals for the Cox model and the FG model, including Cramer-von-Mises (CvM) and Anderson-Darling (AD) type test statistics for checking PH and PSH assumptions.

The remainder of this manuscript is organized as follows. In Section 2 we present the different test statistics implemented by the package and the different methods available to approximate their asymptotic null distributions. Section 3 provides a basic description of the core functions and main subroutines, including details of the programming language and classes. Results from simulation studies assessing the empirical type I error rates of the different tests are reported in Section 4. In Section 5 a case study is given to illustrate the different package functionalities. We end with concluding remarks in Section 6.

2. Methods

2.1. Data and models

Let T denote the time to failure, C the time to censoring, $\varepsilon \in \{1, \cdots, K\}$ the failure cause, and Z a vector of p individual covariates. We assume the data consist of a collection of n independent observations $(X_i, \delta_i\ \varepsilon_i, Z_i)$, where $X_i = T_i \wedge C_i$ is the time to failure or censoring observed for the i^{th} patient, $\delta_i = 1$ if $T_i \leq C_i$ and 0 otherwise. Assuming k is a failure cause of specific interest, we note $F(t|Z)$ the cause-specific cumulative incidence given a set of individual covariates,

$$F(t|Z) = P[T \leq t, \varepsilon = k|Z].$$

Two regression models are considered, the Cox model and the FG model that simply assume that the (sub)hazard rate function $\gamma(t|Z) = -\ dF(t|Z)/\big(1 - F(t|Z)\big)$ takes the form,

$$\gamma(t|Z) = \gamma_0(t)exp(\beta' Z)$$

where $\gamma_0(t)$ denotes an unspecified baseline (sub)hazard function and $\beta = (\beta_1, \cdots, \beta_K)$ a set of unknown regression coefficients. Further let $N_i(t) = I(X_i \leq t, \varepsilon_i = k, \delta_i = 0)$, $Y_i(t)$ =1 $-N_i(t-)$ where $I(.)$ denotes the indicator function, and let $\hat{G}(t)$ be the Kaplan-Meier probability estimator of remaining uncensored up to time t. Assuming (T_i, ε_i) and C_i are conditionally independent given Z_i, a consistent estimate $\hat{\beta}$ for β is obtained by solving the equation $U(\beta) = 0$ where

$$U(\beta) = \sum_{i=1}^{n} \int_0^{\infty} \left[Z_i - \frac{S^{(1)}(\beta, t)}{S^{(0)}(\beta, t)} \right] dN_i(t)$$

and $S^{(j)}(\beta, t) = \sum_{i=1}^{n} W_i(t) Y_i(t) Z_i^j e^{\beta' Z_i}$ for $j = 0{,}1$ with $W_i(t) = I(X_i \leq t)$ if $\varepsilon_i = k$ or $\delta_i = 0$, $W_i(t) = \hat{G}(t)/\hat{G}(t \wedge X_i)$ otherwise ([1],[12]).

2.2. Goodness-of fit test statistics

The package *goftte* implements goodness-of-fit (GOF) tests based on cumulative sums $\hat{S}(t, z)$ for detecting departures from PH/PSH assumptions and deviations from the assumed functional form for a given covariate , where $\hat{S}(t, z) = S\big(\hat{\beta}, t, z\big)$ and

$$S(\beta, t, z) = \sum_{i=1}^{n} \int_0^{t} \left[f(Z_i) I(Z_i \leq z) - \frac{S^{(1)}(\beta, t, z)}{S^{(0)}(\beta, t)} \right] dN_i(t)$$

or equivalently

$$\hat{S}(t,z) = \sum_{i=1}^{n} f(Z_i) I(Z_i \leq z) W_i(t) \widehat{M}_i(t).$$

Here $f(.)$ denotes a known smooth function, $S^{(1)}(\beta,t,z) = \sum_{i=1}^{n} W_i(t) Y_i(t) f(Z_i) I(Z_i \leq z) e^{\beta' Z_i}$, $\widehat{M}_i(t) = N_i(t) - \int_0^t Y_i(u) e^{\hat{\beta}' Z_i} d\hat{\Gamma}(u)$ with $d\hat{\Gamma}(t) = S^{(0)}\left(\hat{\beta},t\right)^{-1} \sum_{i=1}^{n} dN_i(t)$ and $z = \left(z_1,\cdots,z_p\right) \in \mathbb{R}^p$. Setting $f(Z_i) = Z_i$ and $z = (\infty,\cdots,\infty)$, $\hat{S}(t,z)$=$U\left(\hat{\beta},t\right)$ where $U(\beta,t)$ denotes the score process $U(\beta,t) = \sum_{i=1}^{n} \int_0^t \left[Z_i - \frac{S^{(1)}(\beta,u)}{S^{(0)}(\beta,u)}\right] dN_i(u)$. In the sequel, j will denote the index of a covariate of interest, $U_j(\beta,t)$ the j^{th} component of the score process, $\hat{I}_{j,j}(t) = -\frac{\partial}{\partial \beta_j} U_j(\beta,t)_{|\beta=\hat{\beta}}$ and $\delta(u) = \hat{I}_{j,j}(u)/\hat{I}_{j,j}(\infty)$.

The new package makes available three measures of deviation for the user to detect departures from PH/PSH assumptions in covariates. The first is based on KS statistics as in Lin et al. [15] and Li et al. [16]:

$$KS_j = \sup_t \left|U_j\left(\hat{\beta},t\right)\right|.$$

The CvM and AD type statistics are also available as they have been recommended for detecting general non PH alternatives [17]. These have the following form:

$$CvM_j = \int_0^{\infty} U_j\left(\hat{\beta},u\right)^2 d\delta(u)$$

$$AD_j = \int_0^{\infty} \frac{U_j\left(\hat{\beta},u\right)^2}{\delta(u)(1-\delta(u))} d\delta(u)$$

As in *gof* and *crskdiag* packages, GOF tests based on maximal deviation statistics $KS_j = \sup_{z_j} \left|\hat{S}_j\left(z_j\right)\right|$, $\hat{S}_j(z_j) = \sum_{i=1}^{n} I\left(Z_{ij} \leq z_j\right) \widehat{M}_i(\infty)$ with $z_j \in \mathbb{R}$, were made available for assessing the functional form adequacy assumed for covariate Z_{ij}.

2.3. Methods for approximating the null distribution

Let $N_i^C(t) = I(X_i \leq t, \delta_i = 1)$, $G(t) = P[C > t]$, $\Lambda^C(t) = -\partial C(t)/C(t)$, $dM_i^C(t) = dN_i^C(t) - I(X_i \geq t) d\Lambda^C(t)$ and $dM_i(\beta,t) = dN_i(t) - Y_i(u) exp(\beta' Z_i) \gamma_0(u) du$. For any fixed t and z, it can be shown using the same probabilistic arguments as outlined in Lin et al. [15] and Fine and Gray [12] that the statistic $n^{-\frac{1}{2}} S(\beta,t,z)$ is asymptotically equivalent under the model assumptions to a sum $n^{-\frac{1}{2}} \sum_{i=1}^{n} \Psi_i(t,z)$, where

$$\Psi_i(t,z) = \int_0^t \left\{ f(Z_i) - \frac{\tilde{s}^{(1)}(\beta,u,z)}{\tilde{s}^{(0)}(\beta,u,z)} \right\} \tilde{w}_i(u) dM_i(\beta,u) + \int_0^t \frac{\tilde{q}(\beta,u,z)}{\tilde{\pi}(u)} dM_i^C(u)$$

