Computational Statistics (2021) 36:2467–2489
https://doi.org/10.1007/s00180-021-01086-3

ORIGINAL PAPER

Mixture cure rate models with neural network estimated nonparametric components

Yujing Xie[1] · Zhangsheng Yu[2]

Received: 11 July 2020 / Accepted: 9 February 2021 / Published online: 27 March 2021

Abstract
Survival data including potentially cured subjects are common in clinical studies and mixture cure rate models are often used for analysis. The non-cured probabilities are often predicted by non-parametric, high-dimensional, or even unstructured (e.g. image) predictors, which is a challenging task for traditional nonparametric methods such as spline and local kernel. We propose to use the neural network to model the nonparametric or unstructured predictors' effect in cure rate models and retain the proportional hazards structure due to its explanatory ability. We estimate the parameters by Expectation–Maximization algorithm. Estimators are showed to be consistent. Simulation studies show good performance in both prediction and estimation. Finally, we analyze Open Access Series of Imaging Studies data to illustrate the practical use of our methods.

Keywords Consistency · Deep learning · EM algorithm · Survival analysis

1 Introduction

Survival data including potentially cured subjects are common in clinical studies. The population is a mixture of two types of subjects, namely, the 'cured' and the 'non-cured'. For example, Farewell (1986) showed that the Kaplan–Meier curves of treatment A for breast cancer remained at about 73% after a long follow up period, which suggested that a potential cure fraction. Masud et al. (2016) also showed that there was a proportion of children who were non-susceptible to the asthma using the

✉ Zhangsheng Yu
yuzhangsheng@sjtu.edu.cn

Yujing Xie
qaz_sx@sjtu.edu.cn

[1] School of Mathematical Sciences, Shanghai Jiao Tong University, Shanghai, China

[2] Department of Bioinformatics and Biostatistics, SJTU-Yale Joint Center for Biostatistics, School of Mathematical Sciences, Shanghai Jiao Tong University, Shanghai, China

cure rate model. Cancer patients receiving immunotherapy can also have long term survival (Ferrara et al. 2018).

In practice, one usually examines the Kaplan–Meier curve. If there is a long plateau at the later part of Kaplan–Meier curve, we believe that there may be a subgroup of cured subjects. Mixture cure rate models (MCM) and promotion time cure rate models are two major categories of cure rate models, and the former is considered in our article. The often used MCM was proposed by Kuk and Chen (1992), which consists of a parametric logistic regression model to estimate the probability for cured subpopulation and a Cox model that characterize the failure time distribution for non-cured subjects. Extensive researches have been performed for MCM. Yin and Ibrahim (2005) proposed a novel class of cure rate model, which was formulated through a transformation on the population survival function. Othus et al. (2009) studied cure rate models with dependent censoring without making parametric assumptions. Lu (2010) studied an accelerated failure time model with a cure fraction via kernel-based nonparametric maximum likelihood estimation. Masud et al. (2016) proposed a method of variable selection for cure rate model. The above models used parametric logistic regression models to depict the non-cured probability, however, parametric assumption is overly restrictive in practical application, especially in the complex biological field. Hence Wang et al. (2012) proposed a MCM with nonparametric forms for both the cure rate and the hazard rate function and established good theoretical properties for such model. Their model can well estimate the nonparametric effect of a small number of structured covariates. However, when covariate dimension is large and true non-parametric function contains many interactions, the spline model will bring a large number of parameters to be estimated, which may greatly reduce the model efficiency.

On the other hand, estimating the cured probability is clinically important because some patients with a high cure rate may be protected from the additional risks of high-intensity treatment. For example, Liu et al. (2012) found in a breast cancer clinical trial MA.5 that a new adjuvant chemotherapy regimen can effectively improve the cure rate of patients than the classic regimen, but the chemotherapy regimen was associated with some serious toxicities that can bring the risk of albinism. Identifying the cured probability of patients and then performing different treatments based on it can maximize overall benefit. In practice, predictors of the cured probability may include unstructured data such as images in addition to the structured covariates. For example, in cancer immunotherapy study, patients may be cured of the disease and the cure rate is primarily determined by the biomarkers (e.g. programmed death-ligand 1 (PD-L1) positive rate in cancer cells) which are derived from the histological images (Cho et al. 2017). Current prediction models using image derived biomarkers are labor intensive and subject to substantial human reading error. Incorporating the image as an unstructured predictor directly in MCM will provide a more accurate and automated model prediction. However, traditional non-parametric methods are generally not computationally suitable for processing unstructured predictors, such as high-dimensional images or sparse text. Therefore in this study, we aim to study MCM with structured and unstructured predictors in the cure rate component. We recommend using a neural network to fit the component because it can process not only structured but also unstructured predictors such as image and text data.

