INSTITUT DE STATISTIQUE
BIOSTATISTIQUE ET
SCIENCES ACTUARIELLES
(ISBA)

DISCUSSION PAPER

2016/14

Diagnostic checks in mixture cure models with interval-censoring

Scolas, S., Legrand, C., Oulhaj, A. and
A. El Ghouch

Diagnostic checks in mixture cure models with interval-censoring

Sylvie Scolas*[1], Catherine Legrand [1], Abderrahim Oulhaj [2], and Anouar El Ghouch [1]

[1]Institute of Statistics, Biostatistics and Actuarial Sciences (ISBA), Université catholique de Louvain, Louvain-la-Neuve, Belgium
[2]Institute of public health, College of Medicine & Health Sciences, United Arab Emirates University (UAEU), United Arab Emirates (UAE)

Abstract

Models for interval-censored survival data presenting a fraction of "cure" or "immune" patients have recently been proposed in the literature, in particular extending the mixture cure model to the case of interval-censoring. However, little is known about the fit of such models to a given data application. We thus propose to extend the classical Cox-Snell residuals to such models to assess assumptions about the survival distribution. Moreover, as covariates may, in mixture cure models, impact either the probability to experience the event, and/or the survival distribution of the uncured patients, we define deviance residuals allowing to detect non-linearity in covariates in each part of the model. Simulation studies show the behavior of these residuals; they are then applied to an Alzheimer's disease database studying the occurrence of Mild Cognitive Impairment, which may be a precursor of Alzheimer's disease. This event is typically detected between two visits, and is thus interval-censored. Furthermore it is known that not all of the patients will experience this event, leading to a fraction of "cure" or "immune" patients.

1 Introduction

The modelling of "survival data" or more generally "time-to-event data", for which the response of interest is the duration of time between a well defined origin and an event of interest, has a predominant place in medicine. A well known characteristic of these data is that they are typically right-censored, meaning that some patients have not yet experienced the event of interest at the end of the follow-up period. Besides right-censoring, it is also quite frequent that data are interval-censored, with the event only known to have occurred within an interval of time. An example of interval-censored data in Alzheimer's disease can be the duration of time needed for an elderly subject to convert from a healthy to a mild cognitive impairment status (MCI) ([Oulhaj et al., 2009]). Identifying significant risk factors that increase the risk of the conversion from a healthy to an MCI status and also identifying healthy subjects at high risk of conversion to MCI are of great interest, as MCI is known to be a possible precursor of Alzheimer's disease [Ravaglia et al., 2006]. Clearly, such time-to-event data can be both right-censored since some patients have not yet experienced MCI conversion at the end of the follow-up period, and interval-censored since MCI conversion is typically known to have occurred between two successive follow-up visits. As pointed out in [Scolas et al., 2015], an additional feature of these data is that a fraction of the patients will never convert to MCI, whatever the length of follow-up. In the statistical literature, this kind of patients are referred to as "cured individuals", or "long-term survivors" or "non-susceptibles" ([Maller and Zhou, 1996]). Neglecting any of these two particular features, i.e. interval-censoring and/or the presence of a cure fraction, may lead to incorrect inference [Lindsey, 1998, Maller and Zhou, 1996].

In the absence of cure, parametric and semi-parametric approaches have been proposed to handle interval-censoring in the modelling of survival data, see for example [Lindsey, 1998] and [Sun, 2006]. Regarding the presence of a fraction of cured individuals, a common approach is to assume that the population under study is a mixture of cured and uncured individuals [Boag, 1949], leading to the mixture cure model. This

*sylvie.scolas@uclouvain.be

model is constituted of two parts: the incidence part, modelling the probability to be cured, and the latency part, modelling the survival distribution of the event times for uncured observations. A logistic regression is frequently assumed in the incidence part, and popular choices for the latency are the Proportional Hazards (PH) model or the Accelerated Failure Time (AFT) model [Sy et al., 2000, Zhang and Peng, 2012]. Literature combining both interval-censoring and the presence of cure is rather sparse. [Xiang et al., 2011] extend the semi-parametric mixture cure model, with a semi-parametric Cox PH model for the latency part, to the case of interval-censored data and clustered observations. Considering a semi-parametric model leads to complex and computationally intensive estimation procedures, relying, like [Xiang et al., 2011], on the Expectation-Maximization algorithm. To avoid this, [Chen et al., 2013] and [Scolas et al., 2015] rather propose a flexible parametric model assuming an AFT model for the latency along with a flexible distribution for the error term. These papers discussed the estimation method but do not really address the fit of these models. [Chen et al., 2013] shortly discuss a graphical procedure to check the fit of their model, comparing, for fixed covariate values, the fitted global survival curve to a non-parametric estimator, and using standardized residuals to assess to adequacy of the latency part of the model.

In this paper, we focus our attention on the use of residuals, in the case of interval censored data, to assess the fit of mixture cure models considering either a parametric PH or AFT model in the latency part. The inspection of the residuals is indeed one of the usual methods to assess the assumptions of a given model, with a long tradition in linear models. While the definition of residuals for a linear regression model is unambiguous [Seber and Lee, 2012], it becomes more complex in the context of time-to-event analysis mainly due to the presence of censoring. In the context of right-censored data, several types of residuals have been defined with different goals [Collett, 2003]. The most often used are probably the Cox-Snell residuals and the martingale or deviance residuals. In short, Cox-Snell residuals are used to check the fit of the survival distribution, while an inspection of the martingale or deviance residuals may help in detecting if a covariate included in the model needs a transformation. [Farrington, 2000] extend these residuals to evaluate the goodness-of-fit of the Cox PH model in the presence of interval-censored data.

In a mixture cure model, the entire population is characterized by an improper mixed survival distribution; whereas the uncured sub-population follows a proper survival distribution. It therefore seems interesting to be able to use the Cox-Snell residuals to check both the survival distribution of the entire population and of the uncured sub-population. However, it is not obvious that the Cox-Snell residuals applied to the global survival distribution will keep their good properties, due to the improper nature of the distribution. Also, it is not possible to apply the Cox-Snell residuals as such to assess the survival distribution of the uncured sub-population since the uncured status is not observed for right-censored individuals. These residuals can therefore not be computed for all uncured observations. When considering the use of residuals to check the linearity of the covariates, it is important to keep in mind that in a mixture cure model, non-linearity in a covariate may appear in the incidence, in the latency, or in both parts of the model. Residuals should therefore ideally allow a separate diagnostic in the incidence and in the latency component of the model.

In this paper, we aim to extend the use of residuals to perform diagnostic checks in a parametric mixture cure model with right- and interval-censoring. Our first objective is to discuss how to define the Cox-Snell residuals intended to check the survival distribution of the uncured sub-population and of the entire population. To do so, we first study the properties of the Cox-Snell residuals for the entire population. We then define an approach to estimate the status, cured or uncured, of a right-censored observation. This allows us to define Cox-Snell residuals aimed to assess hypothesis on the survival for the uncured. Our second objective is to propose deviance residuals allowing to detect non-linearity in covariates in the incidence and in the latency part of the model, separately.

This paper is organized as follows: Section 2 describes models and notations. Section 3 develops the Cox-Snell residuals in mixture cure models with and without interval-censoring. Residuals aiming at detecting non-linearity are covered in Section 4. Section 5 shows the behavior of the proposed residuals in a simulation study. Section 6 presents the results of the application of our method to a real data set on Alzheimer's disease as mentioned earlier. Finally, our results are discussed in Section 7.

2 The mixture cure model

There are two broad classes of models in the literature that take into account the existence of cured individuals: the promotion time cure model [Tsodikov, 1998], and the mixture cure model, first introduced by [Boag, 1949].

This paper focuses on the latter because of its intuitiveness. The mixture cure model assumes that the entire population of interest is composed of two sub-populations: the uncured and the cured sub-populations. The model consists of two parts, called the incidence and the latency part. The incidence part models the probability to experience the event of interest and the latency part models the event times for uncured individuals only.

Let $t_1, \cdots, t_n$ be realizations of n independent and non-negative random variables $T_1, \cdots, T_n$, denoting the true, but possibly unobserved, time to the event of interest. Unlike standard survival methods, the survival time for cured individuals is infinite and consequently $P(T_i = +\infty) > 0$. Moreover, the survival time t_i is not exactly observed: the event either occurs between two censoring time points, i.e. the observation is interval censored; or occurs later than a censoring time point, i.e. the observation is right-censored. Thus, instead of t_i, an interval $(l_i, r_i]$ such that $l_i < t_i \leq r_i$ is observed. Right-censored observations are covered by allowing r_i to be infinite. Together with (l_i, r_i), the censoring indicator δ_i is also observed: $\delta_i = 1$ means that the individual i experienced the event of interest during the study period, i.e. $0 < l_i < t_i \leq r_i < \infty$; and $\delta_i = 0$ means that the individual i is right-censored, either cured or uncured, i.e. $0 < l_i < t_i \leq r_i = \infty$. In the following, let Y_i be the random variable indicating the uncured status of the individual i, i.e. $Y_i = \mathbb{1}(T_i < +\infty)$. Clearly, $Y_i = 1$ when $\delta_i = 1$, but due to right-censoring, Y_i is unknown when $\delta_i = 0$. In fact, in such a case, one only knows that the true survival time t_i is larger than l_i, but it is impossible to know if the subject in in the cured or uncured group.

Let $S(t_i) = P(T_i \geq t_i | \boldsymbol{X}_i = \boldsymbol{x}_i, \boldsymbol{Z}_i = \boldsymbol{z}_i)$ be the (improper) conditional survival distribution of T_i given the covariate vector $\boldsymbol{X}_i = (1, X_{i1}, \cdots, X_{im}) \in \mathbb{R}^{(m+1)}$ and the covariate vector $\boldsymbol{Z}_i = (1, Z_{i1}, \cdots, Z_{is}) \in \mathbb{R}^{(s+1)}$. The vectors $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$ may share some or all their components. The mixture cure model assumes that

$$S(t) = p_i S_u(t) + 1 - p_i, \quad \forall t \in [0, \infty], \tag{1}$$

where $p_i = P(Y_i = 1 | \boldsymbol{Z}_i = \boldsymbol{z}_i)$ denotes the conditional probability to experience the event of interest given $\boldsymbol{Z}_i$; and $S_u(t_i) = P(T_i \geq t_i | \boldsymbol{X}_i = \boldsymbol{x}_i, Y_i = 1)$ denotes the (proper) conditional survival distribution of the uncured sub-population given $\boldsymbol{X}_i$. Remark that in the absence of cure, i.e. $p_i = 1 \; \forall i$, S coincides with S_u. Let $\nu_i \in \mathbb{R}$ such that

$$p_i \equiv p(\nu_i) = \frac{\exp(\nu_i)}{1 + \exp(\nu_i)}. \tag{2}$$

In the following, ν_i will be modelled through a linear relationship $\nu_i = \boldsymbol{z}_i' \boldsymbol{\gamma}$, where $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, \cdots, \gamma_s)$ is a $(s+1)$-vector of unknown coefficients. This is the classical linear logistic regression model. Concerning S_u, two of the most widely used regression models in survival analysis are the Proportional Hazards (PH) model and the Accelerated Failure Time (AFT). These two models are summarized in the following formula:

$$S_u(t_i) = \begin{cases} S_0(t_i)^{\exp(\mu_i)} & \text{(PH)}, \quad \text{(3a)} \\ S_0\left(\dfrac{\log(t_i) - \mu_i}{\sigma}\right) & \text{(AFT)}. \quad \text{(3b)} \end{cases}$$

In the above equations, $S_0(t_i) \equiv S_0(t_i, \lambda)$ is a given baseline survival distribution assumed to be known up to some finite-dimensional parameter(s) λ common to all individuals; $\sigma > 0$ is a scale parameter; and $\mu_i \in \mathbb{R}$ is a location parameter, typically modelled through the linear relationship $\mu_i = \boldsymbol{x_i}' \boldsymbol{\beta}$, where $\boldsymbol{\beta} = (\beta_0, \beta_1, \cdots, \beta_m)$ is a $(m+1)$-vector of unknown coefficients, with $\beta_0 = 0$ in the PH model to avoid identifiability issues. The AFT model is a useful alternative to the PH model because it does not require the proportional hazards assumption and also because it os often said to have a simpler interpretation: a covariate either accelerates or decelerates the survival time. Although $\boldsymbol{\beta}$ appears in a linear fashion in μ_i in both (3a) and (3b), its interpretation is quite different; see for example [Collett, 2003]. The PH model is commonly written in terms of the hazard function under the form $h_u(t_i) = \exp(\mu_i) h_0(t_i)$, where $h_0(t) = -\frac{d}{dt} \log S_0(t)$ is the baseline hazard function corresponding to the baseline survival function $S_0(t)$, and $h_u(t) = -\frac{d}{dt} \log S_u(t)$. The AFT model is more commonly encountered under the form $\log(T_i) = \mu_i + \sigma \varepsilon_i$, where ε_i is an error term with survival function $S_0(t)$. Commonly used specifications for $S_0(t)$ are the Weibull, the log-Logistic, the log-Normal and the Extended Generalized Gamma (EGG) distributions. The latter is a very flexible family that was recently used in the context of mixture cure models by [Scolas et al., 2015].

