Fitting parametric cure models in R using the packages cuRe and rstpm2

Jensen, Rasmus Kuhr; Clements, Mark; Gjærde, Lars Klingen; Jakobsen, Lasse Hjort

Published in:
Computer Methods and Programs in Biomedicine

DOI (link to publication from Publisher):
10.1016/j.cmpb.2022.107125

Publication date:
2022

Document Version
Publisher's PDF, also known as Version of record

Computer Methods and Programs in Biomedicine 226 (2022) 107125

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

journal homepage: www.elsevier.com/locate/cmpb

Fitting parametric cure models in R using the packages `cuRe` and `rstpm2`

Rasmus Kuhr Jensen[a], Mark Clements[b], Lars Klingen Gjærde[c], Lasse Hjort Jakobsen[a,d,*]

[a] Department of Haematology, Aalborg University Hospital, Sdr. Skovvej 15, Aalborg 9000, Denmark
[b] Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Nobels Väg 12A, Stockholm 171 65, Sweden
[c] Department of Haematology, Rigshospitalet, Copenhagen University Hospital, Copenhagen, Denmark
[d] Department of Mathematical Sciences, Aalborg University, Skjernvej 4A, Aalborg Ø 9220, Denmark

ARTICLE INFO

Article history:
Received 24 May 2022
Revised 9 September 2022
Accepted 10 September 2022

Keywords:
Cure models
Parametric models
Splines
Relative survival
CuRe
Rstpm2

ABSTRACT

Background and objective: Within medical research, cure models are useful for analyzing time-to-event data in the scenario where a proportion of the analyzed individuals are expected to never experience the event of interest. Cure models are also useful for modelling the relative survival in scenarios where a proportion of the individuals are expected to eventually experience a mortality rate similar to that of the general population. Here we present two R packages, `cuRe` and `rstpm2`, that provide researchers with several tools for performing statistical inference using parametric cure models.

Methods: Cure models are commonly used to estimate 1) the proportion of individuals that are cured and 2) the event-time distribution of individuals who are not cured. This can be done using simple parametric distributions for the event-time distribution of the uncured, but our implementations also enable fitting of more flexible spline-based cure models. The parametric framework of both packages ensures that cure models for the relative survival can easily be used.

Results: The `cuRe` package contains two main functions for estimating parametric mixture cure models; one based on simple parametric distributions (e.g. Weibull or exponential) and one utilizing a spline-based formulation of the cure model. The `rstpm2` package enables estimation of spline-based latent cure models, i.e., cure models with no explicit parameters modelling the proportion of cured individuals.

Conclusions: Through the R-packages `cuRe` and `rstpm2`, a wide range of different parametric cure models can be fitted. The `cuRe` package also contains a number of useful post-estimation procedures for computing the time to statistical cure and conditional probability of cure, which may spread the use of cure models in medical research.

1. Introduction

Cure models constitute a wide range of models that are applicable to time-to-event analysis [1]. They have traditionally been considered when a proportion of the subjects under follow-up are expected to never experience the event of interest. This is for instance the case when measuring the time from first to second birth, because some mothers will never have a second child [2]. Here, cure models can be used to answer questions such as: 1) what proportion of first-time mothers will never have a second child? and 2) among women who eventually have a second child, what is the distribution of the time between births?

Cure models may also be employed for analyzing overall survival with the goal of estimating the proportion of individuals that are long-term survivors [3]. In many contexts, analyzing overall survival relative to the survival in the general population from which the individuals originate (*relative survival*) is of interest. The relative survival function is defined as the ratio of the overall survival to the matched general population survival. Particularly within medical research, relative survival is commonly used to quantify net survival, i.e., the survival of a patient population in the hypothetical scenario where only deaths due the disease can occur [4]. For many medical conditions, patient survival is inferior to that of the general population. However, for some diseases, the patient mortality rate reaches the

* Corresponding author at: Department of Mathematical Sciences, Aalborg University, Skjernvej 4A, Aalborg Ø 9220, Denmark.
E-mail address: lasse.j@rn.dk (L.H. Jakobsen).

https://doi.org/10.1016/j.cmpb.2022.107125

mortality rate of the general population with time. In such cases, cure models are useful for estimating the proportion of patients with expected survival similar to that of the general population and to examine the excess mortality for those who do not experience a similar survival.

Despite the early introduction of cure models, they are still not used regularly in medical research. This may be due to several reasons including challenges with interpreting estimated coefficients, the lack of well-maintained software to support the analyses, and a more complicated variable selection process as variables need to be selected for two sets of model parameters.

Cure models are available through a variety of R-packages with the majority focusing on mixture cure models. The `smcure` package fits semi-parametric mixture cure models based on an EM algorithm [5]. The `npcure` [6] and `curephEM` [7] packages provide functionalities for non-parametric inference in mixture cure models. The `nltm` package [8] implements a method for fitting cure models in the framework of non-linear transformation models. The `miCoPTCM` package enables fitting of semi-parametric non-mixture (or promotion time) cure models incorporating measurement errors in covariates. The `GORCure` package [9] enables fitting of mixture cure models based on interval-censored data, while the `geecure` package fits parametric and semi-parametric mixture cure models using a generalized estimation equation scheme that allows for modelling of clustered time-to-event data [10]. The `penPHcure` package implements a variable selection procedure for mixture cure models using a penalization scheme and allows for time-varying covariates [11]. The `mixcure` package fits parametric and semi-parametric cure models utilizing existing R-packages [12], and the `spduration` package contains functions for fitting parametric mixture cure models based on simple distributions such as the Weibull or log-logistic distributions. However, none of these packages enable fitting of cure models for the relative survival function.

The R-package `flexsurvcure`, which builds upon the comprehensive `flexsurv` package [13] enables fitting of cure models based on simple parametric distributions. Additionally, the package allows for specification of hazard rates from the general population, which allows estimation of cure models for the relative survival function. However, the `flexsurvcure` package utilizes simple parametric distributions, which may not be sufficiently flexible to yield a good model fit.

Here, we describe the R-package `cuRe`, which includes a number of useful functionalities for fitting different types of cure models for ordinary time-to-event survival functions and relative survival functions. These include cure models based on simple and composite parametric distributions as well as spline-based cure models. We also introduce latent cure models which can be fitted using the `rstpm2` package. These models are introduced in Section 2. In the present article, we will only describe cure models for the relative survival function, but the implementation in the two packages also enable fitting of cure models for the ordinary survival distribution. In Section 3, we demonstrate how cure models are fitted using the `cuRe` and `rstpm2` packages and we conclude with a discussion in Section 4.