with $\tilde{s}^{(k)}(\beta,t,z) = \lim_{n\to\infty} n^{-1}S^{(k)}(\beta,t,z)$ for $k = 0,1$, $\tilde{\pi}(t) = \lim_{n\to\infty} n^{-1}\sum_{j=1}^{n} Y_j(t)$ and $\tilde{q}(\beta,t,z) = -\lim_{n\to\infty} n^{-1}\sum_{i=1}^{n}\int_0^{\infty}\left[f(Z_i) - \frac{\tilde{s}^{(1)}(\beta,u,z)}{\tilde{s}^{(0)}(\beta,u,z)}\right] I(u \geq t > X_i)\tilde{w}_i(u)dM_i(u)$. Let $\tilde{h}(t,z) = \lim_{n\to\infty} n^{-1}\widehat{H}(t,z)$ with $\widehat{H}(t,z) = -\frac{\partial}{\partial\beta}S(\beta,t,z)_{|\beta=\widehat{\beta}}$ and $\tilde{\imath} = \lim_{n\to\infty} n^{-1}\hat{I}$, with $\hat{I} = -\frac{\partial}{\partial\beta}U(\beta)_{|\beta=\widehat{\beta}}$. Again for any fixed t and z, it can be shown using a Taylor's series expansion of $n^{-\frac{1}{2}}\hat{S}(t,z)$ around β (Lin [15], Boher [22]) that $n^{-\frac{1}{2}}\hat{S}(t,z)$ is asymptotically equivalent to the sum $n^{-\frac{1}{2}}\tilde{S}(t,z) = n^{-\frac{1}{2}}\sum_{i=1}^{n}\tilde{S}_i(t,z)$, where $\tilde{S}_i(t,z) = \Psi_i(t,z) - \tilde{h}(t,z)\tilde{\imath}^{-1}\Sigma_i$ with $\Sigma_i = \Psi_i(\infty,\infty)$ when letting $f(z) = z$. As under the null, $n^{-\frac{1}{2}}\tilde{S}(t,z)$ is essentially a sum of mean-zero independent variables; we approximate the limiting distribution of $n^{-\frac{1}{2}}\hat{S}(t,z)$ by a mean zero Gaussian process.

Two Monte Carlo procedures have been implemented to empirically generate samples representing the null distribution. Following Lin [15] and Liu [23], approximations to the null distributions are obtained by repeatedly sampling random normal deviates $\{G_i^m\}_{i=1,\ldots,n}$ to draw M realizations $n^{-\frac{1}{2}}\tilde{S}_m(t,z) = n^{-\frac{1}{2}}\sum_{i=1}^{n}\hat{S}_i(t,z)G_i^m$. The values $\hat{S}_i(t,z)$ are obtained by replacing the unknown limiting values $\tilde{h}(t,z)$ and $\tilde{\imath}$ in $\tilde{S}_i(t,z)$ with consistent sample estimates, $\hat{S}_i(t,z) = \widehat{\Psi}_i(t,z) - \widehat{H}(t,z)\hat{I}^{-1}\hat{\Sigma}_i$ and setting

$$\widehat{\Psi_\iota}(t,z) = \delta_i I(X_i \leq t)\left[f(Z_i) - \frac{S^{(1)}(\hat{\beta},X_i,z)}{S^{(0)}(\hat{\beta},X_i,z)}\right] + (1-\delta_i)I(X_i \leq t)\frac{Q(\hat{\beta},X_i,z)}{\sum_{j=1}^{n} Y_j(X_i)}$$

or

$$\widehat{\Psi_\iota}(t,z) = \int_0^t\left[f(Z_i) - \frac{S^{(1)}(\hat{\beta},u,z)}{S^{(0)}(\hat{\beta},u,z)}\right]W_j(u)d\widehat{M}_j(u) + \int_0^t \frac{Q(\hat{\beta},u,z)}{\sum_{j=1}^{n} Y_j(u)}d\widehat{M}_j^C(u)$$

with

$$Q(\hat{\beta},t,z) = -\sum_{j=1}^{n} I(t > X_j)\int_t^{\infty}\left\{f(Z_j) - \frac{S^{(1)}(\hat{\beta},u,z)}{S^{(0)}(\hat{\beta},u,z)}\right\}W_j(u)d\widehat{M}_j(u),$$

$d\widehat{M}_i^C(t) = dN_i^C(t) - Y_i(t)\frac{\sum_{j=1}^{n} dN_j^C(u)}{\sum_{j=1}^{n} Y_j(u)}$ and $\widehat{\Sigma_\iota} = \widehat{\Psi_\iota}(\infty,\infty)$ when letting $f(z) = z$. P values are then derived from the randomly perturbed processes $\left\{n^{-\frac{1}{2}}\tilde{S}_m(t,z)\right\}_{m=1,\ldots,M}$ as the sample proportion of simulated KS type statistics greater than the observed value of the test statistic.

3. Package description

As Monte Carlo method is characterized by long lead times, the main computations were carried out using primitive functions using C++ code, and possibly parallelized in the C++ environment. Results are made available to the user by two main generic S3 function, fcov() and prop(), including methods for R objects of class coxph{*survival*} [24], cph{*rms*} [25], crr{*cmprsk*} [26]. Compiled binaries are available for Linux, macOS, Solaris and Windows platforms. The package with detailed documentation is distributed freely by CRAN (Comprehensive R Archive Network). Table 1 summarizes the different testing functionalities made available in the R packages *goftte*, *gof* and *crskdiag*.

4. Type I error

We conducted numerical studies to evaluate the level I properties of GOF tests with level $\alpha = 5\%$ checking departures from PH, PSH or functional linear form in covariate Z under both regression settings. We first checked the Type I error control and compared the performance of *goftte* to other packages. Repeated sets of independent data were generated from a model with a single covariate Z, $Z \sim N(0,1)$, cumulative incidence function for a type 1 failure of interest given by

$$\mathrm{F}(t|Z) = 1 - [1 - a(1 - e^{-t})]^{e^{0.3Z}}$$

and various sample sizes (N=50,100,200). Independent uncensored data were obtained by first drawing each failure type with probability $\Pr(\varepsilon = 1|Z) = 1 - [1 - a]^{e^{0.3Z}}$ and exact failure time from the distribution $\mathrm{F}(t|Z)$ if failure of type 1 or from an exponential distribution with rate $Exp(e^{-0.5Z})$. To assess the performance under both regression settings, the value for a was set to 1 and to a scalar <1 to yield an overall type 1 failure rate under PSH equals to 0.66. Censored data were obtained by drawing independent censoring times from an exponential distribution chosen to yield 0.15, 0.30 and 0.50 overall crude censoring rates. For all scenarios, a total of 2000 repetitions was used to estimate with precision around 1% the true rejection rates of 5%. In Table 2 are reported the empirical sizes for all new testing functionalities using the two methods for p-value approximation: the first called the Lin method and the last implemented in other packages [20] and referred to as the Liu method. Overall all proposed tests maintained the empirical error rates close to the 5% Type I nominal level even in mild to moderate sample sizes. In the few cases where the differences between the observed and theoretical type I error rates exceeded 1%, the Lin method showed some conservatism as opposed to the Liu method who exhibited some anti-conservatism. The Lin method was set as the default method to approximate the p values because of its apparent conservatism in situations where error rates of type I were outside the expected range, except for checking the PSH assumptions where the Liu method better maintained the error rates close to the 5% nominal level. According to these default settings, we evaluated the empirical type I error rates between goftte and *gof* (cf Supplementary Table 1)., *goftte* and *crskdiag* (cf Supplementary Table 2). With respect to KS and CvM statistics, we observed some empirical error rates higher than expected with *gof* as compared to *goftte.* These results are consistent with our findings from Table 2 suggesting that the Lui method rejects too frequently. Note that the CvM statistic for checking the PH assumptions in *gof* differs slighthly from described in paragraph 2.2 with $\delta(t) = dt$ [20]. Overall the rejection rates for KS statistics reported with *crskdiag* were higher than those

reported using *goftte*. In particular, the empirical Type I error rates for PSH with *crskdiag* increased up to 7% with moderate to high censoring rates.