The vector of all unknown parameters to be estimated is $\boldsymbol{\eta} = (\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\gamma})$, with $\boldsymbol{\theta} = \lambda$ for the PH model, and $\boldsymbol{\theta} = (\lambda, \sigma)$ for the AFT model. In the following, when no confusion may arise, we will write p_i for $p(\nu_i)$

and $S_u(t_i)$ for $S_u(t_i|\mu_i;\boldsymbol{\theta})$. As shown by [Li et al., 2001], the parametric mixture cure model as given by (1), (2) and (3a) or (3b) is identifiable. Parameter estimates can be obtained by maximum likelihood. Let $\mathcal{O}_i = (\delta_i, l_i, r_i, \boldsymbol{x}_i, \boldsymbol{z}_i)$ denote the observed data for $i = 1, \cdots, n$. The log-likelihood function in the mixture cure model with interval- and right-censored data is given by:

$$l(\boldsymbol{\eta};\mathcal{O}) = \sum_{i=1}^{n} \delta_i \log\left[p_i(S_u(l_i) - S_u(r_i))\right] + (1-\delta_i)\log\left[p_i S_u(l_i) + (1-p_i)\right]. \tag{4}$$

Maximizing the above likelihood function with respect to $\boldsymbol{\eta}$ leads to a consistent and asymptotically efficient estimate $\hat{\boldsymbol{\eta}} = (\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})$ [Casella and Berger, 2001]. The corresponding asymptotic variance-covariance matrix can be obtained as usual through the Hessian matrix. The plug-in method then leads to the estimators $\hat{\nu}_i = \boldsymbol{z}_i'\hat{\boldsymbol{\gamma}}$, $\hat{\mu}_i = \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}$, $\hat{p}_i \equiv p(\hat{\nu}_i)$, $\hat{S}_u(t_i) \equiv S_u(t_i|\hat{\mu}_i;\hat{\boldsymbol{\theta}})$ and $\hat{S}(t_i) = \hat{p}_i\hat{S}_u(t_i) + 1 - \hat{p}_i$ of ν_i, μ_i, p_i, $S_u(t_i)$ and $S(t_i)$, respectively.

3 Checking the survival functions: Cox-Snell residuals

In this section, Cox-Snell residuals aiming at checking the marginal survival function (S) and the uncured survival function (S_u) in a mixture cure model are defined and studied.

The main idea behind the Cox-Snell residuals is as follows: If a random variable T_u has a proper survival distribution S_u, then $W = -\log(S_u(T_u))$ follows an exponential distribution with unit mean. Since the latter has an identity cumulative hazard function, the plot of $-\log(S_W(t))$ against t, where S_W is the survival function of W should reveal points aligned on a straight line with a unit slope and a zero intercept; see [Collett, 2003] for more details.

For simplicity, let's temporarily assume that no interval-censoring is present. To assess the validity of the hypothesized survival distribution S_u of the uncured sub-population, the Cox-Snell residuals $r_{CS,u}(t_i) = -\log(\hat{S}_u(t_i))$, can be computed for all i with $Y_i = 1$, i.e. only for the uncured observations. Since the uncured status Y_i is unknown for right-censored observations, we propose to replace Y_i with its expected value given the observed data :

$$E\left(Y_i|\mathcal{O}_i\right) = \delta_i + (1-\delta_i)\frac{p_i S_u(l_i)}{p_i S_u(l_i) + 1 - p_i}.$$

Since p_i and S_u are unknown, this expectation is estimated by ξ_i:

$$\xi_i = \delta_i + (1-\delta_i)\frac{\hat{p}_i\hat{S}_u(l_i)}{\hat{p}_i\hat{S}_u(l_i) + 1 - \hat{p}_i}. \tag{5}$$

As ξ_i can take any value between 0 and 1, we need a threshold to predict the uncured status. In this work, we take 0.5 as a threshold and classify an individual with $\xi_i > 0.5$ to the uncured sub-population and classify an individual with $\xi_i \leq 0.5$ to the cured sub-population. We thus define uncured Cox-Snell residuals as

$$r_{CS,u}(t_i) = -\log(\hat{S}_u(t_i)) \text{ for } i : \xi_i > 0.5.$$

A plot of the Cox-Snell residuals on the x-axis versus their Kaplan-Meier or Nelson-Aalen estimated cumulative hazard based on the right censored sample $(\delta_i, r_{CS,u}(t_i))$ on the y-axis, should exhibit points aligned on a straight line with a unit slope and zero intercept. A departure from this line may suggest a model inadequacy but, in such a case, no indication of the cause of this inadequacy is provided by the plot. Despite this drawback, the Cox-Snell residuals plot remains useful, for example, to compare two possible distributions.

The same approach can be used to assess the hypothesized global survival distribution S of the whole population, cured and uncured. For this purpose, we define the Global Cox-Snell residuals as

$$r_{CS}(t_i) = -\log(\hat{S}(t_i)) = -\log\left(\hat{p}_i\hat{S}_u(t_i) + 1 - \hat{p}_i\right),\ i = 1, \cdots, n.$$

This definition is motivated by the fact if T has an improper cumulative distribution $F(T) = pF_u(T)$, where $F_u(T) = 1 - S_u(T)$ is a proper cumulative distribution, then for $0 \leq t < p$,

$$P(F(T) \leq t) = pP(T \leq F_u^{-1}(t/p)|T < +\infty) + (1-p)P(F_u(T) \leq t/p|T = +\infty) = t,$$

and for $t \geq p$, $P(F(T) \leq t) = P(F_u(T) \leq t/p) = 1$. Consequently, if the global survival function S fitted to the data is satisfactory, the global Cox-Snell residuals r_{CS} should behave in $[0, -\log(1-p))$ like a censored sample from a mean one exponential distribution. In this case, as for the uncured observations, a plot of the global Cox-Snell residuals versus their estimated cumulative hazard should reveal points aligned on a straight line of unit slope and zero intercept.

The same argument can also be applied to the case of interval-censored data to check the validity of both S_u and S. In the following, only S_u is discussed, as the same approach can be applied to S. From the observed interval-censored data $(l_i, r_i]$, the interval-censored Cox-Snell residuals are obtained: $(-\log(\hat{S}_u(l_i), -\log(\hat{S}_u(r_i)]$, for subject i, $i = 1, \ldots, n$. Their cumulative hazard function can be estimated by the self-consistency algorithm of Turnbull [Turnbull, 1976]. If the model is adequate, a plot of the estimated function against the residuals interval endpoints should approximately resembles a straight line of unit slope and zero intercept. These residuals are not easy to handle given their interval nature. For this reason, [Farrington, 2000] suggest to replace the interval residuals with their expected values under the unit exponential distribution, leading to the following adjusted Cox-Snell residuals for S_u :

$$r_{CS,u}(l_i, r_i) = \frac{\hat{S}_u(l_i)(1 - \log(\hat{S}_u(l_i))) - \hat{S}_u(r_i)(1 - \log(\hat{S}_u(r_i)))}{\hat{S}_u(l_i) - \hat{S}_u(r_i)}. \tag{6}$$

Once the hypothesis on S_u are validated, the validity of the global survival S can also be assessed via the Cox-Snell residuals $r_{CS}(l_i, r_i)$ obtained from (6) but with $\hat{S}$ instead of $\hat{S}_u$.

4 Detecting non-linearity: Deviance residuals

One way of assessing the adequacy of a model is to compare it with the corresponding saturated model. This is the model with the same distribution and the same structure as the model under study, but with the maximum number of parameters of interest that can be estimated ("perfect" estimable model). In our case, the saturated model allows μ_i and ν_i to be different for each observation, whereas the regression hypothesized model assumes that $\mu_i = \boldsymbol{x_i}'\boldsymbol{\beta}$ and $\nu_i = \boldsymbol{z_i}'\boldsymbol{\gamma}$.

In the mixture cure model, the observed data are $\mathcal{O}_i = (\delta_i, l_i, r_i, \boldsymbol{x}_i, \boldsymbol{z}_i)$. The complete data set is obtained by augmenting the observed data set by the partially unobserved uncured status, y_i: $\mathcal{O}_{i,c} = (\delta_i, l_i, r_i, \boldsymbol{x}_i, \boldsymbol{z}_i, y_i)$. Based on the complete data set, the complete-data log-likelihood function for the saturated model is defined as

$$l_S(\mathcal{O}_c) = \sum_{i=1}^{n} l_i(\boldsymbol{\theta}, \nu_i, \mu_i) \tag{7}$$

where $l_i(\boldsymbol{\theta}, \nu_i, \mu_i)$, $i = 1, \ldots, n$, are the individual log-likelihood contributions given by

$$l_i(\boldsymbol{\theta}, \nu_i, \mu_i) = y_i \log(p_i(\nu_i)) + (1-y_i)\log(1-p_i(\nu_i)) + \delta_i \log(S_u(l_i|\mu_i;\boldsymbol{\theta}) - S_u(r_i|\mu_i;\boldsymbol{\theta})) + (1-\delta_i) y_i \log(S_u(l_i|\mu_i;\boldsymbol{\theta})).$$

Following [Nelder and Wedderburn, 1972], we define the "complete" deviance statistic $D(\mathcal{O}_c)$ as twice the difference between the maximum achievable complete-data log-likelihoods under the saturated model and under the current regression model. Thus,

$$D(\mathcal{O}_c) = 2\sum_{i=1}^{n}(l_i(\hat{\boldsymbol{\theta}}, \tilde{\nu}_i, \tilde{\mu}_i) - l_i(\hat{\boldsymbol{\theta}}, \hat{\nu}_i, \hat{\mu}_i)),$$

where $\tilde{\mu}_i$ and $\tilde{\nu}_i$, $i = 1, \ldots, n$, are the points maximizing the saturated complete-data log-likelihood (7), and where $\hat{\boldsymbol{\theta}}$ is the maximum likelihood estimator under the current model of the unknown nuisance parameter $\boldsymbol{\theta}$. Some easy algebra shows that the maximum attainable value for the individual log-likelihood contribution is

$$l_i(\hat{\boldsymbol{\theta}}, \tilde{\nu}_i, \tilde{\mu}_i) = y_i \log(p(\tilde{\nu}_i)) + (1-y_i)\log(1-p(\tilde{\nu}_i)) + \delta_i \log\left(S_u(l_i|\tilde{\mu}_i;\hat{\boldsymbol{\theta}}) - S_u(r_i|\tilde{\mu}_i;\hat{\boldsymbol{\theta}})\right),$$

where $p(\tilde{\nu}_i) = y_i$ and $\tilde{\mu}_i$ is the root of the equation $\frac{\partial}{\partial \mu_i} S_u(l_i|\mu_i;\hat{\boldsymbol{\theta}}) = \frac{\partial}{\partial \mu_i} S_u(r_i|\mu_i;\hat{\boldsymbol{\theta}})$, $i = 1, \ldots, n$. For the PH model, the solution is such that

$$\exp(\tilde{\mu}_i) = \frac{\log(-\log(S_{u,0}(r_i))) - \log(-\log(S_{u,0}(l_i)))}{\log(S_{u,0}(r_i)) - \log(S_{u,0}(l_i))},$$

whereas for the AFT model, the solutions depends on the assumed distribution. For example, as the log-Normal and the Weibull distribution are special cases of the EGG distribution, calculation for the EGG distribution are given in the Appendix.