2. Relative survival and cure models

2.1. Relative survival

We consider the case where a patient population is followed until they experience the event of interest or become right-censored and we assume that the censoring time is independent from the event time. Let T be the time to event for all patients and let $S(t \mid \boldsymbol{z})$ be the survival function for an individual with covariate vector $\boldsymbol{z}$. Additionally, we denote by $S^*(t \mid \boldsymbol{z})$ the survival function of an individual from the general population with covariate vector $\boldsymbol{z}$. Typically, the general population survival function can be stratified by age, sex, and calendar year. The relative survival is the ratio of the patient survival to the general population survival, i.e.:

$$R(t \mid \boldsymbol{z}) = \frac{S(t \mid \boldsymbol{z})}{S^*(t \mid \boldsymbol{z})}. \tag{1}$$

Using the relationship between the hazard function and the survival function, this corresponds to an additive hazards model:

$$h(t \mid \boldsymbol{z}) = h^*(t \mid \boldsymbol{z}) + \lambda(t \mid \boldsymbol{z}), \tag{2}$$

where h and h^* are the hazard functions corresponding to the patient population and the general population, respectively, while λ is termed the *excess hazard*. There is a one-to-one correspondence between R and λ. Modelling of λ has received much attention with the most popular choices being proportional excess hazards models [14]. For diseases where the mortality reaches the same levels as seen in the general population, e.g., colon cancer and specific lymphoma and leukemia subtypes [15–18], the relative survival will reach a plateau [19]. In that scenario, cure models can be used to quantify (1) the level of the plateau, and (2) the trajectory of the relative survival until the plateau is reached.

If a parametric distribution, such as the Weibull distribution, is specified for R (or λ), then the parameters of the distribution, $\boldsymbol{\theta}$, can be estimated using maximum likelihood estimation. The log-likelihood function for right-censored data is given as

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \delta_i \log\left(h^*(t_i \mid \boldsymbol{z}_i) + \lambda(t_i \mid \boldsymbol{z}_i, \boldsymbol{\theta})\right) + \log\left(R(t_i \mid \boldsymbol{z}_i, \boldsymbol{\theta})\right), \tag{3}$$

where δ_i is the indicator of patient i having experienced the event of interest during the follow-up, and t_i is the observed follow-up time of patient i. Importantly, the log-likelihood function does not require any information from S^*, but only uses $h^*(t_i \mid \boldsymbol{z}_i)$, i.e., the hazard rate in the general population at the observed follow-up time. This can be directly retrieved from national or regional life tables, which are often made publicly available. The Human Mortality Database provides a comprehensive archive of life tables from numerous countries [20]. Even though the likelihood function associated with the ordinary survival function is a special case of (3), where $h^*(t_i \mid \boldsymbol{z}_i) = 0$ for all i, we will only focus on cure models for the relative survival for the remainder of this article.

2.2. Mixture cure models

Among cure models, the mixture cure model is often presented as the most intuitive. It is formulated as

$$R(t \mid \boldsymbol{z}) = \pi(\boldsymbol{z}) + (1 - \pi(\boldsymbol{z}))S_u(t \mid \boldsymbol{z}). \tag{4}$$

That is, the relative survival is modelled as a mixture of two relative survival functions originating from the cured and non-cured proportion of patients, respectively; namely a constant relative survival function of 1 and a relative survival function, S_u, which is estimated from the data. The mixture proportion π is the modelled proportion of cured individuals, which is often of primary interest. Using a parametric formulation of both π and S_u, the cure model can be fitted by optimizing (3). For π, covariates are typically modelled linearly with the use of appropriate link functions, such as logit ($\pi(\boldsymbol{z}) = \exp(\boldsymbol{z}^\top\boldsymbol{\theta})/(\exp(\boldsymbol{z}^\top\boldsymbol{\theta})+1)$) and cloglog (complementary log-log; $\pi(\boldsymbol{z}) = 1-\exp(-\exp(\boldsymbol{z}^\top\boldsymbol{\theta}))$), to ensure fitted values between 0 and 1. Simple parameteric distibutions, such as the Weibull, exponential, or log-normal distribution have been investigated for S_u [21]. More complicated models, such as mixtures of Weibull and exponential models, mixtures of two Weibull models, the generalized modified Weibull, the generalized gamma, and the generalized exponential distribution have been proposed as more flexible alternatives to the simple distributions [22–25]. Note that covariates included in the modelling of π and S_u do not need to be identical.

A more general class of parametric mixture cure models has also been introduced [26]. This class is formulated as in (4) with the specification that

$$g_1(S_u(t \mid \boldsymbol{z})) = \eta_1(t, \boldsymbol{z}, \boldsymbol{\theta}) \quad \text{and} \quad g_2(\pi(\boldsymbol{z})) = \eta_2(\boldsymbol{z}, \boldsymbol{\theta}), \tag{5}$$

where g_1 and g_2 are appropriately chosen link functions, preferably such that fitted values of π and S_u remain between 0 and 1, while η_1 and η_2 denote linear predictors. This generalized cure model allows for the use of splines to model the time-effects associated with S_u, which increases model flexibility compared to the simple parametric distributions. Choosing $g_1(x) = \log(-\log(x))$, and modelling the time-effect in $\eta_1(t, \boldsymbol{z}, \boldsymbol{\theta})$ by a restricted cubic spline function $s_0(t, \boldsymbol{\gamma}_0)$, such that $\eta_1(t, \boldsymbol{z}, \boldsymbol{\theta}) = s_0(t, \boldsymbol{\gamma}_0) + \boldsymbol{z}^\top\boldsymbol{\beta}$ and $\boldsymbol{\theta} = (\boldsymbol{\gamma}_0, \boldsymbol{\beta})$, we arrive at the Royston–Parmar proportional hazards (PH) model for S_u [27]. Likewise, a proportional odds model is obtained by choosing $g_1(x) = \log(x^{-1}-1)$.