Neural networks are popular methods for processing complex data that have been widely employed in recent decades. The earliest neural network originated from the artificial neuron model proposed by McCulloch and Pitts (1943). Rumelhart et al. (1986) described back-propagation learning procedure for networks of neuron-like units to solve the problem of nonlinear classification. Cybenko (1989) first proved the universal approximation theorem for a neural network with sigmoid activation function. Halbert (1990) and Hornik (1991) proved the theorem under less assumptions of the activation functions. Unlike traditional statistical models, neural networks do not need to pre-specify the model form and can be used for a wider range of data types. Besides, Faraggi and Simon (1995) indicated that neural networks are considered by many to be very promising tools for classification and prediction. Therefore, in recent years, there have been some studies dedicated to using neural networks to improve the predictive ability of statistical models. Faraggi and Simon (1995) modeled censored survival data with a simple feed-forward neural network as the basis for a non-linear proportional hazards model. Katzman et al. (2018) applied the deep feedforward neural network to survival data, and proposed a DeepSurv model. They showed that DeepSurv performs as well as or better than other survival analysis methods on survival data with both linear and nonlinear effects from covariates. Ching et al. (2018) have developed a neural network extension of the Cox model, called Cox-nnet. It is optimized for survival prediction from high throughput gene expression data, with comparable or better performance than other conventional methods. Tandon et al. (2006) expanded the linear part of traditional mixed-effects models with neural networks. And they used the proposed model to analyze longitudinal data of Alzheimer's disease. Therefore, we propose to use a neural network to fit a cure rate model with a complex covariant structure, aiming to improve the accuracy of the cure rate estimation. As far as we know, this is the first time that a neural network has been used in a cure rate model. On the other hand, due to the good interpretability and theoretical properties, the proportional hazards structure is preserved in the survival analysis for the non-cured subjects.

We present the model and estimation procedure in Sect. 2. We study consistency and asymptotic properties of the proposed estimators in Sect. 3. We conduct simulation studies to examine the numerical properties of the proposed method in Sect. 4. The method is applied to the analysis of Open Access Series of Imaging Studies (OASIS) data in Sect. 5. We conclude the study with a discussion.

2 Models and estimation

2.1 Model

Define the failure time and censoring time of i th subjects as $\widetilde{T}_i$ and $C_i, i = 1, 2, \ldots, n$. The observed time is $T_i = \min(\widetilde{T}_i, C_i)$. The failure time indicator $\delta_i = 1$ when $T_i = \widetilde{T}_i$ and $\delta_i = 0$ otherwise. Let $Y_i = 1$ be a binary indicator with $Y_i = 1$ denoting the non-cured subjects and $Y_i = 0$ otherwise. We model the non-cured probability $P(Y_i = 1)$ with a nonparametric function $\theta(\cdot)$. Assume that $\boldsymbol{x}_i \in \mathbb{R}^p$ and $\boldsymbol{z}_i \in \mathbb{R}^q$ are the covariates for cure and survival component respectively. The population survival function $S_p(t_i)$ can be expressed as:

$$S_p(t_i) = \{1 - \theta(\boldsymbol{x}_i)\} + \theta(\boldsymbol{x}_i)S(t|\boldsymbol{z}_i), \tag{1}$$

where $S(\cdot)$ is the survival function for the non-cured subjects, given $\boldsymbol{z}_i$. Kuk and Chen (1992) used a logistic regression model to model the probability of a subject in the non-cured group:

$$\theta(\boldsymbol{x}_i) = sig(\boldsymbol{x}_i) = \frac{\exp(\boldsymbol{\beta}^T \boldsymbol{x}_i)}{1 + \exp(\boldsymbol{\beta}^T \boldsymbol{x}_i)}, \tag{2}$$

where $\boldsymbol{\beta} \in \mathbb{R}^p$ is the corresponding regression parameter vector, and $sig(\cdot)$ is sigmoid function. Model (2) assumes a linear effect of covariate $\boldsymbol{x}$ and the logit of probability of cure. Linear assumption is often violated in practice. And in this study, we consider cases where $\boldsymbol{x}$ with large p and can be even unstructured data. As stated in Sect. 1, traditional non-parametric models may be cumbersome unless additional variable screening processes are used in such cases. Hence we propose to use neural network to depict more flexible forms of $\theta(\cdot)$ and $\boldsymbol{x}$. For demonstration, we consider two hidden layers network with k, m finite number of neurons respectively, we write the model as:

$$\theta(\boldsymbol{x}_i) = act[act\{B_2^T act(B_1^T \boldsymbol{x}_i + \boldsymbol{b}_1)^T + \boldsymbol{b}_2\}^T \boldsymbol{\beta}_3 + b_3], \tag{3}$$

where $\boldsymbol{x}_i \in \mathbb{R}^p$ is the covariates, $\boldsymbol{B}_1 \in \mathbb{R}^{p \times k}$ is the parameter matrix of the first hidden layer, $\boldsymbol{B}_2 \in \mathbb{R}^{k \times m}$ is the parameter matrix of the second hidden layer, $\boldsymbol{\beta}_3 \in \mathbb{R}^k$ is the parameter vector of the output layer. $\boldsymbol{b}_1 \in \mathbb{R}^k$, $\boldsymbol{b}_2 \in \mathbb{R}^m$ and b_3 are the parameters of offset neurons. $act(\cdot)$ is activation function, commonly used functions are sigmoid function, relu function, leaky relu function etc. In addition to fully connected neural networks, there are other types of networks, such as Convolutional Neural Networks (CNN) that are well suited for processing image data.

For survival component, due to the need for interpret ability in most medical applications and the good theoretical properties, we still use the Cox model to fit the hazard function $\lambda(t|\boldsymbol{z}_i)$:

$$\lambda(t|\boldsymbol{z}_i) = \lambda_0(t|\boldsymbol{z}_i)e^{\boldsymbol{\gamma}^T \boldsymbol{z}_i}, \tag{4}$$

where $\lambda_0(t|\boldsymbol{z}_i)$ is the baseline function, and $\boldsymbol{\gamma} \in \mathbb{R}^q$ is the coefficient vector for $\boldsymbol{z}$. The cumulative baseline hazard function is $\Lambda_0(t) = \int_0^t \lambda_0(u)du$. The survival function is $S(t) = S_0(t)^{e^{\boldsymbol{\gamma}^T \boldsymbol{z}}}$, where $S_0(t) = e^{-\Lambda_0(t)}$.