Consequently, $D(\mathcal{O}_c) = 2\sum_{i=1}^{n} d_i(y_i)$, where $d_i(y_i)$ is the contribution of the observation i to the deviance and is given by

$$\begin{aligned} d_i(y_i) =& l_i(\hat{\boldsymbol{\theta}}, \tilde{\nu}_i, \tilde{\mu}_i) - l_i(\hat{\boldsymbol{\theta}}, \hat{\nu}_i, \hat{\mu}_i) \\ =& y_i \log\left(\frac{y_i}{\hat{p}_i}\right) + (1-y_i)\log\left(\frac{1-y_i}{1-\hat{p}_i}\right) + \delta_i \log\left(\frac{S_u(l_i|\tilde{\mu}_i;\hat{\boldsymbol{\theta}}) - S_u(r_i|\tilde{\mu}_i;\hat{\boldsymbol{\theta}})}{\hat{S}_u(l_i) - \hat{S}_u(r_i)}\right) \\ & - (1-\delta_i) y_i \log(\hat{S}_u(l_i)). \end{aligned}$$

As the uncured status y_i is not observed for right-censored observations, we propose to replace it by its expected value, ξ_i, defined in Equation (5). This leads to the "observed" deviance $D = 2\sum_{i=1}^{n} d_i$, where d_i is a shortcut for $d_i(\xi_i)$. D measures the closeness of the hypothesized model to the saturated model, but without distinguishing between the latency and the incidence part. A nice feature of this deviance is that it can be split into two additive parts: an incidence part D_{inc} and an latency part D_{lat}. More precisely, D can be written as:

$$D = D_{inc} + D_{lat} = 2\sum_{i=1}^{n} d_{i,inc} + 2\sum_{i=1}^{n} d_{i,lat},$$

where

$$\begin{aligned} d_{i,inc} &= \xi_i \log\left(\frac{\xi_i}{\hat{p}_i}\right) + (1-\xi_i)\log\left(\frac{1-\xi_i}{1-\hat{p}_i}\right), \text{ and} \\ d_{i,lat} &= \delta_i \log\left(\frac{S_u(l_i|\tilde{\mu}_i;\hat{\boldsymbol{\theta}}) - S_u(r_i|\tilde{\mu}_i;\hat{\boldsymbol{\theta}})}{\hat{S}_u(l_i) - \hat{S}_u(r_i)}\right) - (1-\delta_i)\xi_i \log(\hat{S}_u(l_i)). \end{aligned}$$

Observe that if all the individuals are uncured ($\xi_i = \hat{p}_i = 1$), then D reduces to D_{lat}, the classical deviance for interval and right-censored data. Observe also that D_{inc} is the deviance for the classical binary logistic model, with ξ_i as response variable. A "large" value of D_{inc} (D_{lat}) indicates a poorly fitted incidence (latency) part. This may be caused by a missing covariate or a covariate needing a transformation in the incidence part ("large" D_{inc}), in the latency part ("large" D_{lat}) or in both parts.

As in classical regression models (with or without censoring), we propose to measure the individual contribution to the deviance via the deviance residuals. In general, they are defined as $s_i\sqrt{2d_i}$, where d_i is the individual contribution to the deviance. In the classical logistic regression model, s_i is the sign of the difference between the observed response and its estimated expected value. In survival analysis, s_i is the sign of $\delta_i - r_i$, with r_i the Cox-Snell residuals, which can be interpreted as the difference between the observed and expected number of events for the individual i over the interval $(0, t_i)$; see [Collett, 2003, Farrington, 2000] for more details. Therefore, in our case, for an individual i, we define three types of deviance residuals. One aiming at checking the global model, one aiming at checking the latency, and the last one aiming at checking the incidence: :

- Global deviance residuals: $r_D(l_i, r_i) = sgn(\delta_i - r_{CS}(l_i, r_i))\sqrt{2d_i}$,
- Latency deviance residuals: $r_{D,lat}(l_i, r_i) = sgn(\delta_i - r_{CS,u}(l_i, r_i))\sqrt{2d_{i,lat}}$,
- Incidence deviance residuals: $r_{D,inc}(l_i, r_i) = sgn(\xi_i - \hat{p}_i)\sqrt{2d_{i,inc}}$.

In the above expressions, $sgn(\cdot)$ is the sign function and $r_{CS}(l_i, r_i)$ and $r_{CS,u}(l_i, r_i)$ are the Cox-Snell residuals for S and S_u, respectively; see (6). As in classical survival analysis, deviance residuals can, for example, be plotted against the estimated linear predictors ($\hat{\mu}_i$ or $\hat{\nu}_i$) or against the explanatory variables in the linear predictors, and any unusual pattern may indicate an omission of an important covariate or a non linear effect of one or more covariates in the latency part, in the incidence part, or in both parts.

5 Simulations study

With this simulation study, we aim to illustrate the behavior of global and uncured Cox-Snell residuals in different settings; and to show that incidence and latency deviance residuals allow to correctly detect the need for a transformation in a covariate in the incidence and/or latency.

5.1 Simulation settings

The time-to-event and uncured status are generated based on four scenarios:

1. Scenario A: All terms are linear, both in the latency and incidence part of the model

$$\begin{cases} \log(T_u) = -0.5 + X_1 - X_2 - X_3 + 0.5\varepsilon \\ p = \phi(\gamma_0 - X_2 + X_3) \end{cases}$$

2. Scenario B: A quadratic term is included in the latency part of the model

$$\begin{cases} \log(T_u) = 1.5 + X_1 - X_2^2 - X_3 + 0.5\varepsilon \\ p = \phi(\gamma_0 - X_2 + X_3) \end{cases}$$

3. Scenario C: A quadratic term is included in the incidence part of the model

$$\begin{cases} \log(T_u) = -0.5 + X_1 - X_2 - X_3 + 0.5\varepsilon \\ p = \phi(\gamma_0 - 0.5X_2^2 - X_3) \end{cases}$$

4. Scenario D: A quadratic term is included in the latency and in the incidence parts of the model

$$\begin{cases} \log(T_u) = -1 + X_1 - X_2^2 - X_3 + 0.5\varepsilon \\ p = \phi(\gamma_0 - 0.5X_2^2 - X_3) \end{cases}$$

We generate X_1 from a Normal distribution with mean 0.5 and variance 1.5, i.e. $X_1 \sim \mathcal{N}(0.5, 1.5)$; X_2 from an Uniform distribution with support $[-3, 3]$, i.e. $X_2 \sim \mathcal{U}[-3, 3]$; and X_3 from a Bernoulli distribution with success probability 0.5, i.e. $X_3 \sim \mathcal{B}(0.5)$. As for the link function ϕ, in order to study the impact on residuals checking of a misspecified link function in the incidence, we generate our data using either a logit or probit function, while we always use logit in our fitting procedure.

Whenever necessary, we use the notation A_l and A_p to distinguish scenario A using ϕ=logit from scenario A using ϕ=probit, and equivalently for scenarios B, C and D. The uncured status is generated from a Bernouilli random variable with probability p, and γ_0 takes different values to reach different proportions of cured individuals. The error term ε is generated following the extreme-value distribution, so that T_u, the time to event of the uncured observations, follows a Weibull distribution. Note that in this case, our simulation study covers both the PH and the AFT model, as the PH and AFT assumptions are verified if the event times follow a Weibull distribution [Collett, 2003]. The right-censoring distribution is exponential with mean λ. The values of γ_0 and λ have been chosen to reach three different proportions of cured and right-censored individuals: light (20% cured, 30% right-censored), medium (30% cured, 40% right-censored), and heavy (50% cured, 60% right-censored). The values of γ_0 and λ as well as details about the censoring/cure rate are given in the Appendix. This simulation setting leads to a total of 24 settings (four scenarios, two link functions, three right-censoring/cure proportions).

To study the effect of interval-censoring on residuals checking, two other scenarios complement the simulation study: event times from scenario B_l and D_l are interval-censored, leading to six additional settings: two scenarios and three levels of cure/right-censoring. To simulate interval-censored data, each patients was supposed to be followed at regular visits, and we have considered unequally spaced visits, generating the length between two successive visits from a uniform distribution $\mathcal{U}[0.25, 0.5]$. The lower limit of the interval is the last visit at which the event has not yet occurred, while the upper limit is the first visit at which the event occurred. This allows a comparison of the performance of the Cox-Snell residuals with or without interval-censored event times.

The sample size was set to 500, and we replicated 500 datasets. For obvious reasons of space, only main results are presented in the paper; further results can be found in the Appendix.

5.2 Simulation results

5.2.1 Cox-Snell residuals

For each setting, we have fitted a mixture cure model assuming a logistic regression in the incidence; and an AFT model for the latency part with either a log-Normal, a Weibull or an Extended Generalized Gamma (EGG) distribution. All covariates (X_1, X_2 and X_3 for the latency and X_1 and X_2 for the incidence) were included in the model, assuming a linear effect. The following results are supported by Figures 1, 2, 3 and 4, which show the uncured or global Cox-Snell residuals for 500 datasets. This means that 500×500 points are superimposed on each plot.

In the following, it is important to keep in mind that the Weibull and log-Normal distributions are special cases of the EGG distribution. Therefore, since the true generating distribution is Weibull, the plots of the Cox-Snell residuals for the models fitted with the EGG distribution is expected to be similar (or better aligned) to the ones obtained when fitting the models with the Weibull distribution. Furthermore, a plot of Cox-Snell residuals based on the log-Normal distribution is expected to show more departure from the straight line than a plot of Cox-Snell residuals based on the EGG or Weibull distribution.

First, we check via simulations whether the use of the indicator function $\mathbb{1}(\xi_i > 0.5)$ provides a satisfactory estimation of the uncured status of an observation to compute the Cox-Snell residuals. In a simulation setting, the uncured status of the observations are known. It is thus possible to compare the Cox-Snell residuals plots based on the true uncured status with the Cox-Snell residuals plots based on the uncured status estimated by $\mathbb{1}(\xi_i > 0.5)$. As seen in Figure 1, both plots lead to identical conclusions about the fitted model. We therefore conclude that $\mathbb{1}(\xi_i > 0.5)$ can be used as an estimation of the uncured status in the computation of the Cox-Snell residuals. Nevertheless, we noticed some discrepancies between the uncured Cox-Snell residuals based on the true uncured status and those based on the estimated uncured status, with models fitted with the log-Normal distribution, and if there is a misspecification in latency (scenario B, see plots in Appendix). This is not an unexpected behavior, since the fitted model are in fact wrongly defined. Plots for other scenarios are given in Appendix.

Second, since global and uncured Cox-Snell residuals may be impacted in a similar way by a misspecification in the incidence part, we expect that a plot of the global Cox-Snell residuals will lead to the same conclusion as a plot of the uncured Cox-Snell residuals. We therefore use simulations to compare the global and uncured Cox-Snell residuals. In every scenarios, the conclusions are indeed similar: compare for example Figures 1 and 2, showing uncured and global Cox-Snell residuals for Scenario C_l. In the following, we will only discuss results using global Cox-Snell residuals.

Third, we investigate the impact of the right-censoring and cure proportion on the ability of the Cox-Snell residuals to correctly prefer the Weibull distribution as the true generating distribution. In every scenario, and for the three proportions of cured and right-censored observations, the Weibull distribution is always correctly chosen over the log-Normal. This conclusion is supported by Figure 2, related to scenario C_l, and Figures 3 and 4, related to scenario D and discussed hereinafter. However, as the proportion of right-censored observations increases, the distinction becomes less evident: compare the plots in Figure 2 from left to right, i.e. from light to heavy censoring.

We have also investigated whether the Cox-Snell residuals can still be used to identify the most appropriate underlying distribution even in the presence of a misspecification in the incidence or the latency part of the model. In this simulation study, two misspecifications in the incidence part are covered: assuming a logit link instead of a probit link, as in scenarios A_p, B_p, C_p and D_p; and assuming linearity in covariates whereas a transformation is needed, as in scenarios C and D. In the latency part, one type of misspecification is covered: assuming linearity in covariates whereas a transformation is needed, as in scenarios B and D. The Cox-Snell residuals plots support the Weibull and EGG distribution over the log-Normal distribution in every settings: Figure 2 illustrates this conclusion in relation to the non-linear covariate in incidence, Figure 3 in relation to the non-linear covariates in both latency and incidence, and Figure 4 supports the conclusion in relation to the non-linear covariate and the wrong link function in incidence. Note also that a misspecification in the latency part is reflected in the Cox-Snell residuals plot for models fitted with the Weibull distribution (see Figures 3 and 4 based on scenario D). As the EGG distribution is more flexible, the Cox-Snell residuals plots for models fitted with the EGG distribution show less departure from the straight line. We conclude that Cox-Snell residuals in a mixture cure model are effective to distinguish between two possible distributions even in the case of misspecification in incidence or latency.