2.3. Latent cure models

Another type of cure model is the latent cure model, which differs from the mixture cure models described above by not modelling the cure proportion $\pi(\boldsymbol{z})$ explicitly [26]. In these models, the cure proportion is simply computed as the asymptote of the relative survival function. In particular, the relative survival is assumed to be constant after some time point, t_c, i.e, $\pi(\boldsymbol{z}) = R(t \mid \boldsymbol{z}) = R(t_c \mid \boldsymbol{z})$ for all time points after t_c. Using (2), this implies that $h(t \mid \boldsymbol{z}) = h^*(t \mid \boldsymbol{z})$, yielding $\lambda(t \mid \boldsymbol{z}) = 0$ for $t > t_c$. That is, no excess mortality exists beyond t_c. The relative survival of the uncured is then obtained by assuming an underlying mixture cure model (4), yielding

$$S_u(t \mid \boldsymbol{z}) = \frac{R(t \mid \boldsymbol{z}) - \pi(\boldsymbol{z})}{1-\pi(\boldsymbol{z})} = \frac{R(t \mid \boldsymbol{z}) - R(t_c \mid \boldsymbol{z})}{1-R(t_c \mid \boldsymbol{z})}. \tag{6}$$

The latent cure model is formulated similarly to the generalized parametric survival model introduced by Liu et al. [28]. That is,

$$g(R(t \mid \boldsymbol{z})) = \eta(t, \boldsymbol{z}, \boldsymbol{\theta}), \tag{7}$$

where η is a linear predictor modelling the time and covariate effects similarly to η_1 in (5). Since the model is fully parametric, it can be fitted by maximizing the likelihood function in (3). To incorporate a cure proportion into the model, $\eta(t, \boldsymbol{z}, \boldsymbol{\theta})$ needs to be formulated such that it is constant after t_c. However, as the previously introduced parametric distributions do not naturally incorporate such a restriction, other methods for modelling the time-effects in $\eta(t, \boldsymbol{z}, \boldsymbol{\theta})$ have to be employed. In Section 3.4 we will demonstrate how this can be carried out using the functionalities in the `rstpm2` package.

3. R-package functionalities

The R-package `cuRe` contains three main functions for fitting cure models. Firstly, for relative survival analyses, h^* is needed to compute the likelihood function. To extract h^* for each individual, the `general.haz` function can be used. Secondly, once the general population hazard values have been obtained, cure models can be fitted with the `fit.cure.model` or the `GenFlexCureModel` function. The `fit.cure.model` function uses simple parametric distributions to model S_u, but also allows for more complex parametric distributions, such as a Weibull-Weibull mixture or the generalized modified Weibull distribution [21,23]. The `GenFlexCureModel` function uses the link function-based formulation in (5) and is well suited for spline-based modelling of the time effects in S_u.

The `rstpm2` package can be used to fit latent cure models. Cure models are fitted through the `stpm2` function and subsequent calls to the associated `predict` function enables the user to extract estimated cure proportions and the estimates of S_u.

3.1. General population hazards

For illustrative purposes, we will use the `colonDC` dataset from the `cuRe` package, which contains data on 15,564 patients diagnosed with colon cancer. For each patient, disease stage is categorized as either localised, regional, or distant. We exclude patients with localised or missing disease stage, which leads to a total of 6934 patients for analysis. For compatibility with the `stpm2` function presented in this chapter, we create a dummy variable of the stage variable, `stageDistant`, with regional stage as reference, using the following code:

```
> colonDC <- subset(cuRe::colonDC, stage
> colonDC <- transform(colonDC,
+                      stage = factor(stage),
+                      stageDistant = as.numeric(stage == "Distant"))
```

The follow-up time, coded as `FU` in `colonDC`, measures the number of days from diagnosis until death or censoring, and the `status` variable indicates whether the follow-up time is a death time (1) or a censoring time (0). Additionally, patient age measured in days is stored in `agedays`, sex is stored in the variable `sex`, and the diagnosis date is stored in the variable `dx`.

For optimizing the likelihood function in (3), we need the general population hazard function corresponding to each individual. The `general.haz` function extracts the yearly general population hazard values and requires specification of the matching variables corresponding to the stratification level of the applied life table. The life table is inputted as a `ratetable` object. Life tables from the Human Mortality Database can be loaded and formatted to a `ratetable` object by using the `transrate.hmd` function from the `relsurv` package [20,29]. In the following, we will use the `survexp.dk` ratetable. Note that all covariates specified in the `ratetable` should be present in `data` and specified by the `rmap` argument. The hazard values are extracted by using the following:

```
> colonDC$bhaz <- general.haz(time = "FU",
+                             rmap = list(age = "agedays", sex = "sex", year = "dx"),
+                             data = colonDC, ratetable = survexp.dk)
> head(colonDC$bhaz)

[1] 0.05270467 0.08877633 0.07826408 0.07975753 0.05342173 0.05342173
```

3.2. Simple parametric cure models

The `fit.cure.model` function is an implementation of a fitting procedure for the mixture cure model in (4) based on simple parametric distributions. A fully parametric mixture cure model can be fitted using the following code:

```
> fit.wei <- fit.cure.model(formula = Surv(FUyear, status) ~ stageDistant,
+                           formula.surv = list(~ stageDistant, ~ 1),
+                           type = "mixture", dist = "weibull", link = "logit",
+                           bhazard = "bhaz", data = colonDC)
```

The `formula` and `formula.surv` arguments specify the formulas for the cure proportion, π, and the survival of the uncured, S_u, respectively. The argument `formula.surv` is provided as a list of formulas corresponding to the different parameters in the parametric distribution assumed for S_u. In this case, S_u is modelled using the Weibull distribution (see the `dist` argument), $S_u(t) = \exp\left(-\theta_1 t^{\theta_2}\right)$, and the first and last element of the list corresponds to θ_1 and θ_2, respectively. In most cases, the parameters θ are estimated using `log` as link function to ensure $\theta > 0$. In this particular example, θ_1 is modelled by $\log(\theta_1) = \theta_{1,0} + \text{stage} \times \theta_{1,1}$. With θ_2 being independent of covariates, the model for S_u is a PH model. In particular, as we are fitting a cure model for the relative survival, this produces a proportional excess hazards model for S_u. The argument `link` specifies the link function for π, and the general population hazard values are described through the `bhazard` argument. If this latter argument is omitted, the cure model is fitted for the ordinary (overall) survival function.