The observed data for i th subject is $(t_i, \delta_i, \boldsymbol{x}_i, \boldsymbol{z}_i)$. The density function of population is $f_p(t_i) = -\frac{\partial S_p(t_i)}{\partial t_i} = f(t_i) = [\lambda(t_i)]^{\delta_i} S(t_i)$, where $f(\cdot)$ and $S(\cdot)$ is the density function and survival function for the non-cured subjects, respectively.

Write the vector contains all neural network parameters as $\boldsymbol{w} = (Vec(B_1)^T, \boldsymbol{b}_1^T, Vec(B_2)^T, \boldsymbol{b}_2^T, \boldsymbol{\beta}_3^T, b_3)^T$, where Vec is vector operator. The observed likelihood is

$$
\begin{aligned}
&L(\boldsymbol{w}, \boldsymbol{\gamma}, \lambda_0)\\
&= \prod_{i=1}^{n} \{\lambda(t_i)S(t_i)\}^{\delta_i}\{1-\theta(\boldsymbol{x}_i)+\theta(\boldsymbol{x}_i)S(t_i)\}^{1-\delta_i}\\
&= \prod_{i=1}^{n} \{\lambda(t_i)S(t_i)\}^{\delta_i}\{1-act(act(B_2^T act(B_1^T \boldsymbol{x}_i+\boldsymbol{b}_1)^T\\
&\quad +\boldsymbol{b}_2)^T\boldsymbol{\beta}_3+b_3)+act(act(B_2^T act(B_1^T\boldsymbol{x}_i+\boldsymbol{b}_1)^T+\boldsymbol{b}_2)^T\\
&\quad \cdot\boldsymbol{\beta}_3+b_3)S(t_i)\}^{1-\delta_i}, \qquad (5)
\end{aligned}
$$

where $\lambda(t_i)=\lambda_0(t_i)e^{\boldsymbol{\gamma}^T z_i}$.

2.2 Parameter estimation

In order to optimize (5), following Sy and Taylor (2001), we use EM algorithm and treat $y_i s$ as unobserved latent binary variables, where $y_i = 1$ means that the individual i is non-cured, and $y_i = 0$ means cured. Let $O_i = (t_i, \delta_i, \boldsymbol{x}_i, z_i), i = 1, 2, \ldots, n$ denote the set of observed data. Then the complete likelihood can be written as

$$
\begin{aligned}
L_C(\boldsymbol{w}, \boldsymbol{\gamma}, \lambda_0; y, O) &= L(y)L(O|y)\\
&= \prod_{i=1}^{n}[\theta(\boldsymbol{x}_i)^{y_i}\{1-\theta(\boldsymbol{x}_i)^{1-y_i}\}]\prod_{i=1}^{n}[\{\lambda(t_i)\}^{\delta_i}S(t_i)]^{y_i}. \qquad (6)
\end{aligned}
$$

We use the same method as Cai et al. (2012) to simplify the calculation. Set $\log(y_i)\cdot\delta_i = 0$ and notice that $\delta_i \cdot y_i = \delta_i$, the term $[\{\lambda(t_i)\}^{\delta_i}S(t_i)]^{y_i}$ in (6) can be expressed as

$$
\begin{aligned}
[\{\lambda(t_i)\}^{\delta_i}S(t_i)]^{y_i} &= \{\lambda_0(t_i)e^{\boldsymbol{\gamma}^T z_i}\}^{\delta_i y_i}\cdot S_0(t_i)^{e^{\boldsymbol{\gamma}^T z_i}\cdot y_i}\\
&= \{\lambda_0(t_i)e^{\boldsymbol{\gamma}^T z_i+\log(y_i)}\}^{\delta_i}\cdot S_0(t_i)^{e^{\boldsymbol{\gamma}^T z_i+\log(y_i)}}, \qquad (7)
\end{aligned}
$$

which can be viewed as likelihood for standard Cox model with the additional offset variable $\log(y_i)$. After the transformation, solution in a loop can be obtained using existing software package , e.g. *coxph()* in R software.

The log-likelihood rescaled by $1/n$ for (6) is

$$
l_C(\boldsymbol{w}, \boldsymbol{\gamma}, \lambda_0; y, O) = l_{c,1}(\boldsymbol{w}; y, O) + l_{c,2}(\boldsymbol{\gamma}; y, O), \qquad (8)
$$

where

$$
l_{c,1}(\boldsymbol{w}; y, O) = \frac{1}{n}\sum_{i=1}^{n}[y_i\log\{\theta(\boldsymbol{x}_i)\}+(1-y_i)\log\{1-\theta(\boldsymbol{x}_i)\}], \qquad (9)
$$

and

$$l_{c,2}(\boldsymbol{\gamma}; y, O) \quad = \frac{1}{n}\sum_{i=1}^{n}[\delta_i\{\log\lambda_0(t_i) + \boldsymbol{\gamma}^T \boldsymbol{z}_i\} - y_i e^{\boldsymbol{\gamma}^T z_i} \cdot \Lambda_0(t_i)]. \tag{10}$$

We estimate the cumulative baseline function $\Lambda_0(t)$ by Breslow type estimator

$$\Lambda_0(t) = \sum_{t_l \le t} \frac{d_l}{\sum_{r^* \in R_l} y_{r^*} \cdot e^{\boldsymbol{\gamma}^T z_{r^*}}}, \tag{11}$$

where $R_l = \{i : t_i \ge t_l\}$ is the set of individuals who are at risk for failure at time t_l, and $i = 1, 2, \ldots, n$.