5. Examples

As an illustrative example, we consider the dataset “pbc” available in the R package *survival* [24]. The data consists of times to occurrence of a first event (death, transplantation) or censoring in a set of 418 patients with primary biliary cirrhosis of the liver followed at the Mayo Clinic, of whom 161 died and 25 had liver transplant first.

Following Dickson et al. [27] and Li et al. [9], we used multivariate PH and PSH regressions models with terms for edema, age at diagnosis, and log-transformed serum albumin, serum bilirubin and prothrombin time, to model the overall survival and the cumulative incidence of death without transplant. To assess the PH or PSH model assumptions, we first estimated the regression coefficients using the functions coxph{survival} and crr{*cmprsk}* and test for PH/PSH using the functions prop.coxph() and prop.crr() with default resampling methods implemented by *goftte* package. Note that the default number of independent realizations (R=1000) used to approximate p values was over-ridden (R=20000) to improve accuracy. The generic function plot() plotted the observed and the first 50 simulated paths for each component of the score process vs time (see Figure 1 and Figure 2). For each covariate the output displays three tests for checking proportionality and one for functional form. In addition, the Monte-Carlo approximation method selected and the number of simulations used are reported. All the results with default resampling method are summarized in Table 3.

```
# Estimation of regression coefficients,
data(pbc,package=”survival”)
attach(pbc)
fit.coxph <- coxph(Surv(time,status==2) ~ age+edema+log(bili) + log(albumin) + log(protime), ties="breslow")
fit.crr<-crr(ftime=time,fstatus=status,cov1=cbind(age,edema,log(bili),log(albumin),log(protime)),failcode=2)

# Test for PH/PSH assumptions
coxph.prop<-prop(model=fit.coxph,variable=c("age","edema","log(bilirubin)","log(albumin)","log(protime)"),
R=20000, plots=50, seed=10)
crr.prop<-
prop(model=fit.crr,ftime=time,fstatus=status,cov1=cbind(age,edema,log(bili),log(albumin),log(protime)),failcode=2
,variable=c("age","edema","log(bilirubin)","log(albumin)","log(protime)"),R=20000, plots=50,seed=10)

# Print results of tests for PH for log prothrombin time in output windows
print(coxph.prop,idx=5); plot(coxph.prop, idx=5)

Rejection p-values associated to Lin’s approximation for proportional hazards assumption
- - -
Kolmogorov-Smirnov-test : p-value<0.001
Cramer-von-Mises-test : p-value<0.001
Anderson-Darling-test : p-value<0.001
Based on 20000 realizations. Cumulated residuals ordered by log(protime)-variable.
```

```
# Print results of tests for PSH for edema in output windows
print(crr.prop, idx=2); plot(crr.prop, idx=2)
```

Rejection p-values *associated to Liu's approximation for proportional subdistribution hazards assumption*
- - -
Kolmogorov-Smirnov-test : p-value=0.024
Cramer-von-Mises-test : p-value=0.0437
Anderson-Darling-test : p-value=0.04995
Based on 20000 realizations. Cum*ulated residuals* ordered by edema-variable.

KS and CvM test statistics rejected the PH assumptions for log prothrombin time and edema. The same conclusions hold using te package *gof*. The default KS p-values for PH were almost identical, only the differences for CvM statistics were more pronounced because of the lack of homogeneity of the definitions between the packages. For example, using *gof and goftte*, the p value of KS statistics for Edema were respectively p=0.0210 and p=0.0218, while the p value of CvM statistics for Edema were respectively p=0.0244 and p=0.0474. Regarding PSH assumption, the KS test rejected the null hypothesis at the 5% significance level for Edema and log prothrombin time whatever the package used. The CvM and AD statistics as implemented in *gofttte* led to the same conclusions. There were some noticeable differences in p values for PSH between *crskdiag* and *goftte*. As an example, the p value of the KS test for PSH for log(Bilirubin) was p=0.0757 using *crskdiag* and p=0.2868 using *goftte*. As more than 50% of the patients were censored, this large difference can be attributed to the lack of Type I error control observed with *crskdiag* (Supplementary table 2).
Several authors have emphasized the inadequacy of the linear functional form for untransformed bilirubin in Cox regression (Lin & Wei [10]; Leon & Tsai [28]; Kvaloy & Neef [17]) and suggested a model with logarithm transformation of bilirubin as an alternative [15],[16] For illustrative purposes, we estimated the regression coefficients assuming untransformed bilirubin with other covariates left unchanged, and then checked the adequacy of untransformed bilirubin in both regression settings.

```
# Estimation of regression coefficients assuming untransformed bilirubin,
fit2.coxph <- coxph(Surv(time,status==2) ~ age+edema+bili + log(albumin) + log(protime), ties="breslow")
fit2.crr<-crr(ftime=time,fstatus=status,cov1=cbind(age,edema,bili,log(albumin),log(protime)),failcode=2)


# Test for untransformed bilirubin functional form
coxph2.fcov <- fcov(model=fit2.coxph, variable=c("age", "edema","bilirubin","log(albumin)","log(protime)"),
R=20000, plots=50, seed=10)

crr2.fcov <-
fcov(model=fit2.crr,ftime=time,fstatus=status,cov1=cbind(age,edema,bili,log(albumin),log(protime)),failcode=2,var
iable=c("age","edema","bilirubin","log(albumin)","log(protime)"),R=20000, plots=50,seed=10)


# Print summary information of GOF test for untransformed serum bilirubin
```

```
print(coxph2.fcov,idx=3)
```

```
Rejection p-values associated to Lin's approximation for covariate(s) functional form assumption
- - -
Kolmogorov-Smirnov-test : p-value<0.001
Based on 20000 realizations. Cumulated residuals ordered by bilirubin-variable.
- - -
```

```
print(crr2.fcov, idx=3)
```

```
Rejection p-values associated to Lin's approximation for covariate(s) functional form assumption
- - -
Kolmogorov-Smirnov-test : p-value<0.001
Based on 20000 realizations. Cumulated residuals ordered by bilirubin-variable.
- - -
```

As expected, the KS test for untransformed bilirubin rejected the null hypothesis in both regression models $(p < 0.001)$. Deviations of the observed cumulative sums process from simulated paths under the null are depicted in Figure 3 (PH model) and Figure 4 (PSH model). After substituting bilirubin by log(Bilirubin), tests assessing the functional form of covariate remained non statistically significants (Table 3). We did not noticeable difference between the p values derived from *goftte* and *crskdiag*.