The `summary` function is used directly on the fitted cure model and yields the parameter estimates as well as standard errors, z-values (Wald-statistics), and p-values:

```
> summary(fit.wei)

                     Estimate    StdErr  z.value   p.value
pi.(Intercept)      -0.381088  0.083840  -4.5454 5.482e-06 ***
pi.stageDistant     -1.866213  0.101035 -18.4709 < 2.2e-16 ***
theta1.(Intercept)  -0.916647  0.062247 -14.7260 < 2.2e-16 ***
theta1.stageDistant  1.130157  0.064922  17.4079 < 2.2e-16 ***
theta2.(Intercept)  -0.055546  0.012537  -4.4306 9.396e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Type = mixture
Link = logit
LogLik(model) = -7801.27
```

The prefixes in the coefficient names indicate which term the coefficients are related to. For instance, the prefix `pi` indicates that the coefficient is related to the cure proportion, π. We emphasize that the displayed estimates do not correspond to differences in π and θ_i's between regional and distant disease stage, but rather differences on the link-transformed scale. To obtain the cure proportion for a patient with specific covariates, the `predict` function can be used. Here, we predict the cure rate of patients having regional and distant disease stages:

```
predict(fit.wei, data.frame(stageDistant = c(0,1)), type = "curerate")

[[1]]
   Estimate     lower     upper
1 0.4058645 0.3661007 0.4452117

[[2]]
    Estimate      lower     upper
1 0.09558257 0.08593149 0.1058279
```

The predictions are supplemented by 95% confidence intervals computed using the delta method. Through the `type` argument, the `predict` function also enables predictions on other scales such as the relative survival (`type = ''surv''`), the relative survival of the uncured (`type = ''survuncured''`), the excess hazard (`type = ''hazard''`), and the excess hazard of the uncured (`type = ''hazarduncured''`).

Table 1
Estimated cure rate, odds ratio, 5 year survival of the uncured and relative survival, and excess hazard ratio of the uncured for each of the two disease stages estimated by the three fitted mixture models with corresponding 95% confidence intervals. The model `fit.gen.timevar` estimates the excess HR as time-varying and so the excess HR is not provided. OR: Odds ratio. HR: Hazard ratio.

Model	Stage	Cure rate	OR	S_u (5y)	R (5y)	Excess HR
fit.wei	Regional	41% (37–45%)	1.00 (ref.)	16% (13–20%)	50% (47–53%)	1.00 (ref.)
	Distant	10% (9–11%)	0.15 (0.13–0.19)	0% (0–0%)	10% (9–11%)	3.10 (2.73–3.52)
fit.gen	Regional	35% (29–42%)	1.00 (ref.)	25% (19–33%)	52% (49–54%)	1.00 (ref.)
	Distant	8% (7–10%)	0.17 (0.13–0.22)	1% (1–2%)	10% (9–10%)	3.22 (2.44–4.25)
fit.gen.timevar	Regional	41% (36–46%)	1.00 (ref.)	14% (8–24%)	49% (47–52%)	
	Distant	8% (7–9%)	0.12 (0.10–0.16)	2% (1–3%)	9% (9–10%)	

3.3. Spline-based cure models with `cuRe`

The `GenFlexCureModel` function fits the model formulated in (5). Although some simple parametric cure models are included in this general model formulation, the function is primarily intended for fitting cure models where the time-effects are modelled using smoothers, such as splines. Using this function, we fit a generalized mixture cure model:

```
> fit.gen <- GenFlexCureModel(formula = Surv(FUyear, status) ~ stageDistant,
+                             smooth.formula = ~ ns(log(FUyear), df = 3),
+                             cr.formula = ~ stageDistant,
+                             bhazard = "bhaz", data = colonDC)
```

The argument `bhazard` in `GenFlexCureModel` is identical to the corresponding argument in `fit.cure.model`. The cure proportion in the above mixture cure model is modelled similarly to the cure proportion in the Weibull cure model in Section 3.2, but the cure proportion formula is now instead specified by the `cr.formula` argument. The `formula` provides the time-invariant covariate effects for S_u, whereas time effects for S_u are described in the `smooth.formula` argument. The survival of the uncured is modelled using the `ns` function, which provides the basis functions for natural cubic splines. The `cuRe` package also includes the function `cb`, which compute the basis functions for the restricted cubic splines described by Royston and Parmar [27]. The above model for S_u is thus specified as

$$g_1(S_u(t \mid \boldsymbol{z})) = s_0(x, \boldsymbol{\gamma}_0) + \beta z, \tag{8}$$

where z is an indicator of regional or distant disease stage, $x = \log(t)$, and s_0 is a natural cubic spline with coefficients $\boldsymbol{\gamma}_0$. The default link function, g_1, is the log-log link function corresponding to a PH model, since no time-varying covariate effects are included. Similar to the Weibull model, this corresponds to an excess proportional hazards model with the baseline hazard modelled using natural cubic splines.

Coefficient estimates can be displayed using the `summary` function:

```
> summary(fit.gen)

Call - pi:
Surv(FUyear, status) ~ stageDistant
             Estimate   StdErr  z.value   p.value
(Intercept)  -0.60727  0.14087  -4.3107 1.628e-05 ***
stageDistant -1.77471  0.14159 -12.5342 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Call - surv - baseline: Surv(FUyear, status) ~ stageDistant + ns(log(FUyear), df = 3)
                          Estimate    StdErr  t.value   p.value
(Intercept)              -4.551612  0.140875  -4.3107 1.628e-05 ***
stageDistant              1.168904  0.141589 -12.5342 < 2.2e-16 ***
ns(log(FUyear), df = 3)1  3.244219  0.099868 -45.5764 < 2.2e-16 ***
ns(log(FUyear), df = 3)2  7.304823  0.081190  14.3971 < 2.2e-16 ***
ns(log(FUyear), df = 3)3  4.723612  0.055489  58.4661 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Type = mixture
Link - pi = logit
Link - surv =  PH
LogLik(model) = 7696.629
```

As the modelling of S_u is now based on natural cubic splines, additional parameters (compared to the Weibull model) for the spline terms, `ns(log(FUyear), df = 3)`, are included in the output. Due to the proportional excess hazard formulation of S_u, the parameter estimates for disease stage in the Weibull model `fit.wei` and the generalized model `fit.gen` can be interpreted similarly, as we will do in Table 1.

Additionally, the `GenFlexCureModel` function enables time-varying covariate effects. These can be included in the model for S_u by also specifying a `tvc.formula` argument, which controls all time-varying covariate effects:

```
> fit.gen.timevar <- GenFlexCureModel(formula = Surv(FUyear, status) ~ stageDistant,
+                                     smooth.formula = ~ ns(log(FUyear), df = 3),
+                                     tvc.formula = ~ ns(log(FUyear), df = 3):stageDistant,
+                                     cr.formula = ~ stageDistant,
+                                     bhazard = "bhaz", data = colonDC)
```

This specifies a mixture cure model in which S_u is modelled as:

$$g_1(S_u(t \mid \boldsymbol{z})) = s_0(x, \boldsymbol{\gamma}_0) + s_1(x, \boldsymbol{\gamma}_1)z. \tag{9}$$

Thus, the coefficient corresponding to z (indicator of regional or distant disease stage) is now time-varying with the use of a natural cubic spline, s_1. Note, the number of knots (or degrees of freedom) of the splines for the baseline hazard and the time-varying covariate effects do not need to be identical.