We first initialize $(\boldsymbol{w}^{(0)}, \boldsymbol{\gamma}^{(0)})$, let $y_i^{(0)} = \delta_i$, then estimate $\Lambda_0^{(0)}(t_i)$ by (11). In the mth EM step, denote the current parameter estimation as $\Omega^{(m)} = (\boldsymbol{w}^{(m)}, \boldsymbol{\gamma}^{(m)}, \Lambda_0^{(m)})$, the E-step update latent $y_i s$ by

$$y_i^{(m+1)} = \delta_i + (1-\delta_i)\frac{\theta(\boldsymbol{x}_i)^{(m)} S_0^{(m)}(t_i)^{e^{\gamma^{(m)T} z_i}}}{1 - \theta(\boldsymbol{x}_i)^{(m)}\{1 - S_0^{(m)}(t_i)^{e^{\gamma^{(m)T} z_i}}\}}. \tag{12}$$

With $y^{(m+1)}$ plugged in, the M-step maximizes (9) and (10) respectively to update the parameters for $\boldsymbol{w}^{(m+1)}$ and $\boldsymbol{\gamma}^{(m+1)}$. Repeat iterative process until $||\boldsymbol{w}^{(m+1)} - \boldsymbol{w}^{(m)}||_2^2 \to 0$, $||\boldsymbol{\gamma}^{(m+1)} - \boldsymbol{\gamma}^{(m)}||_2^2 \to 0$ and $||\Lambda_0^{(m+1)}(t_i) - \Lambda_0^{(m)}(t_i)||_2^2 \to 0$, and $||\cdot||_2^2$ is L_2- norm$||\cdot||_2^2$.

3 Asymptotic properties

In this section, we will present consistency of the estimated non-cured probability $\hat{\theta}(\cdot)$ and survival parameters $\hat{\boldsymbol{\gamma}}$, and asymptotic covariance matrix for $\hat{\boldsymbol{\gamma}}$. We use a full connected neural network with single hidden layer containing k neurons to model $\theta(\cdot)$. Let $\boldsymbol{\gamma}_0$ denote the true value of $\boldsymbol{\gamma}$.

3.1 Consistency

The parameter vector $\boldsymbol{w}_{n,k}$ of $\theta(\cdot)$ are marked as necessary to avoid ambiguity, where n is the sample size and k is the number of neurons associated. Our basic denotations are the following.

Let $\theta(\boldsymbol{x}; \hat{\boldsymbol{w}}_{n,k})$ and $\hat{\boldsymbol{\gamma}}$ be minimizers of the following two negative log-likelihoods respectively:

$$\begin{aligned} &- l_{c,1}(\boldsymbol{w}; \hat{y}, O) \\ &\quad = -\frac{1}{n}\sum_{i=1}^{n}[\hat{y}_i \log\{\theta(\boldsymbol{x}_i)\} + (1-\hat{y}_i)\log\{1-\theta(\boldsymbol{x}_i)\}], \end{aligned} \tag{13}$$

$$
\begin{aligned}
&- pl_{c,2}(\boldsymbol{\gamma}; \hat{y}, O)\\
&\quad = -\frac{1}{n}\sum_{i=1}^{n}\delta_i\{\boldsymbol{\gamma}^T z_i - \log\sum_{j=1}^{n}\hat{y}_i e^{\boldsymbol{\gamma}^T z_j}\mathbb{1}(t_i \le t_j)\},
\end{aligned} \tag{14}
$$

where

$$
\hat{y}_i = \delta_i + (1-\delta_i)\frac{\hat{\theta}(\boldsymbol{x}_i)\hat{S}_0(t_i)^{e^{\hat{\boldsymbol{\gamma}}^T z_i}}}{1-\hat{\theta}(\boldsymbol{x}_i)\{1-\hat{S}_0(t_i)^{e^{\hat{\boldsymbol{\gamma}}^T z_i}}\}},
$$

and $\hat{S}_0(t_i) = \exp(-\sum_{t_l \le t_i}\frac{d_l}{\sum_{r^* \in R_l}\hat{y}_{r^*}\cdot e^{\hat{\boldsymbol{\gamma}}^T z_{r^*}}})$. Write $\hat{\theta}(\boldsymbol{x}) = \theta(\boldsymbol{x}; \hat{\boldsymbol{w}}_{n,k})$.