Finally, the last objective of this simulation study is to discuss the impact of interval-censoring on residuals checking. Due to the nature of the Turnbull estimator, a Cox-Snell residuals plot for interval-censored data resemble a step function. For the sake of visibility, we show the Cox-Snell residuals plot for one dataset only. A comparison is then possible with the equivalent Cox-Snell residuals plot without interval-censored data. For scenario C_l, Figure 5 (without interval-censoring) can be compared to Figure 6 (with interval-censored event times): interval-censored residuals show more departure from a straight line. The distinction between two possible distributions is then more complicated but remains possible.

5.2.2 Deviance residuals

In this subsection, a mixture cure model is fitted to each generated dataset assuming a logistic regression in the incidence and a Weibull survival distribution in the latency, and assuming that covariates have a linear effect in both parts of the model. Three types of deviance residuals plots are displayed: global, latency and incidence deviance residuals, versus the covariate X_2, which needs a transformation for scenario B, C and D. For one dataset, it is common to plot the deviance residuals on the y-axis versus the covariates values on the x-axis. To help distinguishing a pattern, one generally adds a lowess smoothing curve to the plot [Cleveland, 1979].

In this simulation study, each plot concerns 500 datasets. This means 500×500 residuals points and 500 lowess lines are superimposed on each plot. To enhance visibility, points are deleted from the graphs, so that only the 500 lowess curves are shown on each plot.

Latency deviance residuals are expected to show a quadratic trend for scenarios B and D; and incidence deviance residuals are expected to show a quadratic trend for scenarios C and D.

First, global, latency and incidence deviance residuals show a trend when appropriate. This can be seen by looking at Figures 7, 8, and 9, giving deviance residuals for scenarios B_l, C_l and D_p, respectively. Other plots are given, as said before, in the Appendix. However, when a transformation is needed in both the latency and the incidence part, a trend may be hidden in the global deviance residuals plot, as can be seen on the global deviance residuals plot for medium censoring in Figure 9. Hence the importance of checking the latency and incidence deviance residuals as well.

Second, latency deviance residuals are somewhat affected by the increase of the right-censored proportion: a high censoring proportion can hide a trend in the smoothing curve (see latency deviance residuals of Figure 7: the lowess lines does not reveal a curve for heavy censoring), or suggest an other transformation than the true one (see latency residuals of Figure 9, where some curves maight suggest a cubic transformation). When looking at a deviance residual plot for only one data set, as given in Figure 10 for example, a trend is highly detectable when right-censored and uncensored observations are plotted in different colors. Plotting right-censored residuals with a different symbol may thus circumvent the issue, and, following [Collett, 2003], we recommend to do so.

Third, the same issue may arise with incidence deviance residuals with a too low cured proportion (see incidence deviance residuals of Figure 8 for light censoring). Indeed, less cured observations implies less information for the modelling of the incidence part [Scolas et al., 2015].

6 Application on real data: Oxford Project To Investigate Memory and Aging (OPTIMA)

We apply our approach to a data set related to a study on Alzheimer's disease [Oulhaj et al., 2009]. As explained in [Scolas et al., 2015], the main objective of that study was to identify a set of cognitive scores that predict the probability of conversion from healthy to Mild Cognitive Impairment (MCI) stage in elderly subjects. MCI often represents the pre-dementia stage of a neuro-degenerative disorder, including Alzheimer disease (AD), vascular dementia (VaD), or other dementia syndromes (ODS) and hence early detection of its onset is of great relevance for patients, carers and government. For that study, a cohort of 241 normal elderly volunteers was followed for up to 20 years with regular assessments of their cognitive abilities using the Cambridge Cognitive Examination (CAMCOG). Among them, 91 converted to MCI (37.8%), and the other 150 (62.2%) were right-censored. The CAMCOG score ranges from 0 to 107 with high scores indicating higher abilities. It is comprised of sub-tests including orientation, comprehension, expression, recent memory, remote memory, learning, abstract thinking, perception, praxis, attention, and calculation.

Criteria for diagnosis of MCI and control were carried out according to international guidelines. For more details see ([Oulhaj et al., 2009]). To summarize, conversion to MCI was determined by a neuropsychologist at each visit, which took place in average every year and a half. The data are clearly interval-censored since conversion actually occurred between visits, and the exact date was not known.

Furthermore, it is known that a fraction of these individuals will actually never experience MCI conversion. More details can be found in [Oulhaj et al., 2009] and [Scolas et al., 2015]. The later analyzed these data using a mixture cure model, with an AFT model with EGG distribution for the error term in the latency part and a logistic regression in the incidence part. Our analysis evaluates the association between CAMCOG score and the time to MCI conversion, adjusting for gender, age at baseline, number of folate cells, total Homo-cysteine (tHcy) rate, Mini-Mental State Examination (MMSE) score, and expression of the APO E4 gene (APOEE4). Maximum likelihood estimates are given in Table 1, and the R-code used to obtain estimates is available from the first author. We use this EGG-AFT mixture cure model to illustrate the use of the Cox-Snell and deviance residuals.

The EGG distribution in the latency part is well supported by the data, as seen on Figure 11, but small departures of the Cox-Snell residuals from a straight line could suggest a missing covariate, for example. Global deviance residuals do not exhibit strong pattern when plotted against any of the five covariates Age, CAMCOG, Folate, MMSE or tHcy. The same conclusion is reached for latency and incidence deviance residuals plots; see Figures 12 and 13.

In conclusion, with data and covariates at hand, and based on Cox-Snell and deviance residuals, there is no need to reconsider the use of the EGG distribution in the mixture cure model, nor to reconsider linearity in latency and incidence.

7 Conclusion

Although widely used in linear regression, and also in classical survival analysis, diagnostic checks based on residuals have not been studied in the context of interval-censored data with a sub-population of cured individuals. In parametric mixture cure models, assumptions are made on the survival distribution of the uncured observations, and we discuss the use of Cox-Snell residuals to detect if those assumptions are correct. It is common to enter the covariates in a linear way, both in the incidence and the latency part of the model but in practice covariates may have a non-linear effect in either or both parts of the model. We thus also propose deviance-based residuals allowing to detect if a covariate needs a transformation, and in which part of the model.

We have shown that global Cox-Snell residuals, based on the improper survival distribution of the entire population, still follows a mean one exponential distribution if the model is correctly fitted. Furthermore, we have also defined uncured Cox-Snell residuals, computed based on an estimation of the uncured status for each observation. Our simulation study demonstrates that global and uncured Cox-Snell residuals can suitably be used to assess the hypothesis about the survival distribution in the latency part, even if the incidence part is incorrectly specified.

In addition, we have shown that non-linearity in latency and in incidence is correctly detected by plotting the latency and incidence deviance residuals against a covariate. As expected, a heavy right-censored proportion may hide a trend in the plot and prevent detecting the need for a transformation in the latency. A low cured fraction has an identical effect in incidence. Like in other models, such residuals plots should be interpreted with care since, as stated by [Collett, 2003], no pattern in the residuals plot does not imply a correct model but rather the absence of reasons to think it is incorrect.

The subjective nature of residual checking techniques is sometimes criticized, and the use of goodness-of-fit hypothesis test may be advocated. However in practice, these graphical checks are still widely used. For example, to check the normality of a variable, one generally agrees that a Q-Q plot or a simple histogram is more informative than a formal normality test. In classical survival analysis, several authors stress the importance of residuals checking in the modelling process [Therneau et al., 1990, Collett, 2003, Klein and Moeschberger, 2003, Lawless, 2003].

To develop a formal test with the same objective as the residuals we have discussed, one could for example extend the semi-nonparametric (SNP) method proposed by [Nysen et al., 2012] for right and interval-censored data to the case of a presence of a cure fraction to develop a formal goodness-of-fit test. A simple graphical check of the Cox-Snell residuals should, however, give a first idea as whether the model is seriously

wrong or rather adequate enough. Concerning non-linearity of covariates, one could think of extending the cumulative residuals proposed by [Lin et al., 1993, Lin and Spiekerman, 1996] to provide a formal test to detect a needed transformation in the presence of a cure fraction, but the extension to interval-censored data is unclear, as [Sun, 2006] stated. Indeed, these tests are based on counting processes which are, by nature, not easily extendable to interval-censored data. Furthermore, if the cumulative residuals test conclude to a needed transformation, a plot of the usual martingale (or deviance) residuals might still be needed to help in detecting the form of the transformation [Lin et al., 1993]. Moreover, these formal testing methods need further development, are more complex and will require more computing power.

The residuals presented in this paper have been implemented in parametric mixture cure models, following the work of [Chen et al., 2013] and [Scolas et al., 2015] on such models. An extension to semi-parametric mixture cure models should however be straightforward, as only the final estimates of the model are required in the definition of Cox-Snell and deviance residuals. These residuals can therefore also be used to check a semi-parametric mixture cure model, such as the ones discussed in [Sy et al., 2000] and [Zhang and Peng, 2012]. However, Cox-Snell residuals are not as useful in a semi-parametric context, due to the fact that the baseline survival distribution has to be estimated non-parametrically [Collett, 2003], so that the approximation to the unit exponential distribution is less likely to hold. Deviance residuals on the other hand do not suffer the same issues and can be used in a semi-parametric context to check the linearity of the covariates.

Lastly, other types of residuals could to be extended to the mixture cure model, with or without interval-censored data. Residuals to detect outliers and influential values have been extended [Ortega et al., 2008] to the case of cure models, but the extension of residuals to check proportionality of hazards in a PH mixture cure model or the assumptions of the AFT in an AFT mixture cure model is subject of future work.

Acknowledgment

The first three authors acknowledges financial support from the IAP research network P7/06 of the Belgian Government (Belgian Science Policy), and from the contract "Projet d'Actions de Recherche Concertées" (ARC) 11/16-039 of the "Communauté française de Belgique", granted by the "Académie Universitaire Louvain".
The principal grant support for OPTIMA came from Bristol-Myers Squibb, Merck & Co. Inc., Medical Research Council, Charles Wolfson Charitable Trust, Alzheimer's Research Trust, and Norman Collisson Foundation.
We are grateful to Professor A. David Smith, University of Oxford for permission to use some of unpublished data from the OPTIMA cohort.

References

[Boag, 1949] Boag, J. (1949). Maximum likelihood estimates of the proportion of patients cured by cancer therapy. *Journal of the Royal Statistical Society. Series B (Methodological)*, 11(1):15–53.

[Casella and Berger, 2001] Casella, G. and Berger, R. (2001). *Statistical Inference*. Duxbury Resource Center.

[Chen et al., 2013] Chen, C.-h., Tsay, Y., Wu, Y., and Horng, C. (2013). Logistic AFT location-scale mixture regression models with nonsusceptibility for left -truncated and general interval-censored data. *Statistics in medicine*, 32(24):4285–4305.

[Cleveland, 1979] Cleveland, W. S. (1979). Robust Locally Weighted Regression and Smoothing Scatterplots. *Journal of the American Statistical Association*, 74:829–836.

[Collett, 2003] Collett, D. (2003). *Modelling Survival Data in Medical Research, Second Edition*. Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis.

[Farrington, 2000] Farrington, C. P. (2000). Residuals for proportional hazards models with interval-censored survival data. *Biometrics*, 56(2):473–482.

[Klein and Moeschberger, 2003] Klein, J. and Moeschberger, M. (2003). *Survival Analysis: Techniques for Censored and Truncated Data*. Springer.

[Lawless, 2003] Lawless, J. F. (2003). *Statistical models and methods for lifetime data*. Wiley.

[Li et al., 2001] Li, C.-S., Taylor, J. M. G., and Sy, J. P. (2001). Identifiability of cure models. *Statistics & Probability Letters*, 54(4):389–395.

[Lin and Spiekerman, 1996] Lin, D. Y. and Spiekerman, C. F. (1996). Model checking techniques for parametric regression with censored data. *Scandinavian Journal of Statistics*, 23:157–177.

[Lin et al., 1993] Lin, D. Y., Wei, L. J., and Ying, Z. (1993). Checking the Cox model with cumulative sums of martingale-based residuals. *Biometrika*, 80(3):557–572.

[Lindsey, 1998] Lindsey, J. K. (1998). A Study of Interval Censoring in Parametric Regression Models. *Lifetime data analysis*, 4:329–354.

[Maller and Zhou, 1996] Maller, R. A. and Zhou, X. (1996). *Suvival analysis with long term survivor*. John Wiley & Sons, Inc.

[Nelder and Wedderburn, 1972] Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society, Series A, General*, 135:370–384.