Again, coefficients can be displayed using the `summary` function:

```
> summary(fit.gen.timevar)

Call - pi:
Surv(FUyear, status) ~ stageDistant
             Estimate   StdErr  z.value   p.value
(Intercept)  -0.36165  0.11078  -3.2645  0.001097 **
stageDistant -2.09066  0.13363 -15.6455 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Call - surv - baseline: Surv(FUyear, status) ~ stageDistant + (ns(log(FUyear), df = 3) +
    ns(log(FUyear), df = 3):stageDistant)
                                     Estimate   StdErr  t.value   p.value
(Intercept)                          -4.34567  0.11078  -3.2645 0.0010967 **
stageDistant                          0.93665  0.13363 -15.6455 < 2.2e-16 ***
ns(log(FUyear), df = 3)1              3.61526  0.24361 -17.8390 < 2.2e-16 ***
ns(log(FUyear), df = 3)2              6.79812  0.25246   3.7101 0.0002072 ***
ns(log(FUyear), df = 3)3              5.44629  0.23241  15.5552 < 2.2e-16 ***
stageDistant:ns(log(FUyear), df = 3)1 -0.38732  0.43220  15.7291 < 2.2e-16 ***
stageDistant:ns(log(FUyear), df = 3)2  0.48057  0.28925  18.8291 < 2.2e-16 ***
stageDistant:ns(log(FUyear), df = 3)3 -0.91685  0.23904  -1.6203 0.1051676
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Type = mixture
Link - pi = logit
Link - surv =  PH
LogLik(model) = 7666.222
```

The `predict` function can again be used to obtain predictions for a wide range of different scales. Because the fitted models, `fit.wei`, `fit.gen`, and `fit.gen.timevar`, use a logit link function for the cure proportion, the odds ratio of being cured can be retrieved by applying $exp()$ to the coefficient estimates corresponding to the cure rate, π. The excess hazard ratio of the uncured can be interpreted as the degree to which colon cancer patients with distant stage also experience a worse survival within the uncured group. It is obtained by applying $exp()$ to the coefficients corresponding to S_u, although this only applies to models with a proportional hazards formulation for S_u, excluding the `fit.gen.timevar` model. The `fit.gen.timevar` model demonstrates a large difference between the cure proportion among patients with regional and distant stage, and even within uncured patients, the relative survival is worse among patients with distant stage (Fig. 1). Similar conclusions were made based on the `fit.wei` and `fit.gen` models (Table 1).

3.4. Spline-based cure models with `rstpm2`

The `rstpm2` package contains functions for fitting a general class of parametric survival models, including the use of penalized splines. The package can also be used for fitting latent cure models. These can be fitted using the `stpm2` function:

```
> fit.lat <- stpm2(formula = Surv(FUyear, status) ~ stageDistant + bhazard(bhaz),
+                  df = 4, cure = TRUE,
+                  data = colonDC)
```

The call fits the model in (7) using the default log-log link function and by modelling the time-effect in the linear predictor using natural cubic splines. Instead of using `ns`, the function uses an extended implementation, namely `nsx`. If `cure = TRUE`, `nsx` uses a QR-decomposition approach which ensures that the splines become constant beyond the last knot [26]. A similar extension, `bsx`, has been made to `bs` and can be found in the `cuRe` package. Additionally, Andersson et al. [30] altered the basis functions of the restricted cubic splines by Royston and Parmar such that the spline becomes constant beyond its last knot. These splines are implemented in the `cbc` function of the `cuRe` package. The `bsx` and `cbc` functions from the `cuRe` package are compatible with `rstpm2` package.

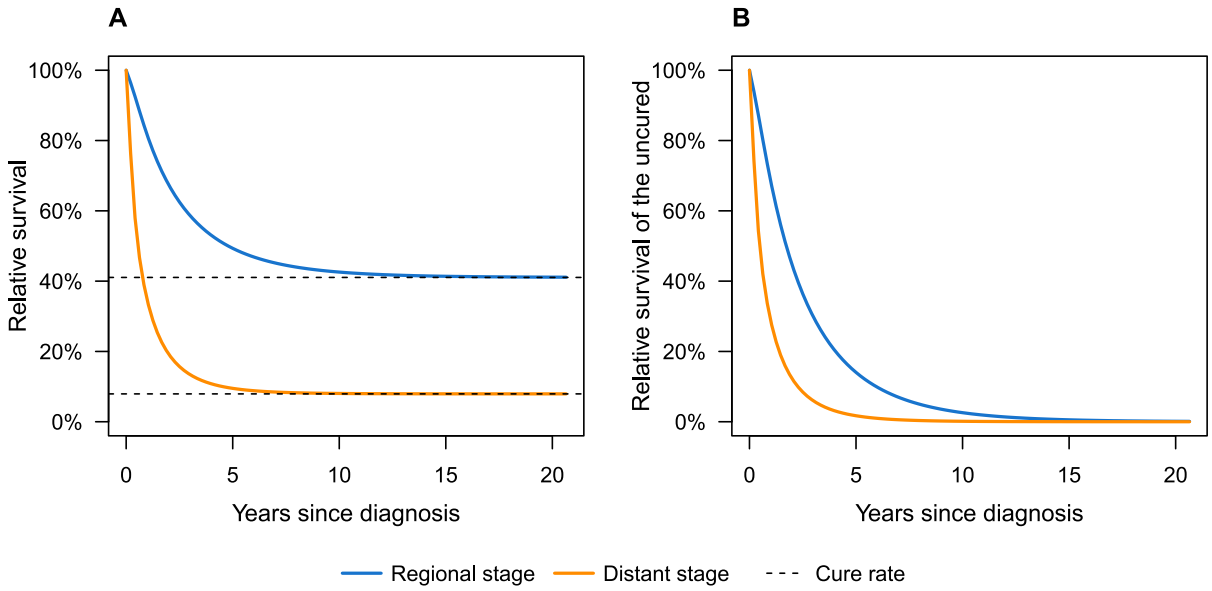

Fig. 1. Relative survival (A) and relative survival of the uncured proportion of patients (B) for different disease stages, estimated using the mixture cure model `fit.gen.timevar`.