Let $\theta(\boldsymbol{x}; \widetilde{\boldsymbol{w}}_{n,k})$ and $\widetilde{\boldsymbol{\gamma}}$ be minimizers of the following two negative log-likelihoods respectively:

$$
\begin{aligned}
&- l^0_{c,1}(\boldsymbol{w}; y^*_0, O)\\
&\quad = -\frac{1}{n}\sum_{i=1}^{n}[y^*_{0i}\log\{\theta(\boldsymbol{x}_i)\} + (1-y_{0i})\log\{1-\theta(\boldsymbol{x}_i)\}],
\end{aligned} \tag{15}
$$

$$
\begin{aligned}
&- pl^0_{c,2}(\boldsymbol{\gamma}; y^*_0, O)\\
&\quad = -\frac{1}{n}\sum_{i=1}^{n}\delta_i\{\boldsymbol{\gamma}^T z_i - \log\sum_{j=1}^{n}y^*_{0i}e^{\boldsymbol{\gamma}^T z_j}\mathbb{1}(t_i \le t_j)\},
\end{aligned} \tag{16}
$$

where

$$
y^*_{0i} = \delta_i + (1-\delta_i)\frac{\theta^*_0(\boldsymbol{x}_i)\widetilde{S}_0(t_i)^{e^{\boldsymbol{\gamma}_0^T z_i}}}{1-\theta^*_0(\boldsymbol{x}_i)\{1-\widetilde{S}_0(t_i)^{e^{\boldsymbol{\gamma}_0^T z_i}}\}},
$$

where $\widetilde{S}_0(t_i) = \exp(-\sum_{t_l \le t_i}\frac{d_l}{\sum_{r^* \in R_l}y^*_{0r^*}\cdot e^{\gamma_0{}^T z_{r^*}}})$, $\theta^*_0(\boldsymbol{x})$ is the true nonparametric function for $\theta(\boldsymbol{x})$ to generate the data.

We follow the notation used in Fine and Mukherjee (1999). We call (15) the empirical error of the network. Let $\mathcal{S}_\epsilon = \{\widetilde{\boldsymbol{w}}_{n,k} : \nabla\{-l^0_{c,1}(\widetilde{\boldsymbol{w}}_{n,k}; y^*_0, O)\} = \mathbf{0}\}$ denote the set of stationary points and M_ϵ denote the set of (local and global) minima respect to (15). Let generalization error $e_g(\boldsymbol{w}_{n,k}; y^*_0, O)$ be $E\{-l^0_{c,1}(\boldsymbol{w}_{n,k}; y^*_0, O)\}$, the expectation here is respect to covariate $\boldsymbol{x}$. Similarly, we define M_{e_g} to be the set of minima of $e_g(\boldsymbol{w}_{n,k}; y^*_0, O)$ and the set of stationary points $\mathcal{S}_{e_g} = \{\widetilde{\boldsymbol{w}}_{n,k} : \nabla e_g(\boldsymbol{w}_{n,k}; y^*_0, O)|_{\widetilde{\boldsymbol{w}}_{n,k}} = \mathbf{0}\}$. The parameter estimates returned by a training algorithm to minimize the empirical error (15) is given by $\widetilde{\boldsymbol{w}}_{n,k}$, and $\widetilde{\boldsymbol{w}}_0(\widetilde{\boldsymbol{w}}_{n,k}) \in M_{e_g}$ is the nearest neighbor minimum of $\widetilde{\boldsymbol{w}}_{n,k}$ in set M_{e_g}. Write $\widetilde{\theta}(\boldsymbol{x}) = \theta(\boldsymbol{x}; \widetilde{\boldsymbol{w}}_{n,k})$ and $\theta_0(\boldsymbol{x}) = \theta(\boldsymbol{x}; \widetilde{\boldsymbol{w}}_0(\widetilde{\boldsymbol{w}}_{n,k}))$. Notice that $\theta_0(\boldsymbol{x})$ is an optimal estimate of the true function $\theta^*_0(\boldsymbol{x})$ that minimizes the generalization error under a given number of finite neurons k.

We follow the notation of Gu (2013). Set $u(\theta; y^*_0, O) = \mathrm{d}l^0_{c,1}(\theta; y^*_0, O)/\mathrm{d}\theta$ and $w(\theta; y^*_0, O) = \mathrm{d}^2 l^0_{c,1}(\theta; y^*_0, O)/\mathrm{d}\theta^2$. Assume $\mathrm{E}u(\theta; y^*_0, O) = 0$ and $Eu^2(\theta; y^*_0, O) = \sigma^2 Ew(\theta; y^*_0, O)$, where σ is constant. Write $N(t) = I(T \le t, \delta = 1)$, $Y(t) = I(T \ge$

t), and $M(t) = N(t) - \int_0^t Y(u) y_0^* e^{\boldsymbol{\gamma}_0^T \boldsymbol{z}} du$. Then $M(t)$ is a martingale conditional on $\boldsymbol{x}$ and $\boldsymbol{z}$.

We will need the following conditions:

Condition 1 *The domains $\mathcal{X}$ and $\mathcal{Z}$ of covariates $\boldsymbol{x}$ and $\boldsymbol{z}$ are finite and compact, and parameter space $\mathcal{W}$ for $\boldsymbol{w}$ is convex. The two activation functions in the layers (hidden layer and output layer) making up the network are twice continuously differentiable.*

Condition 2 *Given the covariates $\boldsymbol{x}$ and $\boldsymbol{z}$, censoring time C is independent of true failure time T^*. Assume the observations are in a finite time interval [0, τ]. True baseline hazard function $\lambda_0(t) > 0$ is bounded.*

Condition 3 $E[\boldsymbol{z}\exp(\boldsymbol{\gamma}_{\boldsymbol{0}}^{\boldsymbol{T}} \boldsymbol{z})]^2$ *is bounded uniformly in a neighborhood of $\boldsymbol{\gamma}_{\boldsymbol{0}}$.*