[Nysen et al., 2012] Nysen, R., Aerts, M., and Faes, C. (2012). Testing goodness of fit of parametric models for censored data. *Statistics in medicine*, 31(21):2374–85.

[Ortega et al., 2008] Ortega, E. M. M., Cancho, V. G., and Lachos, V. H. (2008). Assessing influence in survival data with a cure fraction and covariates. *SORT*, 32(2):115–140.

[Oulhaj et al., 2009] Oulhaj, A., Wilcock, G. K., Smith, a. D., and de Jager, C. a. (2009). Predicting the time of conversion to MCI in the elderly: role of verbal expression and learning. *Neurology*, 73(18):1436–42.

[Ravaglia et al., 2006] Ravaglia, G., Forti, P., Maioli, F., Martelli, M., Servadei, L., Brunetti, N., Pantieri, G., and Mariani, E. (2006). Conversion of mild cognitive impairment to dementia: predictive role of mild cognitive impairment subtypes and vascular risk factors. *Dementia and geriatric cognitive disorders*, 21(1):51–58.

[Scolas et al., 2015] Scolas, S., El Ghouch, A., Legrand, C., and Oulhaj, A. (2015). Variable selection in a flexible parametric mixture cure model with interval-censored data. *Statistics in Medicine*.

[Seber and Lee, 2012] Seber, G. and Lee, A. (2012). *Linear Regression Analysis*. Wiley Series in Probability and Statistics. Wiley, New York.

[Sun, 2006] Sun, J. (2006). *The statistical analysis of interval-censored failure time data*. Springer-Verlag New York.

[Sy et al., 2000] Sy, J. P., Taylor, J. M. G., Way, D. N. A., and Francisco, S. S. (2000). Estimation in a Cox Proportional Hazard Cure Model. *Biometrics*, 56(1):227–236.

[Therneau et al., 1990] Therneau, T. M., Grambsch, P. M., and Fleming, T. R. (1990). Martingale-based residuals for survival models. *Biometrika*, 77:147–160.

[Tsodikov, 1998] Tsodikov, A. (1998). A proportional hazards model taking account of long-term survivors. *Biometrics*, 54(4):1508–16.

[Turnbull, 1976] Turnbull, B. W. (1976). The Empirical Distribution Function with Arbitrarily Grouped , Censored and Truncated Data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 38(3):290–295.

[Xiang et al., 2011] Xiang, L., Ma, X., and Yau, K. K. W. (2011). Mixture cure model with random effects for clustered interval-censored survival. *Statistics in Medicine*, 30:995–1006.

[Zhang and Peng, 2012] Zhang, J. and Peng, Y. (2012). Semiparametric estimation methods for the accelerated failure time mixture cure model. *Journal of the Korean Statistical Society*, 41:415–422.

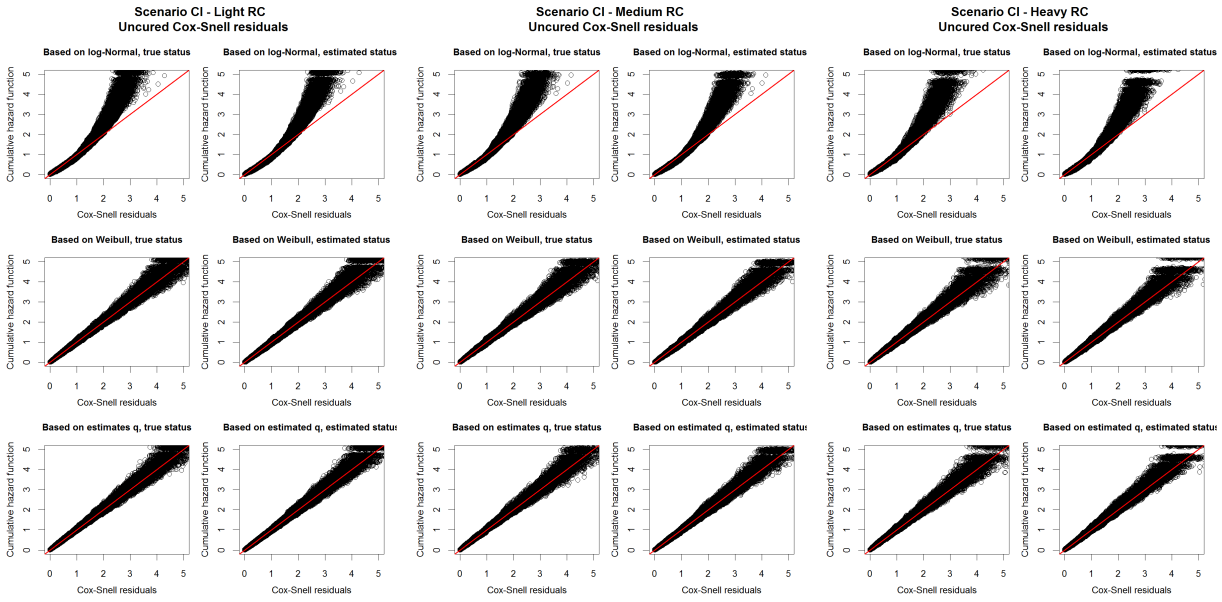

Figure 1: Uncured Cox-Snell residuals for scenario C_l. Left to right: light to heavy cured/right-censored proportion; above to bottom: based on log-Normal, Weibull, and EGG(q). Residuals for 500 datasets are superimposed.

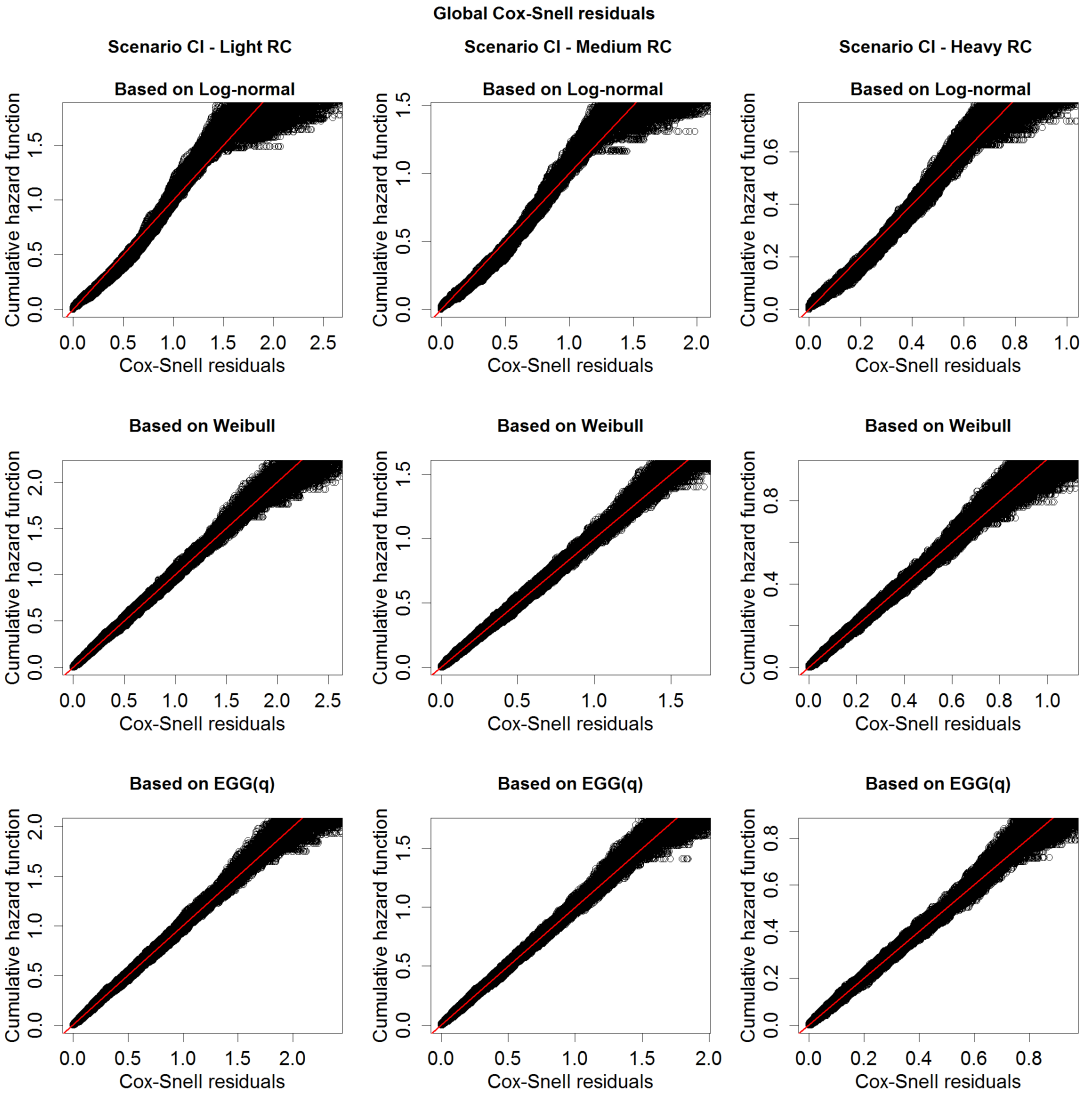

Figure 2: Global Cox-Snell residuals for scenario C_l. Left to right: light to heavy cured/right-censored proportion; above to bottom: based on log-Normal, Weibull, and EGG(q). Residuals for 500 datasets are superimposed.

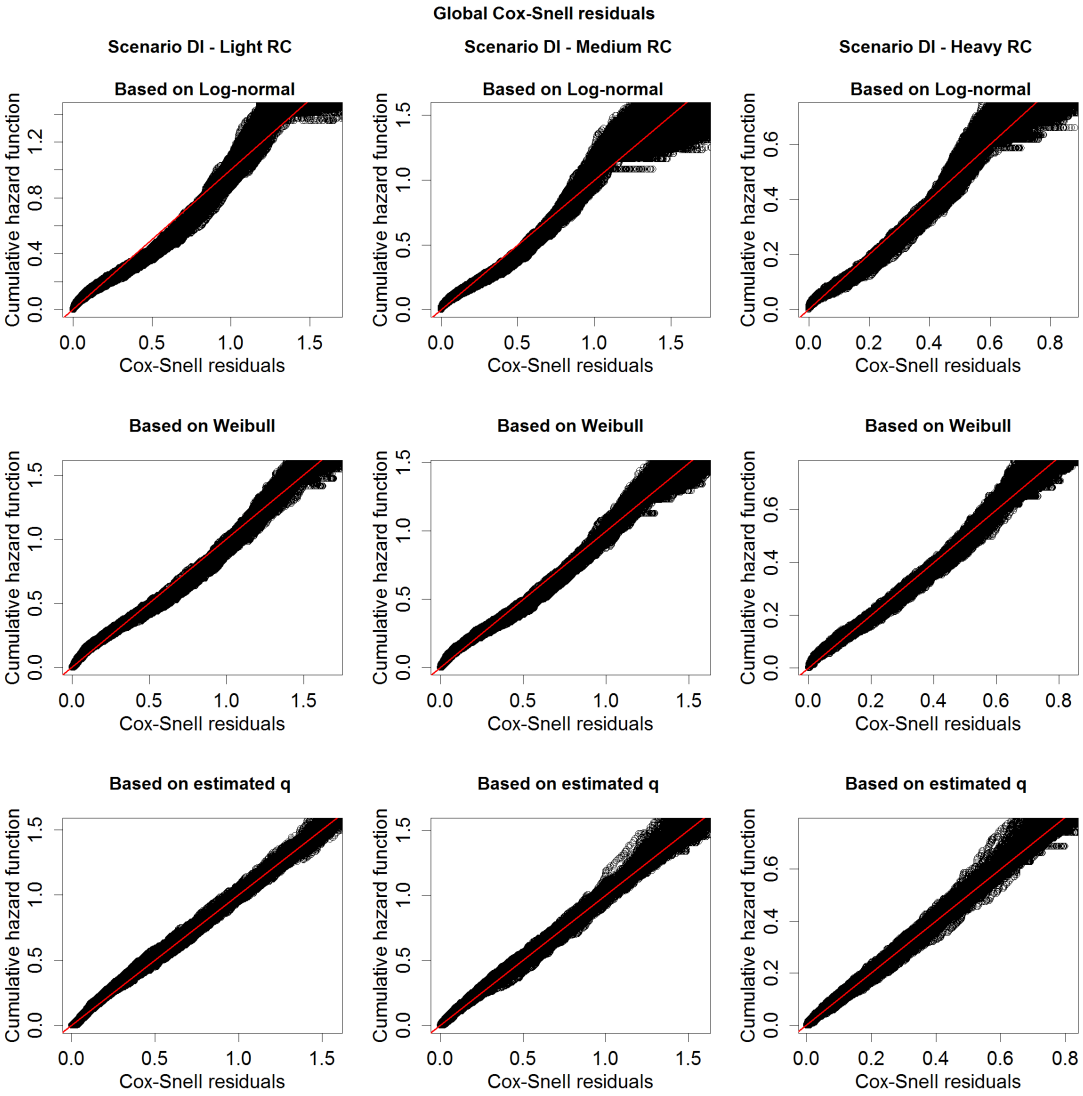

Figure 3: Global Cox-Snell residuals for scenario D_l. Left to right: light to heavy cured/right-censored proportion; above to bottom: based on log-Normal, Weibull, and EGG(q). Residuals for 500 datasets are superimposed.