Similarly to the `fit.gen.timevar`, time-varying covariate effects can be modelled through interaction terms between covariates and a spline on the time scale. The time-varying coefficients are explicitly specified in the `tvc` argument:

```
> fit.lat.timevar <- stpm2(formula = Surv(FUyear, status) ~ stageDistant + bhazard(bhaz),
+                          df = 4, cure = TRUE,
+                          tvc = list(stageDistant = 4),
+                          data = colonDC)
```

When specifying the `tvc` argument, the time-varying coefficients are modelled using the `nsx` function. Estimation of the cure proportion π and the survival of the uncured S_u is not explicitly performed in the latent cure models. However, since the relative survival is restricted to be constant after the last knot, the cure proportion can be estimated as the relative survival at the boundary knot. This can be achieved by using the `predict` function:

```
> bdr.knot <- exp(max(as.list(as.list(attr(fit.lat.timevar@terms, "predvars"))[[4]])$Boundary.knots))

> predict(fit.lat, newdata = data.frame(FUyear = bdr.knot, stageDistant = 0),
+         type = "surv", se.fit = TRUE)

   Estimate     lower     upper
1 0.4972138 0.4692979 0.5267903
```

That is, the cure proportion is approximately 50% for patients with regional stage colon cancer. The relative survival of the uncured specified in (6) is also available though the `predict` function by specifying `type = ''uncured''` and the time point at which predictions should be made. Here, the two-year relative survival of the uncured is computed:

```
> predict(fit.lat, newdata = data.frame(FUyear = 2, stageDistant = 0),
+         type = "uncured", se.fit = TRUE)
   Estimate     lower     upper
1 0.3093215 0.2902189 0.3296814
```

The entire relative survival curve and survival curve of the uncured are displayed in Fig. 2. It is seen from this figure that according to both latent models, cure is reached within five years of diagnosis for almost all of the patients who are eventually cured. Both models attribute a large effect of disease stage on the probability of eventually being cured, with almost none of the patients having distant stage colon cancer being cured.

3.5. Conditional probability of cure

The cure models presented so far yield estimates of the cure proportion (probability) and the relative survival for the uncured patients. Additionally, in some scenarios it would be useful to estimate the conditional probability of cure given survival up to a certain time point. This is also enabled through the `cuRe` and `rstpm2` packages.

If we let Y be a stochastic variable taking the value 1 if a patient is cured (i.e., have the same survival expectancy as the general population) and 0 if uncured, we have $\pi(\boldsymbol{z}) = \mathbb{P}(Y = 1 \mid \boldsymbol{z})$. Some algebra yields the following equality:

$$\mathbb{P}(Y = 1 \mid T > t, \boldsymbol{z}) = \frac{\pi(\boldsymbol{z})}{R(t \mid \boldsymbol{z})}. \tag{10}$$

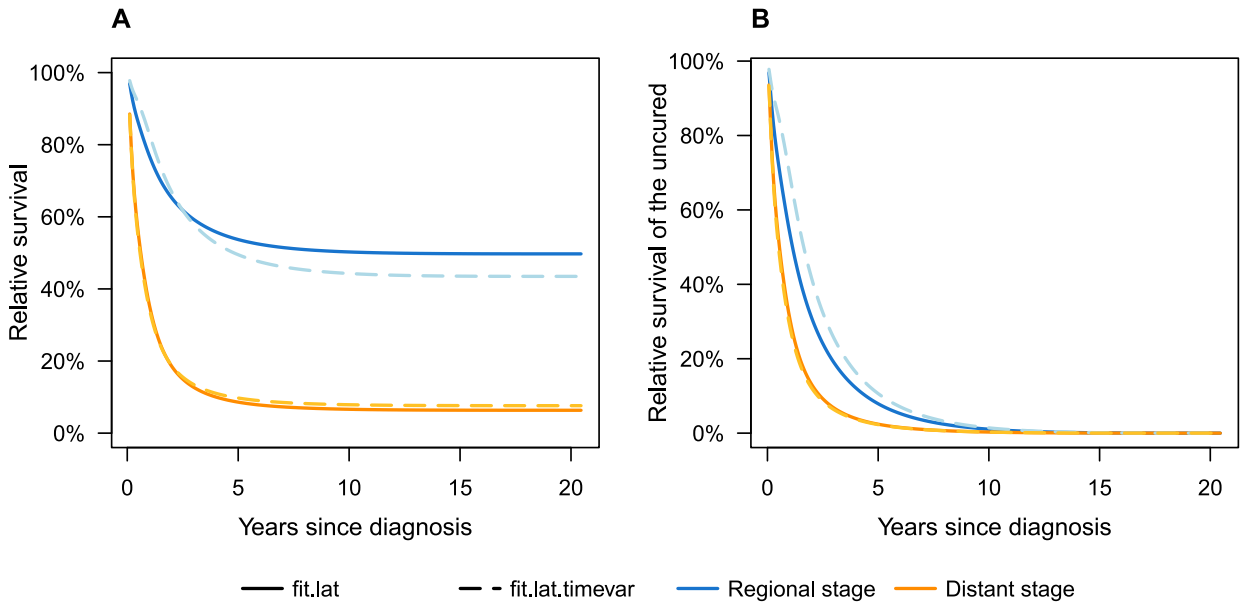

Fig. 2. Relative survival (A) and relative survival of the uncured proportion of patients (B) for different disease stages, estimated using the latent cure models `fit.lat` and `fit.lat.timevar`.

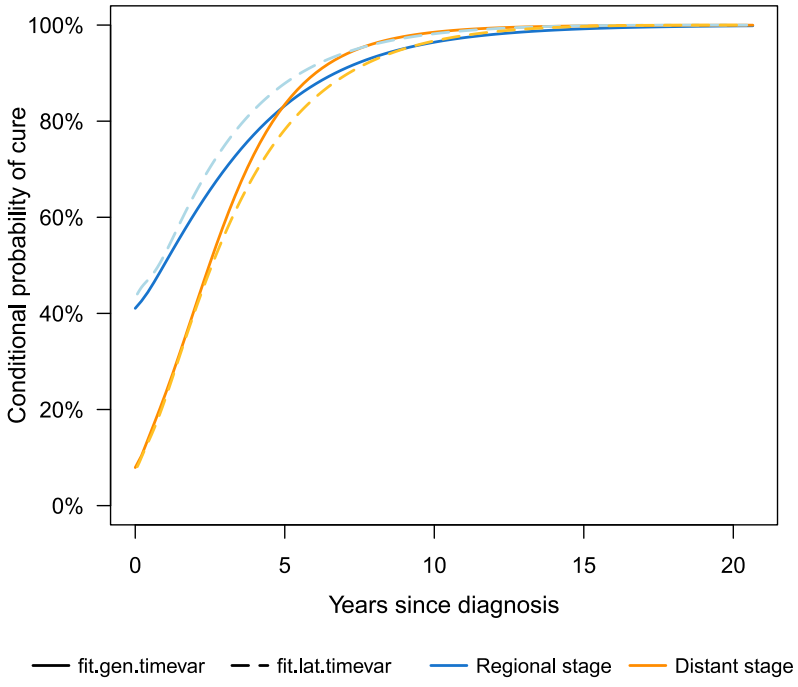

Fig. 3. Conditional probability of cure for different disease stages, estimated using the mixture cure model `fit.gen.timevar` and the latent cure model `fit.lat.timevar`.