Condition 4 *For $\theta(\boldsymbol{x})$ in a convex set B_0 around $\theta_0^*(\boldsymbol{x})$ containing $\hat{\theta}(\boldsymbol{x})$ and $\widetilde{\theta}(\boldsymbol{x})$, $c_1 w(\theta_0^*(\boldsymbol{x}); y_0^*, O) \le w(\theta(\boldsymbol{x}_i); y_0^*) \le c_2 w(\theta_0^*(\boldsymbol{x}_i); y_0^*, O)$ and $c_1^{'}/\theta_0^*(1-\theta_0^*)(\boldsymbol{x}) \le 1/\hat{\theta}(1-\hat{\theta})(\boldsymbol{x}) \le c_2^{'}/\theta_0^*(1-\theta_0^*)(\boldsymbol{x})$, $\forall \boldsymbol{x} \in \mathcal{X}$, for some c_1, c_2, $c_1^{'}$ and $c_2^{'} > 0$.*

Condition 5 (1) *Assume Fourier series expansions for θ exists and can be expressed as $\theta = \sum_\mu a_\mu \phi_\mu$, $Var[\phi_\mu(X) \cdot \phi_\nu(X) w(\theta_0^*(X); y_0^*)] \le c_3$, and $Var\,[\phi_\mu(X)\phi_\nu(X) \frac{S(t;\gamma_0,z)}{[1-\theta_0^*(\boldsymbol{x})(1-S(t;\gamma_0,z))](\theta_0^*(1-\theta_0^*))(\boldsymbol{x})}] \le c_3^{'}$ for some $c_1, c_3^{'} < \infty$, $\forall \mu, \nu$.*

(2) $\int_{\mathcal{U}} m(u) \int_{\mathcal{T}} (\boldsymbol{z}_1^T \boldsymbol{\gamma})^2 (\boldsymbol{z}_2^T \boldsymbol{\gamma})^2 e^{h \boldsymbol{\gamma}_0^T \boldsymbol{z}} P(T \ge t|\boldsymbol{z}) y_0^* dt \le d_1$, *and* $\int_{\mathcal{U}} m(u) \int_{\mathcal{T}} (\boldsymbol{z}_1^T \boldsymbol{\gamma})^2 (\boldsymbol{z}_2^T \boldsymbol{\gamma})^2 e^{h \boldsymbol{\gamma}_0^T \boldsymbol{z}} P(T \ge t|\boldsymbol{z}) y_0^* \cdot \frac{S(t;\gamma_0,z)}{[1-\theta_0^*(\boldsymbol{x})(1-S(t;\gamma_0,z))]} dt \le d_1^{'}$, *for some* $d_1, d_1^{'} < \infty$, $\forall z_1, z_2 \in \mathcal{Z}$, $h = 1, 2$.

Condition 6 *Select a positive sequence δ_n converging to 0 and terminate the optimization algorithm when the following conditions are met: $||\nabla\{-l_{c,1}^0(\hat{\boldsymbol{w}}_{n,k}; y_0^*, O)\}|| < \delta_n$, $\lim_{n\to\infty} \sqrt{n}\delta_n = 0$ and $H_n(\hat{\boldsymbol{w}}_{n,k})$ positive define, where H_n is Hessian matrix.*

Condition 7 *e_g has a finite set M_{e_g} of minima located in the interior of $\mathcal{W}$, and they are all stationary points: $M_{e_g} = \{\widetilde{\boldsymbol{w}}_{n,k}^1, \widetilde{\boldsymbol{w}}_{n,k}^2, \ldots, \widetilde{\boldsymbol{w}}_{n,k}^m\} \subset S_{e_g}$. The Hessian matrix $H_e = \nabla\nabla e_g$ is positive define at each of these interior minima.*

Condition 8 (1) *There exists $\delta > 0$, $\rho < \infty$, such that $H_e(\widetilde{\boldsymbol{w}})$ positive definite and $||\nabla e_g(\widetilde{\boldsymbol{w}})|| < \delta$.* (2) *The first derivative $\nabla\{-l_{c,1}^0(\hat{\boldsymbol{w}}_{n,k}; y_0^*, O)\}_j$ of the neural network respect to the jth parameter $\widetilde{w}_j$ is uniformly bounded (in magnitude) by B_j. For each component $\widetilde{w}_j$ of $\widetilde{\boldsymbol{w}}$ the family*

$$D_j\{I_{(\alpha,\infty)}(\nabla\{-l_{c,1}^0(\boldsymbol{w}_{n,k}; y_0^*, O)\}_j) : \boldsymbol{w}_{n,k} \in \mathcal{W}, |\alpha| \le B_j\}$$

has finite Vapnik-Chervonenkis (VC) dimension.

Condition 1 is an assumption about the boundedness of covariates and parameters of neural networks. Condition 2 assumes non-informative censoring and the boundedness of the baseline hazard. Condition 4 requires big neighborhood B_0 enough to contain the true and the estimated values. Condition 5 is similar to Condition 9.3.4 and

Condition 9.4.4 of Gu (2013). Conditions 6–8 come from Fine and Mukherjee (1999) to derive the convergence of neural network parameters. Condition 6 is a constraint on the neural network training algorithm, that is to ensure that the network is based on gradient training, because one of the conditions for us to determine the minimum weight estimate is that the gradient at this point is very small. Conditions 7–8 are constraints on generalization error. Note that conditions 6–8 are not strictly mathematical conditions, but we must use them to ensure that Theorem 1 of Fine and Mukherjee (1999) holds. Then we have the following theorem.