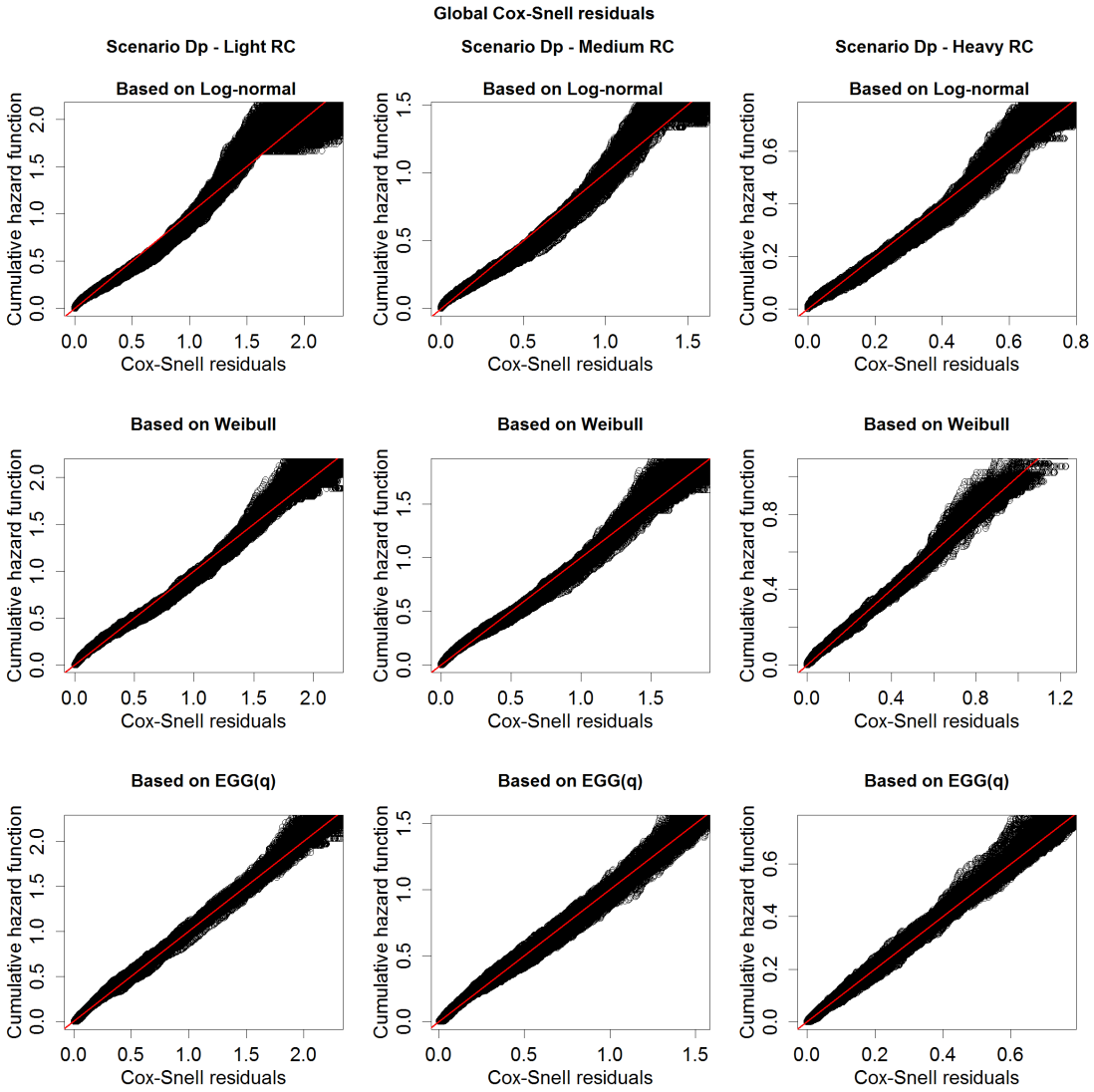

Figure 4: Global Cox-Snell residuals for scenario D_p. Left to right: light to heavy cured/right-censored proportion; above to bottom: based on log-Normal, Weibull, EGG(q). Residuals for 500 datasets are superimposed.

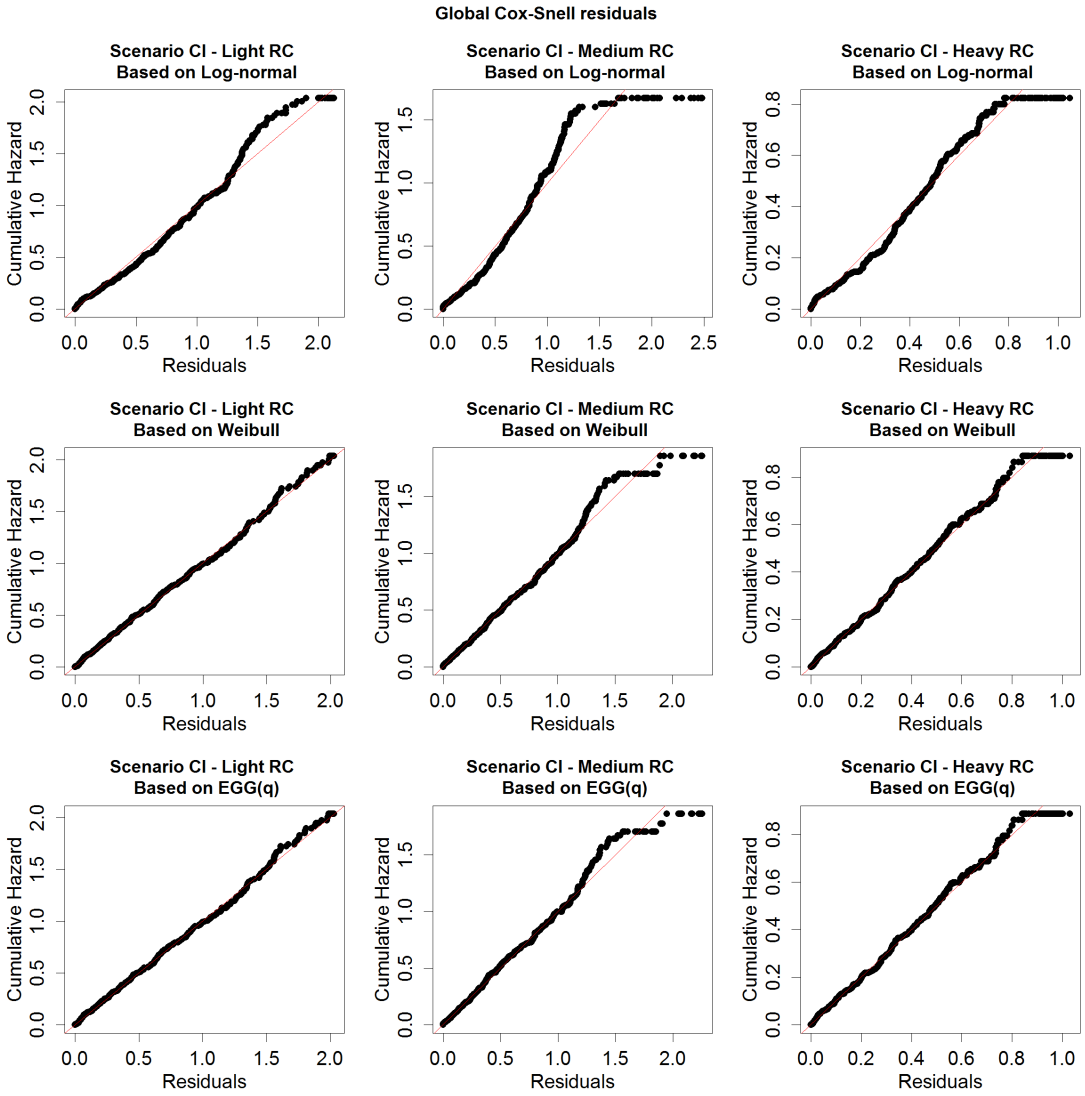

Figure 5: Global Cox-Snell residuals for scenario C_l, without interval-censored data. Left to right: light to heavy cured/right-censored proportion; above to bottom: based on log-Normal, Weibull, and EGG(q). Residuals for one dataset only.

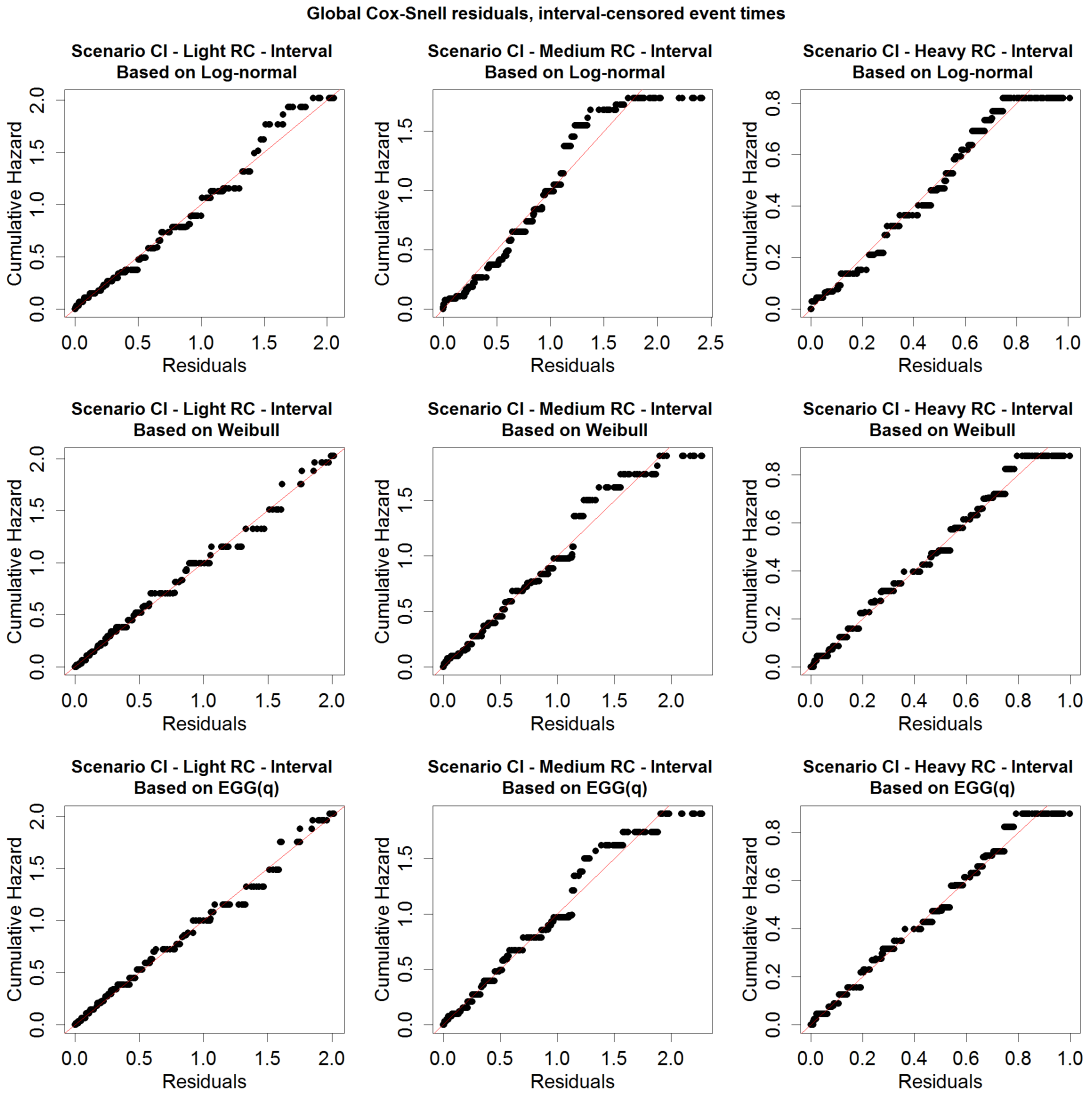

Figure 6: Global Cox-Snell residuals for scenario C_l, with interval-censored data. Left to right: light to heavy cured/right-censored proportion; above to bottom: based on log-Normal, Weibull, and EGG(q). Residuals for one dataset only.