6. Discussion

The Cox model [1] and the Fine and Gray model [12] are the most widely used statistical regression models for analyzing right-censored survival and competing risks. Two different R packages, respectively the package *gof* [18] and the package *crskdiag* [19], implemented omnibus KS tests checking the Cox and the Fine and Gray models with cumulative sums of residuals. The package *goftte* [21] can be viewed as add-on to the package *gof* allowing assessing misspecification in the functional form of a covariate as proposed in Lin et al. [15] and detecting general non PH alternatives using AD type statistic. It also provides diagnostic tools and goodness-of-fit methods based on cumulative sums of residuals for checking the Fine and Gray model assumptions, including CvM and AD statistics against general non PSH alternatives not available from *crskdiag* [19],.

Two resampling methods have been implemented in *goftte* to derive the p values, a first method pioneered in Lin &al. for checking the model assumptions in the Cox model [15], and a similar method implemented in *gof* and *crskdiag* [20]. Numerical results showed that all different test statistics have empirical rejection rates close to the expected 5% error rate. In few cases where the empirical rejection rates were outside of the expected range, the Lin method exhibited some conservative properties as opposed to the Liu method.

In Cox regression settings, unlike *gof*, the package *goftte* allows to test the functional form of a covariate and to test against general non PH alternatives using AD test statistics [17]. In FG regression settings, the package *goftte* allows to test against general non PSH alternatives using different test statistics, including CvM and AD type statistics. More important, our numerical results showed that KS statistics in *crskdiag* failed to control the Type I error rate under moderate and heavy censoring.

Despite the computational burden, the use of the C++ environment for parallelization of the code allows efficient computing and thereby a relatively short execution time enabling one to increase the number of iterations in the Monte Carlo step in order to get same p values rounded up to 3 digits with different seeds in large samples.

7. Conclusion

This paper introduces a new package *goftte* for R to check model assumptions in semi-parametric Cox-type regression for standard failure time data and competing risks data. The package available from CRAN R-like repositories performs model diagnostics and goodness-of-fit tests based on cumulative sums of residuals to detect departures from proportionality assumptions or functional misspecifications in regression covariates. In the present manuscript we explored the ability of our package to identify departures from modeling assumptions through real applications where violations in non-proportionality assumptions or functional misspecifications in specific regression covariate have been clearly evidenced in the literature.

The package *goftte* for R propose new testing functionalities, including AD test statistics against general non PH and non PSH alternatives. Unlike *crskdiag,* the package *goftte* controls the Type I error under moderate and heavy censorings. Empirical studies aimed to assess the type II level properties and the sensitivity of the different testing capabilities under specific alternatives will be reported in a future manuscript.

The development of this package work was originally oriented to medical applications. The scope of applications could be extended to other fields where Cox-type regression is now becoming a more popular analytical tool. In particular in industrial engineering studies where the Cox model assumptions are often checked graphically [30].

7.1. Availability

The package can be load on CRAN repository at https://CRAN.R-project.org/package=goftte.

8. Conflict of interest

The authors have no conflict of interest to disclose regarding this work.

9. Acknowledgement

This work is funded by the Institute National du Cancer (Inca), grant number 2012-081.

10. References

[1] D. R. Cox, “Regression Models and Life-Tables,” *J. R. Stat. Soc. Ser. B Methodol.*, vol. 34, no. 2, pp. 187–220, 1972.

[2] M.-C. Lee, “Business Bankruptcy Prediction Based on Survival Analysis Approach,” *Int. J. Comput. Sci. Inf. Technol.*, vol. 6, no. 2, pp. 103–119, Apr. 2014.

[3] J.-B. Pingault, S. M. Côté, E. Lacourse, C. Galéra, F. Vitaro, and R. E. Tremblay, “Childhood Hyperactivity, Physical Aggression and Criminality: A 19-Year Prospective Population-Based Study,” *PLoS ONE*, vol. 8, no. 5, p. e62594, May 2013.

[4] D. Weatherburn, “The effect of prison on adult re-offending,” *Contemp. Issues Crime Justice*, no. 143, Apr. 2010.

[5] X.-S. Si, “An Adaptive Prognostic Approach via Nonlinear Degradation Modeling: Application to Battery Data,” *IEEE Trans. Ind. Electron.*, vol. 62, no. 8, pp. 5082–5096, Aug. 2015.

[6] Z. Zhang, X. Si, C. Hu, and Y. Lei, “Degradation data analysis and remaining useful life estimation: A review on Wiener-process-based methods,” *Eur. J. Oper. Res.*, vol. 271, no. 3, pp. 775–796, Dec. 2018.

[7] S. W. Lagakos and D. A. Schoenfeld, “Properties of proportional-hazards score tests under misspecified regression models,” *Biometrics*, vol. 40, no. 4, pp. 1037–1048, Dec. 1984.

[8] S. W. Lagakos, “The Loss in Efficiency from Misspecifying Covariates in Proportional Hazards Regression Models,” *Biometrika*, vol. 75, no. 1, p. 156, Mar. 1988.

[9] C. A. Struthers and J. D. Kalbfleisch, “Misspecified Proportional Hazard Models,” *Biometrika*, vol. 73, no. 2, p. 363, Aug. 1986.

[10] D. Y. Lin and L. J. Wei, “The Robust Inference for the Cox Proportional Hazards Model,” *J. Am. Stat. Assoc.*, vol. 84, no. 408, p. 1074, Dec. 1989.

[11] T. A. Gerds and M. Schumacher, “On functional misspecification of covariates in the Cox regression model,” *Biometrika*, vol. 88, no. 2, pp. 572–580, Jun. 2001.

[12] J. P. Fine and R. J. Gray, “A Proportional Hazards Model for the Subdistribution of a Competing Risk,” *J. Am. Stat. Assoc.*, vol. 94, no. 446, p. 496, Jun. 1999.

[13] A. Latouche, V. Boisson, S. Chevret, and R. Porcher, “Misspecified regression model for the subdistribution hazard of a competing risk,” *Stat. Med.*, vol. 26, no. 5, pp. 965–974, Feb. 2007.

[14] N. Grambauer, M. Schumacher, and J. Beyersmann, “Proportional subdistribution hazards modeling offers a summary analysis, even if misspecified,” *Stat. Med.*, vol. 29, no. 7–8, pp. 875–884, Mar. 2010.

[15] D. Y. Lin, L. J. Wei, and Z. Ying, “Checking the Cox Model with Cumulative Sums of Martingale-Based Residuals,” *Biometrika*, vol. 80, no. 3, p. 557, Sep. 1993.

[16] J. Li, T. H. Scheike, and M.-J. Zhang, “Checking Fine and Gray subdistribution hazards model with cumulative sums of residuals,” *Lifetime Data Anal.*, vol. 21, no. 2, pp. 197–217, Apr. 2015.

[17] J. T. Kvaløy and L. R. Neef, “Tests for the proportional intensity assumption based on the score process,” *Lifetime Data Anal.*, vol. 10, no. 2, pp. 139–157, Jun. 2004.

[18] K. K. Holst, *gof: Model-diagnostics based on cumulative residuals. R package version 0.9.1*. 2014.

[19] J. Li, *crskdiag: Diagnostics for Fine and Gray Model. R package version 1.0.1*. 2016.

[20] K. Holst, “Model Diagnostics Based on Cumulative Residuals: The R-package gof.,” 2015.

[21] P. Sfumato and J.-M. Boher, *goftte: Goodness-of-Fit for Time-to-Event Data. R package version 1.0.5*. 2017.