Theorem 1 *Under Condition 1–8 and the above denotations, as $n \to \infty$, $\exists k \in \mathcal{R}$ such that $||\hat{\theta} - \theta_0^*|| \xrightarrow{P} 0$ and $||\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0|| \xrightarrow{P} 0$.*

This theorem states that when the sample size n tend to infinity, there is a finite neuron number k such that the estimates of $\hat{\theta}$ and $\hat{\boldsymbol{\gamma}}$ respectively converge to their true values in probability. See Web Appendix A for the detailed proof.

3.2 Asymptotic variance estimation

This article only considers the estimation of the asymptotic variance of survival parameters $\hat{\boldsymbol{\gamma}}$. Nielsen et al. (1992), Murphy et al. (1997) and Fang et al. (2005) proposed methods for estimating the asymptotic variance of the parameters for traditional cure rate model. Due to the particularity of our proposed model, we use the method suggested by Sy and Taylor (2001), that is to invert an information matrix derived from the observed likelihood. Law et al. (2002) pointed out that the observed data information is not a by-product of the EM algorithm and adopted the conclusion of Louis (1982) to rewrite the information of the observed likelihood as the difference between the information of the complete likelihood and the missing data:

$$
\begin{aligned}
-\frac{\partial^2}{\partial \boldsymbol{\gamma}^2} \log L(\Omega; O) = &- \frac{\partial^2}{\partial \boldsymbol{\gamma}^2} E\{\log L_c(\Omega; O, y)|O, \hat{\Omega}\} \\
&- Var\Big\{\frac{\partial}{\partial \boldsymbol{\gamma}} \log L_c(\Omega; O, y)|O, \hat{\Omega}\Big\}, \qquad (17)
\end{aligned}
$$

where $\Omega = (\boldsymbol{w}, \boldsymbol{\gamma}, \Lambda_0)$, $O = (t, \delta, \boldsymbol{x}, z)$.

The expectation in the first term can be rewritten as

$$
\begin{aligned}
&E\{\log L(\Omega; O, y)|O, \hat{\Omega}\} \\
&\quad = \sum_{i=1}^{n} E\{\log L_c(\Omega; O_i, y_i)|O, \hat{\Omega}\} \\
&\quad = \sum_{i=1}^{n} \{\log L_c(\Omega; O_i, y_i = 1) \cdot P(y_i = 1|O_i, \hat{\Omega}) \\
&\qquad + \log L_c(\Omega; O_i, y_i = 0) \cdot P(y_i = 0|O_i, \hat{\Omega})\}.
\end{aligned}
$$

For a subject with $\delta_i = 0$,

$$\begin{aligned}
P(y_i = 1|O_i, \hat{\Omega}) &= \frac{L_c(y_i = 1, \delta_i = 0)}{L(y_i = 1, \delta_i = 0)}\Big|_{\Omega=\hat{\Omega}} \\
&= \frac{\theta(\boldsymbol{x}_i)S_0(t_i)^{e^{\gamma^T z_i}}}{1 - \theta(\boldsymbol{x}_i)(1 - S_0(t_i)^{e^{\gamma^T z_i}})}\Bigg|_{\Omega=\hat{\Omega}}, \\
P(y_i = 0|O_i, \hat{\Omega}) &= \frac{L_c(y_i = 0, \delta_i = 0)}{L(y_i = 0, \delta_i = 0)}\Big|_{\Omega=\hat{\Omega}} \\
&= \frac{1 - \theta(\boldsymbol{x}_i)}{1 - \theta(\boldsymbol{x}_i)(1 - S_0(t_i)^{e^{\gamma^T z_i}})}\Bigg|_{\Omega=\hat{\Omega}}, \\
\log L_c(y_i = 1, \delta_i = 0) &= \log\theta(\boldsymbol{x}_i) + e^{\gamma^T z_i} \cdot \log S_0(t_i), \\
\log L_c(y_i = 0, \delta_i = 0) &= \log(1 - \theta(\boldsymbol{x}_i)).
\end{aligned}$$

Therefore,

$$\begin{aligned}
&E\{\log L_c(\Omega; O_i, y_i)|O, \hat{\Omega}\} \\
&\quad= \left\{\frac{(\log\theta(\boldsymbol{x}_i) + e^{\gamma^T z_i} \cdot \log S_0(t_i))\theta(\boldsymbol{x}_i)S_0(t_i)^{e^{\gamma^T z_i}}}{1 - \theta(\boldsymbol{x}_i)(1 - S_0(t_i)^{e^{\gamma^T z_i}})}\right\}\Bigg|_{\Omega=\hat{\Omega}} \\
&\qquad+ \left\{\log(1 - \theta(\boldsymbol{x}_i))\frac{1 - \theta(\boldsymbol{x}_i)}{1 - \theta(\boldsymbol{x}_i)(1 - S_0(t_i)^{e^{\gamma^T z_i}})}\right\}\Bigg|_{\Omega=\hat{\Omega}}. \qquad (18)
\end{aligned}$$

For a subject with $\delta_i = 1$,

$$\begin{aligned}
E\{\log L_c(\Omega; O_i, y_i)|O\} &= \log L_c(y_i = 1, \delta_i = 1) \\
&= \log\theta(\boldsymbol{x}_i) + \log\lambda_0(t_i) + \boldsymbol{\gamma}^T \boldsymbol{z}_i,
\end{aligned}$$

the second derivative of the above term respects to $\boldsymbol{\gamma}$ is zero matrix, so it is omitted.