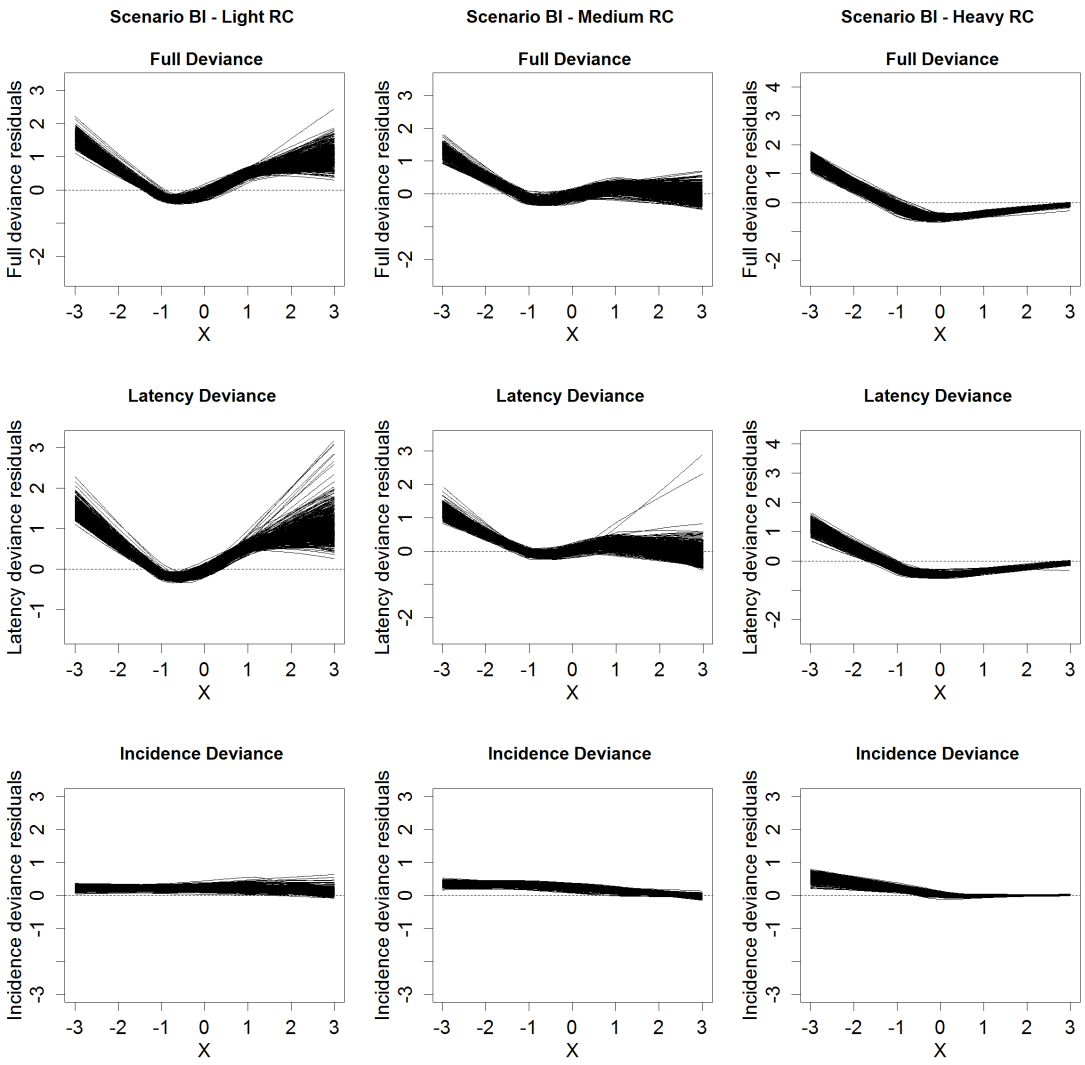

Figure 7: Deviance residuals for scenario B_l. Left to right: for light to heavy cured/right-censored proportion; above to bottom: global, latency and incidence deviance residuals. Lowess smoothing curve for 500 datasets are superimposed.

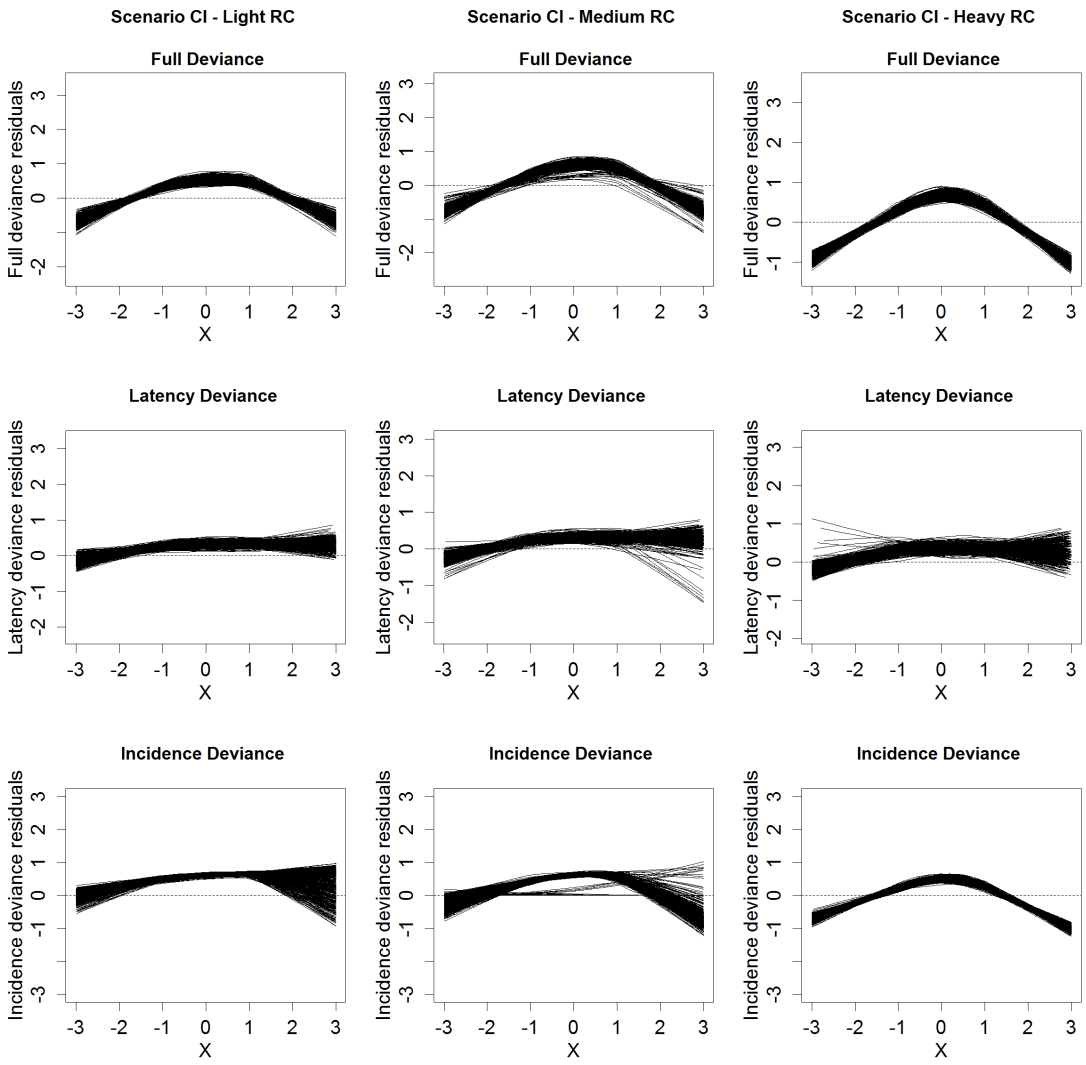

Figure 8: Deviance residuals for scenario C_l. Left to right: for light to heavy cured/right-censored proportion; above to bottom: global, latency and incidence deviance residuals. Lowess smoothing curve for 500 datasets are superimposed.

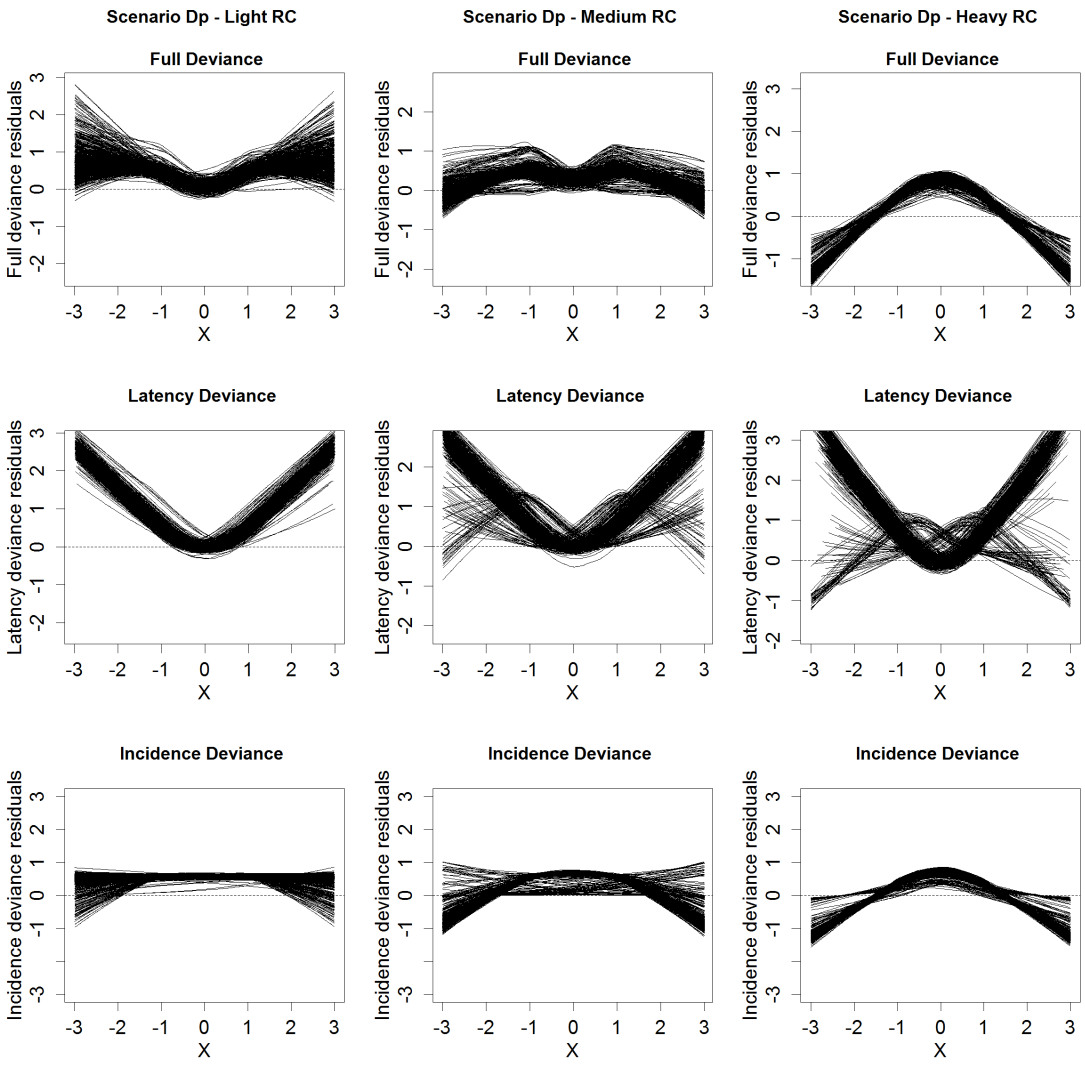

Figure 9: Deviance residuals for scenario D_p. Left to right: for light to heavy cured/right-censored proportion; above to bottom: global, latency and incidence deviance residuals. Lowess smoothing curve for 500 datasets are superimposed.