That is, an estimate of the conditional probability of cure can be derived from the fitted cure models. If R is decreasing, the conditional probability of cure monotonously approaches 1 as time progresses.

For cure models obtained from the `cuRe` and `rstpm2` packages, the conditional probability of cure can be computed by the `predict` function by specifying `type = ''probcure''`:

```
> predict(fit.lat.timevar, newdata = data.frame(FUyear = 2, stageDistant = 0),
+         type = "probcure", se.fit = TRUE)
   Estimate     lower     upper
1 0.6495372 0.6038926 0.6913162
```

Here, we predict the probability of cure for a patient with regional disease stage, conditional on survival up to 2 years. Evaluating this probability at multiple time points produces the curves in Fig. 3. Note, for the latent cure model, the conditional probability of cure is equal to 1 beyond t_c, i.e., the time point from which $R(t)$ is equal to π [26].

Combining the equality in (10) and the definition $\mathbb{P}(Y = 1 \mid \boldsymbol{z}) = \pi(\boldsymbol{z})$, the cure proportion can also be extracted by using the `type = ''probcure''` argument and specifying `FUyear = 0`.

4. Discussion

We presented functionalities in R for fitting parametric cure models applicable to both the total survival and relative survival. The `cuRe` package contains two functions for computing parametric mixture cure models, namely `fit.cure.model` and `GenFlexCureModel`, which enable modelling using a wide range of parametric distributions and splines. The implementation of parametric latent cure models in the `rstpm2` package provides a flexible framework for fitting spline-based cure models via the `stpm2` function. Both packages are available on the Comprehensive R Archive Network (CRAN).

A substantial part of the cure model literature is concerned with a third type of cure models, namely promotion time cure models, also known as non-mixture cure models. Details on this model class can be found elsewhere [21,31]. Fitting of non-mixture cure models is possible through both the `fit.cure.model` and `GenFlexCureModel` functions of the `cuRe` package. Both functions contain the argument `type`, which should be set to `nmixture` if a non-mixture cure model is of interest. The `cuRe` package also contains functions for computing a number of useful post-estimation measures, such as likelihood ratio tests through the `lrtest` function, which can be used on nested models to assess the significance of a chunk of parameters. This can be used for testing if the model improves significantly by including time-varying coefficients for a variable. Furthermore, functions for estimating the loss of lifetime and the time to statistical cure, i.e., the time at which the relative survival reaches a plateau, are implemented in the package. However, the latter comes with a number of challenges which were outlined in our recent work [19].

The package does not contain functions for conducting model diagnostics as the literature on this within the cure model framework is sparse. The Cox-Snell-like residuals introduced by Scolas et al. [32] can be used, but these are not immediately applicable to cure models for the relative survival.

Although we have only focused on cure models for the relative survival in the present article, the `cuRe` and `rstpm2` packages also enable fitting of parametric cure models for the overall survival function. For the overall survival function, the cure proportion equals the probability of never experiencing the event of interest. Within medical research, investigation of outcomes that may never occur are often subject to competing events, particularly death. In competing risks analyses, focus commonly changes from the survival function to the cumulative incidence function. Thus, for many applications it would be of interest to estimate a cure proportion on the cumulative incidence scale instead of the survival function. For instance, it may be of interest to estimate the proportion of cancer patients that will never receive a disability pension. In such a case, retirement pension and death act as competing events to the primary event of interest (i.e., disability pension) [33]. The `cuRe` and `rstpm2` packages do not at present contain functionalities for computing the cure proportion on the cumulative incidence scale. However, the Stata command `stpm2cr` enables this by providing a framework for fitting spline-based models on the cumulative incidence function from which the cure proportion can be estimated [34].

Although the spline-based functionalities available through the `rstpm2` package provide a flexible framework for fitting parametric latent cure models, the use of splines requires the user to specify the number and location of the knots. To mitigate this problem, penalized splines can be utilized as proposed by Xiu et al. [28]. Future work will include the development of penalized splines that are directly applicable in latent cure models.

Declaration of Competing Interest

Authors declare that they have no conflict of interest.