[22] J.-M. Boher, T. Filleron, R. Giorgi, A. Kramar, and R. J. Cook, “Goodness-of-fit test for monotone proportional subdistribution hazards assumptions based on weighted residuals,” *Stat. Med.*, vol. 36, no. 2, pp. 362–377, Jan. 2017.

[23] M. Liu, W. Lu, and Y. Shao, “A Monte Carlo approach for change-point detection in the Cox proportional hazards model,” *Stat. Med.*, vol. 27, no. 19, pp. 3894–3909, Aug. 2008.
[24] T. Therneau, *A Package for Survival Analysis in S. version 2.43-3*. 2018.
[25] F. E. Harrell, *rms: Regression Modeling Strategies. R package version 4.3-0*. 2015.
[26] B. Gray, *cmprsk: Subdistribution Analysis of Competing Risks. R package version 2.2-7*. 2014.
[27] E. R. Dickson, P. M. Grambsch, T. R. Fleming, L. D. Fisher, and A. Langworthy, “Prognosis in primary biliary cirrhosis: model for decision making,” *Hepatol. Baltim. Md*, vol. 10, no. 1, pp. 1–7, Jul. 1989.
[28] L. F. León and C.-L. Tsai, “Functional form diagnostics for Cox’s proportional hazards model,” *Biometrics*, vol. 60, no. 1, pp. 75–84, Mar. 2004.
[29] Z. Li, S. Zhou, S. Choubey, and C. Sievenpiper, “Failure event prediction using the Cox proportional hazard model driven by frequent failure signatures,” *IIE Trans.*, vol. 39, no. 3, pp. 303–315, Mar. 2007.
[30] S. Madeira, P. Infante, and F. Didelet, “Use of survival models in a refinery,” *Revstat Stat. J.*, vol. 11, no. 1, pp. 45–65, Mar. 2013.

Table legends

Table 1: Comparison between the packages main features

Table 2: Empirical sizes (%) for 2000 simulations of goftte tests under (H0) based on cumulative sums of residuals (R=1000).

Table 3: Rejection p-values for goftte on Mayo Clinic Primary Biliary Cirrhosis (PBC) data under proportionality and functional form assumptions (R=20000).

Supplementary Table 1: Empirical sizes (%) in a Cox model under proportional hazard assumption for 2000 simulations and n=200 for *goftte* and *gof* tests based on cumulative sums of residuals (R=1000).

Supplementary Table 2: Empirical sizes (%) in a Fine and Gray model under proportional sub-distribution hazard and functional form assumptions for 2000 simulations and n=200 for *goftte* and *crskdiag* tests based on cumulative sums of residuals (R=1000).

Figure legends

Figure 1: Checking PH assumption for log(protime)

Figure 2: Checking PSH model assumption for Edema

Figure 3: Checking functional form of Bilirubin in PH model

Figure 4: Checking functional form of Bilirubin in PSH model

Figure 1. Checking PH assumption for log(protime)