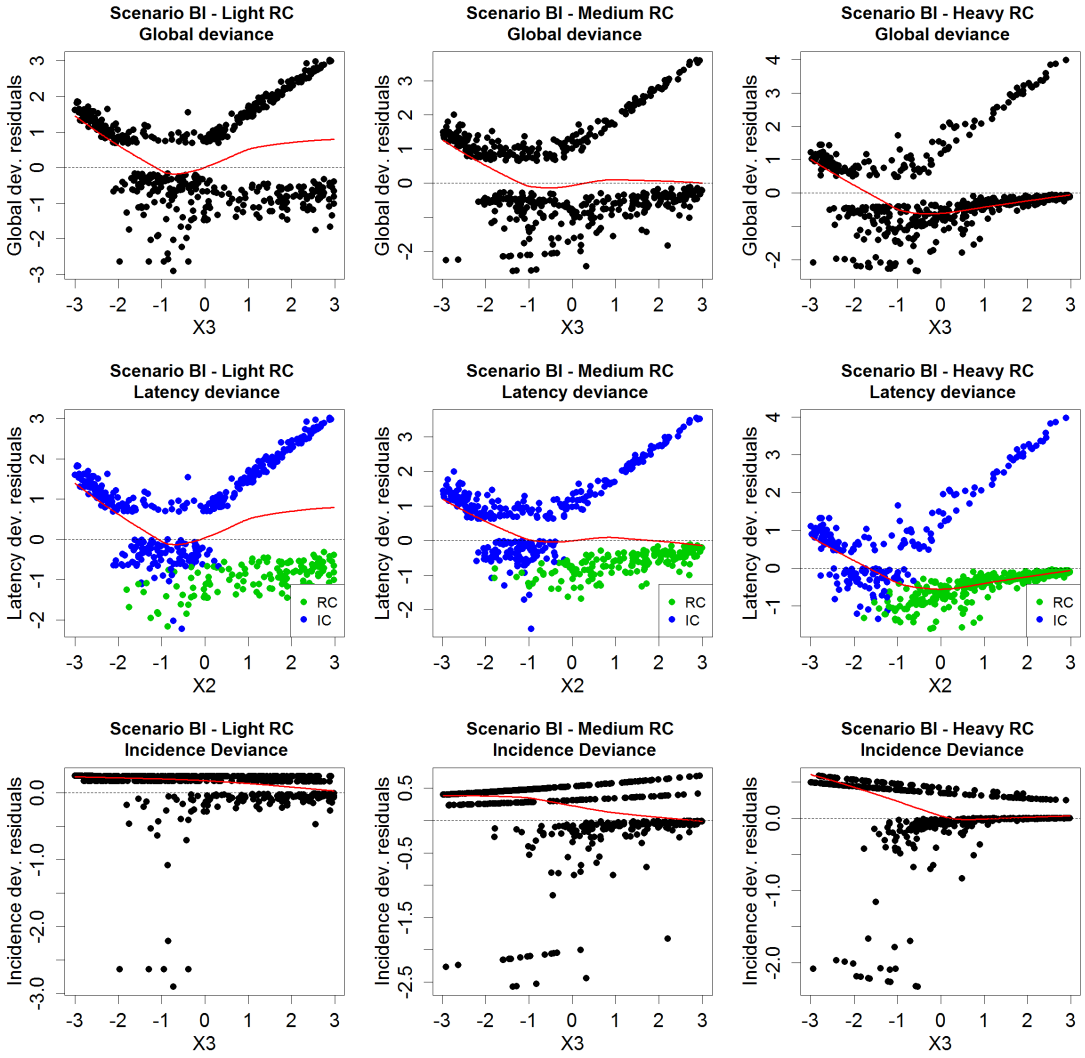

Figure 10: Deviance residuals for scenario B_l. Left to right: for light to heavy cured/right-censored proportion; above to bottom: global, latency and incidence deviance residuals. Residuals and lowess smoothing curve for one dataset only.

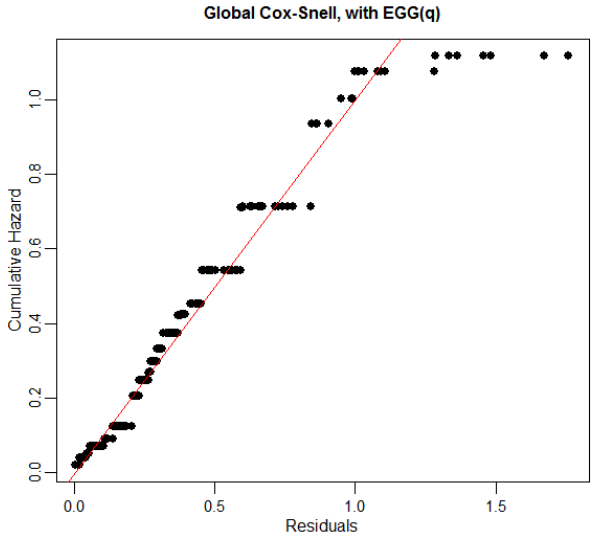

Figure 11: Global Cox-Snell residuals for the MCI database, based on EGG-AFT mixture cure model. EGG(q) distribution is well supported by the data.

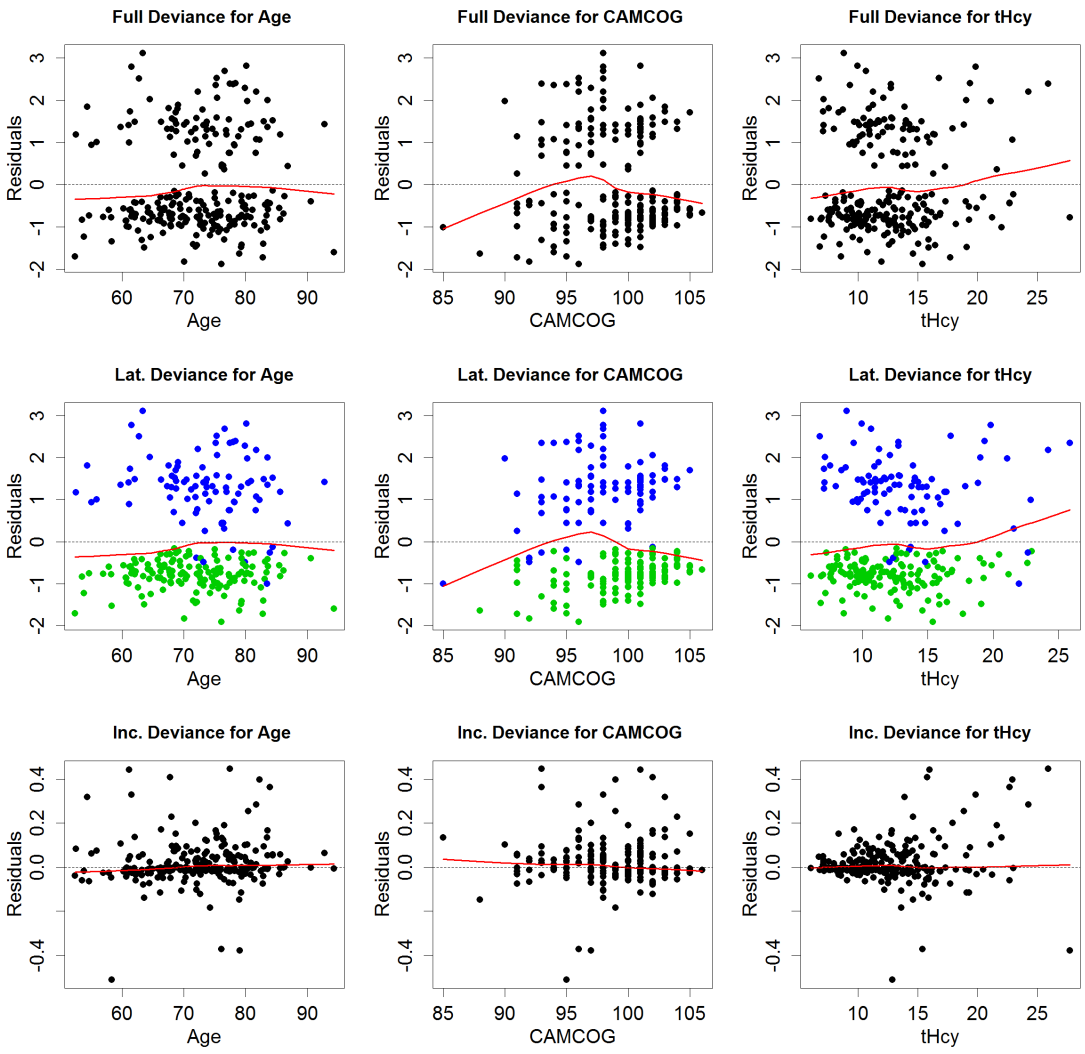

Figure 12: Global, latency and incidence deviance residuals for the MCI database, based on EGG-AFT mixture cure model, for Age, CAMCOG and tHcy.

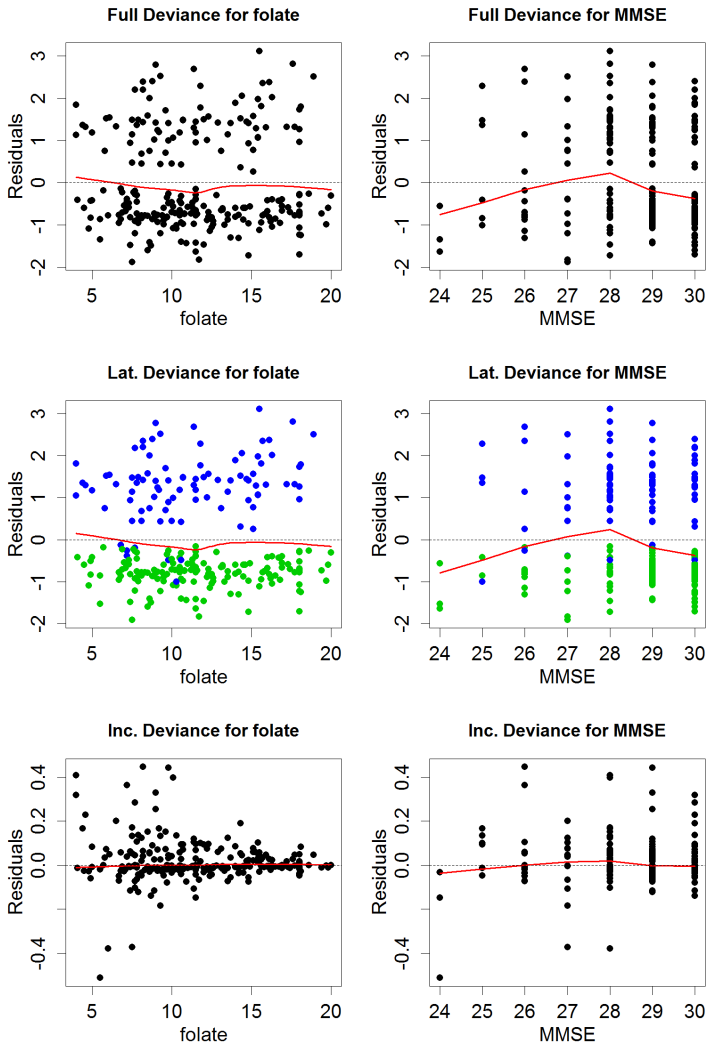

Figure 13: Global, latency and incidence deviance residuals for the MCI database, based on EGG-AFT mixture cure model, for folate and MMSE.

List of Tables

1 Parameter estimates and standard errors of the EGG-AFT mixture cure model for the AD database 27

Table 1: Parameter estimates and standard errors of the EGG-AFT mixture cure model for the AD database

Parameter	Estimate	Std. Error
Latency Part: EGG-AFT model		
q	1,73	1,44
Intercept Latency	-4,92	0,97
Age	-0,03	0,01
CAMCOG	0,12	0,03
tHcy	-0,02	0,02
Folate	-0,01	0,02
MMSE	-0,04	0,01
Gender	-0,14	0,15
APOEE4	-0,16	0,16
$\log(\sigma)$	-0,82	0,73
Incidence Part: Logistic regression		
Intercept Incidence	11,2	58,6
Age	0,08	0,29
CAMCOG	-0,28	1,02
tHcy	-0,33	0,84
Folate	0,09	0,94
MMSE	0,73	1,91
Gender	-1,09	4,22
APOEE4	-0,91	3,54