References

[1] J. Boag, Maximum likelihood estimates of the proportion of patients cured by cancer therapy, J. R. Stat. Soc. Ser. B (Methodological) 11 (1949) 15–53.
[2] V. Bremhorst, M. Kreyenfeld, P. Lambert, Fertility progression in Germany: an analysis using flexible nonparametric cure survival models, Demogr. Res. 35 (2016) 505–534.
[3] M. Othus, B. Barlogie, M. Leblanc, J. Crowley, Cure models as a useful statistical tool for analyzing survival, Clin. Cancer Res. 18 (14) (2012) 3731–3736.
[4] M.P. Perme, J. Stare, J. Estëve, On estimation in relative survival, Biometrics 1 (68) (2012) 113–120.
[5] C. Cai, Y. Zou, Y. Peng, J. Zhanga, Smcure: an R-package for estimating semiparametric mixture cure models, Comput. Methods Programs Biomed. 108 (3) (2012) 1255–1260.
[6] A. Löpez-Cheda, M.A. Jácome, I.L. de Ullibarri, Npcure: an R package for nonparametric inference in mixture cure models, R. J. 13 (1) (2021) 21–41.
[7] J. Hou, C.D. Chambers, R. Xu, A nonparametric maximum likelihood approach for survival data with observed cured subjects, left truncation and right-censoring, Lifetime Data Anal. 24 (2018) 612–651.
[8] A. Tsodikov, Semiparametric models: a generalized self-consistency approach, J. R. Stat. Soc. 65 (3) (2003) 759–774.
[9] J. Zhou, J. Zhang, W. Lu, Computationally efficient estimation for the generalized odds rate mixture cure model with interval-censored data, J. Comput. Graph. Stat. 27 (1) (2018) 48–58.
[10] Y. Niu, X. Wang, Y. Peng, Geecure: an R-package for marginal proportional hazards mixture cure models, Comput. Methods Programs Biomed. 161 (2018) 115–124.
[11] A. Beretta, C. Heuchenne, penPHcure: variable selection in proportional hazards cure model with time-varying covariates, R. J. 13 (1) (2021) 116–129.
[12] Y. Peng, B. Yu, Cure Models: Methods, Applications, and Implementation, Chapman and Hall/CRC, 2021.
[13] C.H. Jackson, Flexsurv: a platform for parametric survival modeling in R, J. Stat. Softw. 70 (8) (2016) 1–33.
[14] P.W. Dickman, A. Sloggett, M. Hills, T. Hakulinen, Regression models for relative survival, Stat. Med. 23 (1) (2004) 51–64.
[15] L. Jakobsen, M. Bøgsted, P. Brown, B. Arboe, J. Jørgensen, T. Larsen, M. Juul, L. Schurmann, L. Højberg, O. Bergmann, T. Lassen, P. Josefsson, P. Jensen, H. Johnsen, T. El-Galaly, Minimal loss of lifetime for patients with diffuse large B-cell lymphoma in remission and event free 24 months after treatment: a Danish population-based study, J. Clin. Oncol. 35 (7) (2017) 778–784, doi:10.1200/JCO.2016.70.0765.
[16] L.H. Jakobsen, F. Ellin, K.B. Smeland, T. Wästerlid, J.H. Christensen, J.M. Jørgensen, P.L. Josefsson, A.K. Övlisen, H. Holte, Y.N. Blaker, J.H. Grauslund, J. Bjørn, D. Molin, I. Lagerlöf, K.E. Smedby, K. Colvin, G. Thanarajasingam, M.J. Maurer, T.M. Habermann, K.W. Song, K.Y. Zhu, A.S. Gerrie, C.Y. Cheah, T.C. El-Galaly, Minimal relapse risk and early normalization of survival for patients with burkitt lymphoma treated with intensive immunochemotherapy: an international study of 264 real-world patients, Br. J. Haematol. 189 (4) (2020) 661–671, doi:10.1111/bjh.16425.
[17] M.J. Maurer, H. Ghesquières, J.-P. Jais, T.E. Witzig, C. Haioun, C.A. Thompson, R. Delarue, I.N. Micallef, F. Peyrade, W.R. Macon, T.J. Molina, N. Ketterer, S.I. Syrbu, O. Fitoussi, P.J. Kurtin, C. Allmer, E. Nicolas-Virelizier, S.L. Slager, T.M. Habermann, B.K. Link, G. Salles, H. Tilly, J.R. Cerhan, Event-free survival at 24 months is a robust end point for disease-related outcome in diffuse large B-cell lymphoma treated with immunochemotherapy, J. Clin. Oncol. 32 (10) (2014) 1066–1073.
[18] J.L. Biccler, L.S.G. Östgård, M.T. Severinsen, C.W. Marcher, P. Møller, C. Schöllkopf, L.S. Friis, M. Bøgsted, L.H. Jakobsen, T.C. El-Galaly, J.M. Nørgaard, Evolution of relative survival for acute promyelocytic leukemia patients alive at landmark time-points: a population-based study, Leukemia 32 (10) (2018) 2263–2303.

[19] L.H. Jakobsen, T.M.-L. Andersson, J.L. Biccler, L. Poulsen, M.T. Severinsen, T.C. El-Galaly, M. Bøgsted, On estimating the time to statistical cure, BMC Med. Res. Methodol. 20 (71) (2020).
[20] Human Mortality Database, 2017. University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany). Accessed 15 September 2017, http://www.mortality.org.
[21] P. Lambert, J. Thompson, C. Weston, P. Dickman, Estimating and modeling the cure fraction in population-based cancer survival analysis, Biostatistics 8 (3) (2007) 576–594.
[22] P. Lambert, P. Dickman, C. Weston, J. Thompson, Estimating the cure fraction in population-based cancer studies by using finite mixture models, J. R. Stat. Soc. 59 (1) (2010) 35–55.
[23] E. Martinez, J. Achcar, A. Jácome, J. Santosc, Mixture and non-mixture cure fraction models based on the generalized modified Weibull distribution with an application to gastric cancer data, Comput. Methods Programs Biomed. 112 (3) (2013) 343–355.
[24] P. Wang, S. Pal, A two-way flexible generalized gamma transformation cure rate model, Stat. Med. 41 (2022) 2427–2447, doi:10.1002/SIM.9363. https://pubmed.ncbi.nlm.nih.gov/35262947/
[25] K. Davies, S. Pal, J.A. Siddiqua, Stochastic em algorithm for generalized exponential cure rate model and an empirical study, J. Appl. Stat. 48 (2020) 2112–2135, doi:10.1080/02664763.2020.1786676. https://www.tandfonline.com/doi/abs/10.1080/02664763.2020.1786676
[26] L. Jakobsen, M. Bøgsted, M. Clements, Generalized parametric cure models for relative survival, Biom. J. 62 (4) (2020) 989–1011.
[27] P. Royston, M. Parmar, Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects, Stat. Med. 21 (15) (2002) 2175–2197.
[28] X.-R. Liu, Y. Pawitan, M. Clements, Parametric and penalized generalized survival models, Stat. Methods Med. Res. 27 (5) (2016) 1531–1546.
[29] M.P. Perme, K. Pavlic, Nonparametric relative survival analysis with the R package relsurv, J. Stat. Softw. 87 (8) (2018) 1–27.
[30] T. Andersson, P. Dickman, S. Eloranta, P. Lambert, Estimating and modelling cure in population-based cancer studies within the framework of flexible parametric survival models, BMC Med. Res. Methodol. 11 (1) (2011) 96.
[31] A.D. Tsodikov, J.G. Ibrahim, A.Y. Yakovlev, Estimating cure rates from survival data: an alternative to two-component mixture models, J. Am. Stat. Assoc. 98 (464) (2003) 1063–1078.
[32] S. Scolas, C. Legrand, A. Oulhaj, A. El Ghouch, Diagnostic checks in mixture cure models with interval-censoring, Stat. Methods Med. Res. 27 (7) (2018) 2114–2131.
[33] P. Pedersen, M. Aagesen, L.H. Tang, N.H. Bruun, A.-D. Zwisler, C.M. Stapelfeldt, Risk of being granted disability pension among incident cancer patients before and after a structural pension reform: aDanish population-based, matched cohort study, J. Am. Stat. Assoc. 98 (464) (2003) 1063–1078.
[34] S.I. Mozumder, M.J. Rutherford, P.C. Lambert, A flexible parametric competing-risks model using a direct likelihood approach for the cause-specific cumulative incidence function, Stata J. 17 (2) (2017) 462–489.