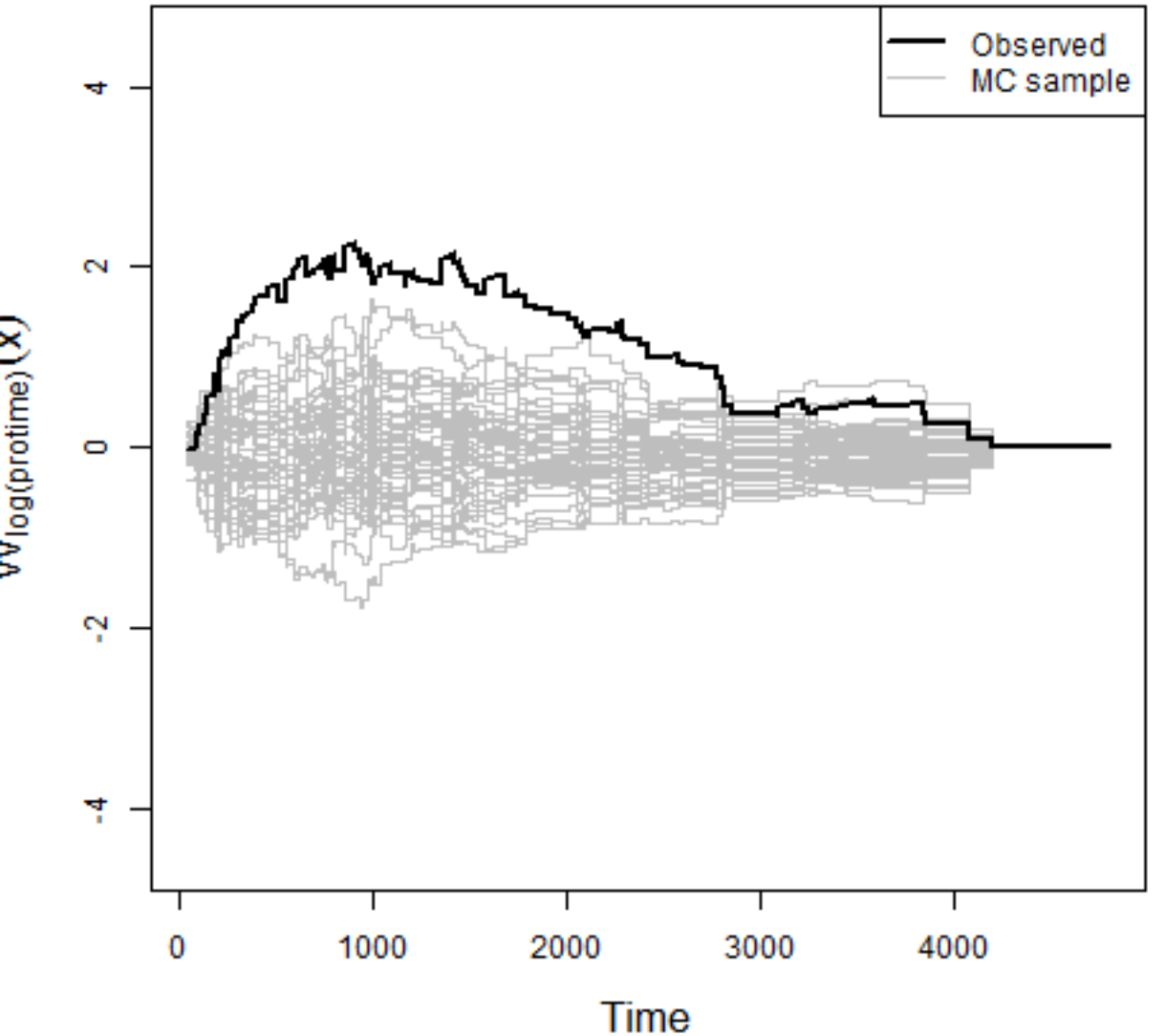

Figure 2. Checking PSH assumption for Edema

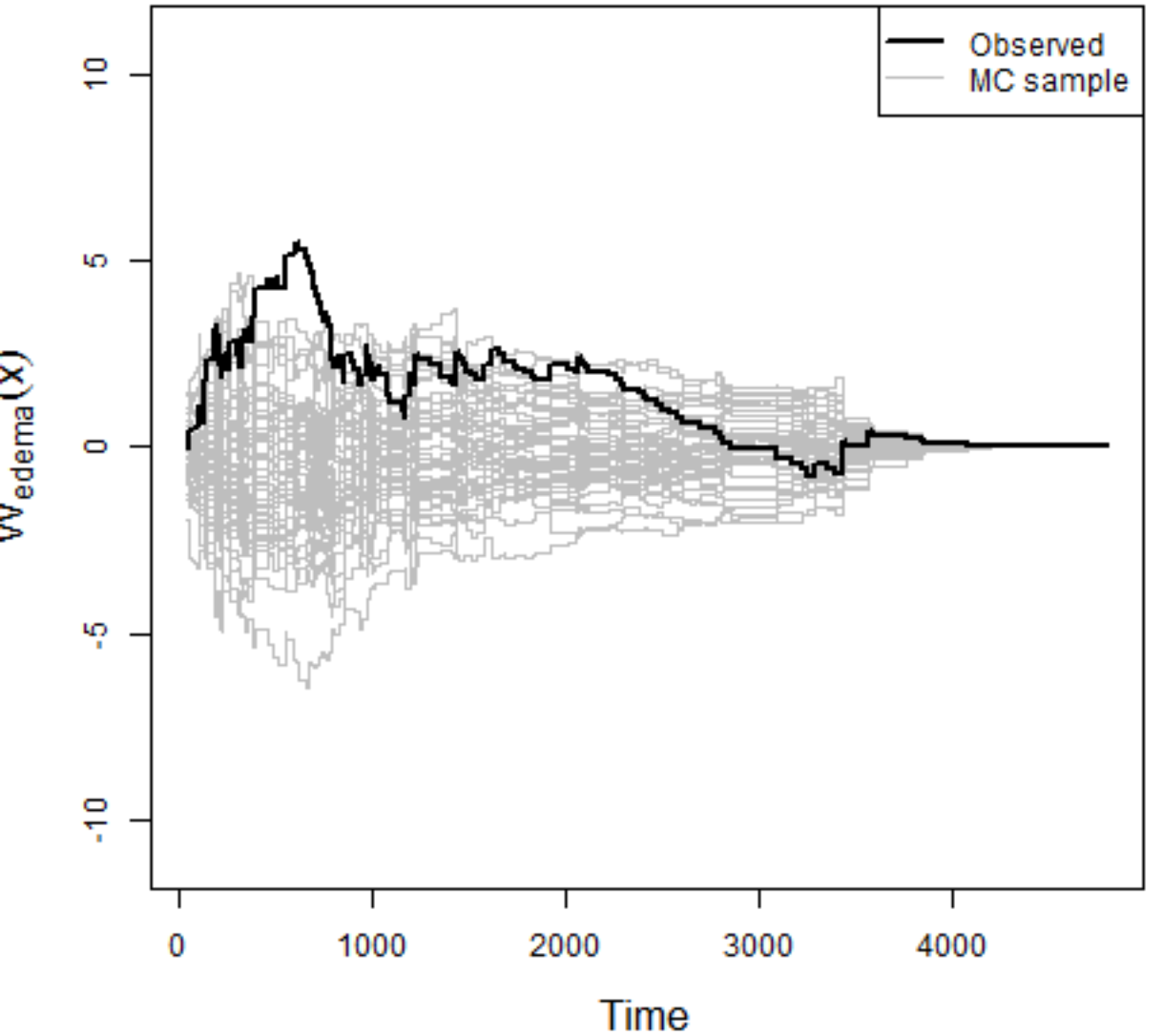

Figure 3. Checking functional form of Bilirubin in PH model

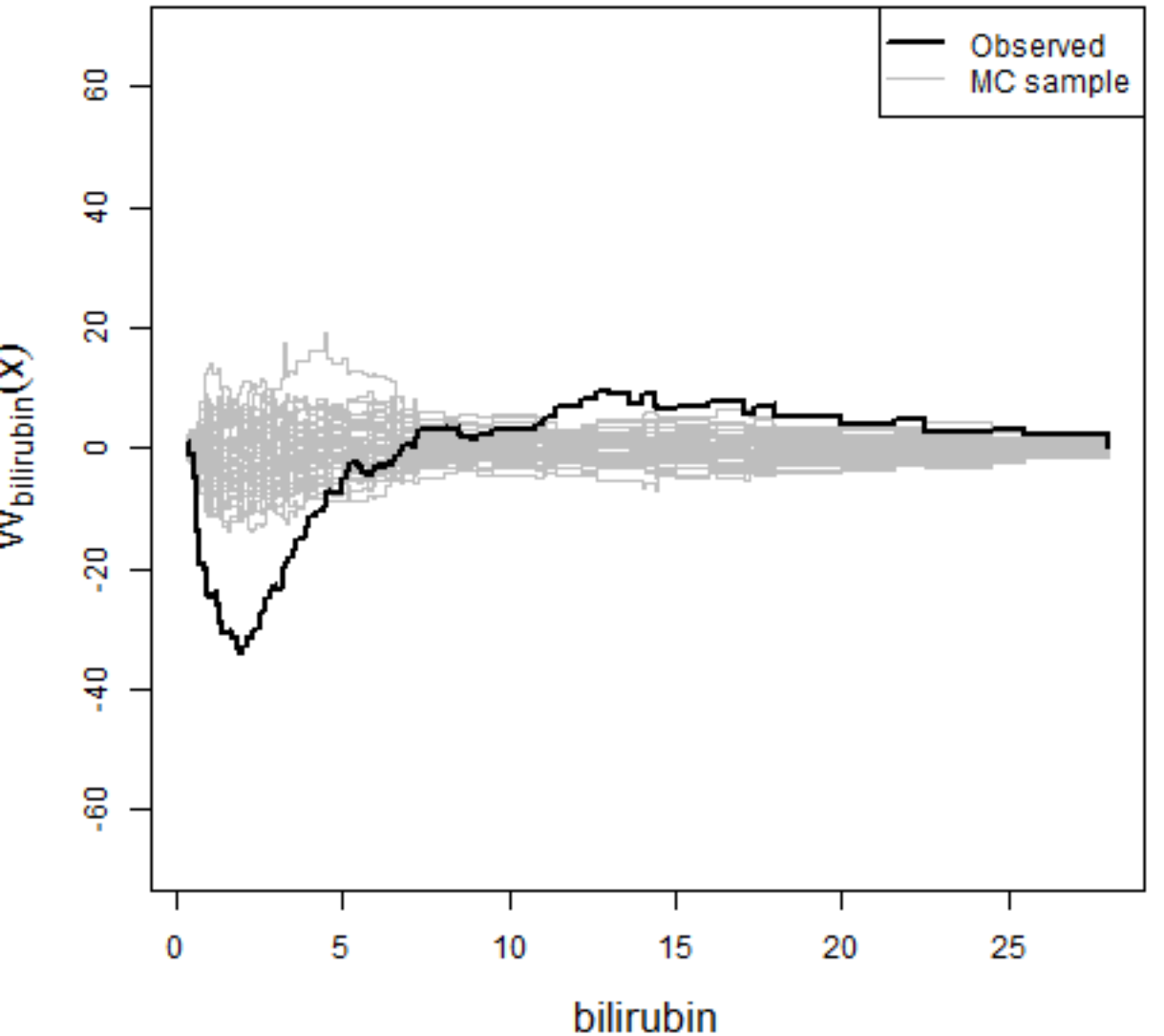

Figure 4. Checking functional form of Bilirubin in PSH model

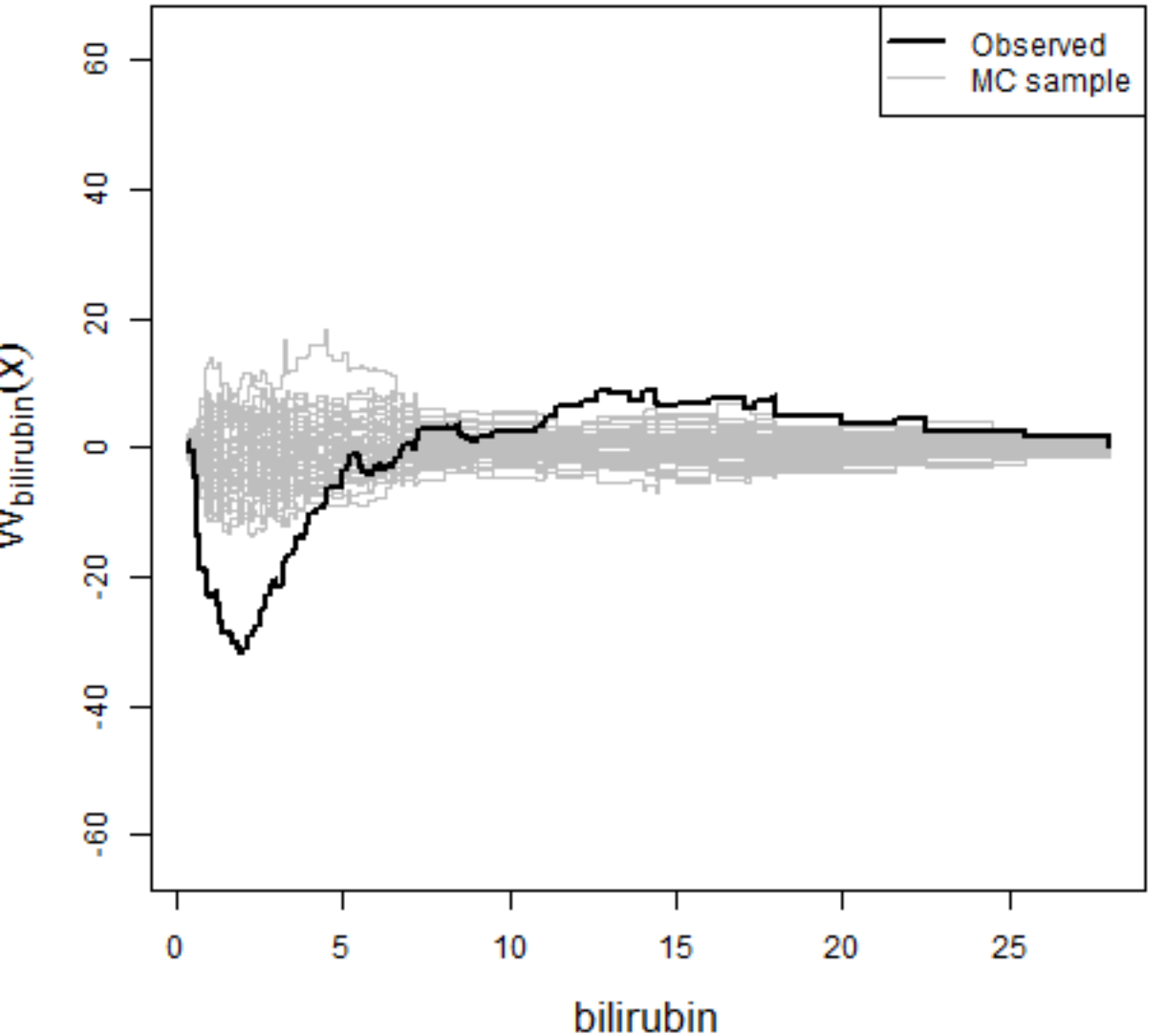

Table 1: Comparison between the packages main features.

Assumption	Model	Statistic	Packages		
			Goftte	Gof	Crskdiag
Proportionality	Cox	KS	x	x	x
		CvM	x	x	
		AD	x		
	Fine & Gray	KS	x		x
		CvM	x		
		AD	x		
Functionnal form	Cox	KS	x		x
	Fine & Gray	KS	x		x

Table 2: Empirical sizes (%) for 2000 simulations of *goftte* tests under (H0) based on cumulative sums of residuals (R=1000).

Model	Event Rate (type 1)	Censoring Rate	Sample Size	Proportionality assumption						Functional form assumption	
				KS		CvM		AD		KS	
				Lin	Liu	Lin	Liu	Lin	Liu	Lin	Liu
Cox	1	0	50	5.20	5.75	4.70	***6.20****	4.35	***6.30****	***3.00****	4.30
			100	5.90	5.85	4.65	***6.15****	***3.95****	***6.10****	4.15	5.25
			200	5.70	5.95	5.70	6.00	5.75	***6.55****	4.60	5.20
	0.85	0.15	50	4.05	4.85	4.15	5.60	***3.60****	5.55	***3.40****	5.65
			100	***6.05****	***6.25****	4.65	5.60	***3.95****	5.95	4.30	5.05
			200	5.95	***6.25****	5.40	5.40	4.95	5.50	4.95	5.25
	0.7	0.3	50	4.10	5.55	4.00	5.35	***3.50****	5.30	***3.90****	***6.60****
			100	6.00	***6.65****	5.05	***6.15****	5.00	***6.40****	4.40	5.90
			200	4.95	5.25	5.10	4.90	4.65	5.35	5.30	5.95
	0.5	0.5	50	4.65	5.30	4.40	5.90	***3.75****	5.40	***3.55****	***6.25****
			100	4.80	5.25	5.10	5.80	5.00	5.60	5.20	***7.00****
			200	4.20	4.75	4.20	4.75	4.05	4.95	5.15	5.50
Fine and Gray	0.66	0	50	5.30	5.10	4.70	5.00	4.10	4.70	***3.75****	5.25
			100	4.80	4.90	5.55	5.65	4.60	5.10	5.35	***6.20****
			200	4.50	4.60	4.20	4.25	4.05	4.00	5.20	5.70
	0.56	0.15	50	4.10	5.10	4.30	4.65	***3.45****	4.35	***3.25****	5.30
			100	4.55	5.00	4.85	5.30	4.40	4.75	5.00	***6.40****
			200	4.90	5.10	4.20	4.60	***3.80****	4.25	4.60	5.60
	0.46	0.3	50	4.25	4.75	4.30	5.30	***3.15****	4.25	***3.30****	5.90
			100	***3.80****	4.75	4.35	4.70	4.50	4.50	4.50	***6.50****
			200	5.65	6.00	4.80	5.45	4.40	5.50	5.10	***6.05****
	0.32	0.5	50	***3.80****	4.65	***3.50****	5.30	***2.55****	4.05	***2.50****	4.65
			100	***3.85****	4.45	4.00	4.65	***3.35****	4.40	4.55	***6.10****
			200	4.85	5.40	4.35	4.70	4.10	4.25	4.40	5.45

KS : Kolmogorov-Smirnov ; **CvM** : Cramer-von-Mises ; **AD** : Anderson-Darling; **R** : Number of Monte-Carlo simulations to generate the limiting null distribution of the statistics;
*: Empirical sizes outside of the 1% range from the expected nominal level (5%)

Table 3: Rejection p-values for goftte on Mayo Clinic Primary Biliary Cirrhosis (PBC) data under proportionality and functional form assumptions (R=20000).

	Proportionality assumption			Functional form assumption
Variable analyzed	KS	CvM	AD	KS
PBC data (Model 1)				
Overall survival				
Age	0.4219	0.57315	0.66045	0.396
Edema	0.0218	0.04745	0.06175	0.3313
log(Bilirubin)	0.09775	0.2037	0.2303	0.0511
log(Albumin)	0.51905	0.52415	0.55935	0.58165
log(Protime)	<0.001	<0.001	<0.001	0.38485
PBC data (Model 2)				
Death incidence				
Age	0.84635	0.6919	0.61285	0.19225
Edema	0.024	0.0437	0.04995	0.29705
log(Bilirubin)	0.2868	0.2736	0.2935	0.09425
log(Albumin)	0.23975	0.1325	0.1101	0.4897
log(Protime)	0.00435	0.00415	0.0034	0.2148

KS : Kolmogorov-Smirnov ; **CvM** : Cramer-von-Mises ; **AD** : Anderson-Darling; **R** : Number of Monte-Carlo simulations to generate the limiting null distribution of the statistics;

Supplementary Table 1: Empirical sizes (%) in a Cox model under proportional hazard assumption for 2000 simulations and n=200 for *goftte* and *gof* tests based on cumulative sums of residuals (R=1000).

Event Rate (type 1)	Censoring Rate	Proportionality assumption				
		Goftte			Gof	
		KS	CVM	AD	KS	CVM
1	0	5.70	5.70	5.75	5.90	***6.80****
0.15	0.15	5.95	5.40	4.95	***6.30****	5.75
0.3	0.3	4.95	5.10	4.65	5.25	5.95
0.5	0.5	4.20	4.20	4.05	4.30	5.25

KS : Kolmogorov-Smirnov ; **CvM** : Cramer-von-Mises ; **AD** : Anderson-Darling; **R** : Number of Monte-Carlo simulations to generate the limiting null distribution of the statistics; **n** : sample size for each simulation;
*: Empirical sizes outside of the 1% range from the expected nominal level (5%)

Supplementary Table 2: Empirical sizes (%) in a Fine and Gray model under proportional sub-distribution hazard and functional form assumptions for 2000 simulations and n=200 for *goftte* and *crskdiag* tests based on cumulative sums of residuals (R=1000).

Event Rate (type 1)	Censoring Rate	Proportionality assumption				Functional form assumption	
		Goftte			Crskdiag	Goftte	Crskdiag
		KS	CVM	AD	KS	KS	KS
0.66	0	4.60	4.25	4.00	4.65	5.20	5.75
0.56	0.15	5.10	4.60	4.25	5.65	4.60	5.75
0.46	0.3	6.00	5.45	5.50	***6.55****	5.10	5.50
0.32	0.5	5.40	4.70	4.25	***7.60****	4.40	5.45

KS : Kolmogorov-Smirnov ; **CvM** : Cramer-von-Mises ; **AD** : Anderson-Darling; **R** : Number of Monte-Carlo simulations to generate the limiting null distribution of the statistics; **n** : sample size for each simulation;
*****: Empirical sizes outside of the 1% range from the expected nominal level (5%)

Table 3: Rejection p-values for goftte on Mayo Clinic Primary Biliary Cirrhosis (PBC) data under proportionality and functional form assumptions (R=20000).

	Proportionality assumption			Functional form assumption
Variable analyzed	KS	CvM	AD	KS
PBC data (Model 1)				
Overall survival				
Age	0.4219	0.57315	0.66045	0.396
Edema	0.0218	0.04745	0.06175	0.3313
log(Bilirubin)	0.09775	0.2037	0.2303	0.0511
log(Albumin)	0.51905	0.52415	0.55935	0.58165
log(Protime)	<0.001	<0.001	<0.001	0.38485
PBC data (Model 2)				
Death incidence				
Age	0.84635	0.6919	0.61285	0.19225
Edema	0.024	0.0437	0.04995	0.29705
log(Bilirubin)	0.2868	0.2736	0.2935	0.09425
log(Albumin)	0.23975	0.1325	0.1101	0.4897
log(Protime)	0.00435	0.00415	0.0034	0.2148

KS : Kolmogorov-Smirnov ; **CvM** : Cramer-von-Mises ; **AD** : Anderson-Darling; **R** : Number of Monte-Carlo simulations to generate the limiting null distribution of the statistics;

Supplementary Table 1: Empirical sizes (%) in a Cox model under proportional hazard assumption for 2000 simulations and n=200 for *goftte* and *gof* tests based on cumulative sums of residuals (R=1000).

Event Rate (type 1)	Censoring Rate	Proportionality assumption				
		Goftte			Gof	
		KS	CVM	AD	KS	CVM
1	0	5.70	5.70	5.75	5.90	***6.80****
0.15	0.15	5.95	5.40	4.95	***6.30****	5.75
0.3	0.3	4.95	5.10	4.65	5.25	5.95
0.5	0.5	4.20	4.20	4.05	4.30	5.25

KS : Kolmogorov-Smirnov ; **CvM** : Cramer-von-Mises ; **AD** : Anderson-Darling; **R** : Number of Monte-Carlo simulations to generate the limiting null distribution of the statistics; **n** : sample size for each simulation;
*: Empirical sizes outside of the 1% range from the expected nominal level (5%)

Supplementary Table 2: Empirical sizes (%) in a Fine and Gray model under proportional sub-distribution hazard and functional form assumptions for 2000 simulations and n=200 for *goftte* and *crskdiag* tests based on cumulative sums of residuals (R=1000).

Event Rate (type 1)	Censoring Rate	Proportionality assumption				Functional form assumption	
		Goftte			Crskdiag	Goftte	Crskdiag
		KS	CVM	AD	KS	KS	KS
0.66	0	4.60	4.25	4.00	4.65	5.20	5.75
0.56	0.15	5.10	4.60	4.25	5.65	4.60	5.75
0.46	0.3	6.00	5.45	5.50	***6.55****	5.10	5.50
0.32	0.5	5.40	4.70	4.25	***7.60****	4.40	5.45

KS : Kolmogorov-Smirnov ; **CvM** : Cramer-von-Mises ; **AD** : Anderson-Darling; **R** : Number of Monte-Carlo simulations to generate the limiting null distribution of the statistics; **n** : sample size for each simulation;
*****: Empirical sizes outside of the 1% range from the expected nominal level (5%)

Table 3: Rejection p-values for goftte on Mayo Clinic Primary Biliary Cirrhosis (PBC) data under proportionality and functional form assumptions (R=20000).

	Proportionality assumption			Functional form assumption
Variable analyzed	KS	CvM	AD	KS
PBC data (Model 1)				
Overall survival				
Age	0.4219	0.57315	0.66045	0.396
Edema	0.0218	0.04745	0.06175	0.3313
log(Bilirubin)	0.09775	0.2037	0.2303	0.0511
log(Albumin)	0.51905	0.52415	0.55935	0.58165
log(Protime)	<0.001	<0.001	<0.001	0.38485
PBC data (Model 2)				
Death incidence				
Age	0.84635	0.6919	0.61285	0.19225
Edema	0.024	0.0437	0.04995	0.29705
log(Bilirubin)	0.2868	0.2736	0.2935	0.09425
log(Albumin)	0.23975	0.1325	0.1101	0.4897
log(Protime)	0.00435	0.00415	0.0034	0.2148

KS : Kolmogorov-Smirnov ; **CvM** : Cramer-von-Mises ; **AD** : Anderson-Darling; **R** : Number of Monte-Carlo simulations to generate the limiting null distribution of the statistics;