For the second term on the right side of Eq. (17),

$$\begin{aligned}
&Var\Big\{\frac{\partial}{\partial\boldsymbol{\gamma}}\log L_c(\Omega; O, y)|O, \hat{\Omega}\Big\} \\
&\quad= \sum_{i=1}^{n} E\{(\frac{\partial}{\partial\boldsymbol{\gamma}}\log L_{ci})^2|O, \hat{\Omega}\} \\
&\qquad- \sum_{i=1}^{n}\{E(\frac{\partial}{\partial\boldsymbol{\gamma}}\log L_{ci}|O, \hat{\Omega})\}^2,
\end{aligned}$$

the second term of the above equation is 0 at MLE. Since $\frac{\partial}{\partial\boldsymbol{\gamma}}\log L_{ci} = y_i\delta_i\boldsymbol{z}_i + y_i e^{\boldsymbol{\gamma}^T z_i} \cdot \log S_0(t_i)\boldsymbol{z}_i$, for a subject with $\delta_i = 0$,

$$
\begin{aligned}
&\frac{\partial}{\partial \boldsymbol{\gamma}} \log L_{ci}(y_i = 1, \delta_i = 0) = \{e^{\boldsymbol{\gamma}^T z_i} \cdot \log S_0(t_i)\} z_i, \\
&\frac{\partial}{\partial \boldsymbol{\gamma}} \log L_{ci}(y_i = 0, \delta_i = 0) = \mathbf{0}, \\
&E\{(\frac{\partial}{\partial \boldsymbol{\gamma}} \log L_{ci})^2 | O, \hat{\Omega}\} \\
&\quad = \left\{ \frac{(e^{\boldsymbol{\gamma}^T z_i} \cdot \log S_0(t_i))^2 \theta(\boldsymbol{x}_i) S_0(t_i)^{e^{\gamma^T z_i}}}{1 - \theta(\boldsymbol{x}_i)(1 - S_0(t_i)^{e^{\gamma^T z_i}})} \right\} z_i z_i^T \bigg|_{\Omega = \hat{\Omega}}. \qquad (19)
\end{aligned}
$$

For a subject with $\delta_i = 1$, which means that $y_i = 1$, we have

$$
\begin{aligned}
&\frac{\partial}{\partial \boldsymbol{\gamma}} \log L_{ci}(y_i = 1, \delta_i = 1) = \{1 + e^{\boldsymbol{\gamma}^T z_i} \cdot \log S_0(t_i)\} z_i, \\
&E\{(\frac{\partial}{\partial \boldsymbol{\gamma}} \log L_{ci})^2 | O, \hat{\Omega}\} = \{1 + e^{\boldsymbol{\gamma}^T z_i} \cdot \log S_0(t_i)\}^2 z_i z_i^T \qquad (20)
\end{aligned}
$$

Substituting the above derivations into (17) and inverting the information matrix, the asymptotic variance of $\hat{\boldsymbol{\gamma}}$ can be obtained.

4 Simulation study

We conduct some simulations to evaluate the performance of the proposed method. Specifically, for the cure rate component, we consider two settings: structure and unstructured predictors. We provide access to the corresponding code in Supplementary Material.

4.1 Structured predictors

In this subsection, both the predictors of the survival component and the cure rate component are structured. We consider $p = 10$ in Scenario 1 and $p = 30$ in Scenario 2. In addition, we consider the case of $n \leq p$ in Scenario 3.

4.1.1 Data generation

We set $\boldsymbol{x} = \boldsymbol{z}$ in Scenarios 1 and 2. Scenario 1 considers $p = q = 10$, sample size $n = 100, 1000, 5000$ and 10,000. Scenario 2 considers $p = q = 30$, sample size $n = 300, 1000, 5000$ and 10,000. In Scenarios 1 and 2, $n =$ 10,000 is used to explore whether a large sample size can fit the data well when the true $\theta(\cdot)$ is very complicated with a lot of interaction terms. We consider the case where the sample size is less than the dimension of $\boldsymbol{x}$ in Scenario 3. At this time, $p = 500, q = 20$, $n = 300, 500$. Failure times are generated from a Weibull distribution with a survival function $S(t; a, b) = \exp(-\frac{t}{b})^a$, where a is the shape parameter, $b = \exp(\boldsymbol{\gamma}^T \boldsymbol{z})^{-\frac{1}{a}}$. The non-cured probability $\theta(\cdot)$ is generated from completely nonparametric models, achieving a average cure rate of 50%. The censoring times are generated from exponential distribution independent of the time to failure, achieving a total censoring rate of 60% in average. True $\theta(\cdot)$ includes a large number of interaction terms and non-smooth features such as indicators and absolute values in Scenarios 1 and 2. The detailed settings for $\theta(\cdot)$, a and $\boldsymbol{\gamma}$ are presented in Web Appendix B. 500 replicates are generated for each scenario. Test set with the same sample size as the training set are generated for each scenario for model evaluation.

4.1.2 Simulation results

We use the proposed method with neural network to estimate $\theta(\cdot)$. Besides, we also fit MCM with linear and spline θ estimates and compare the results of different models. We use R package 'smcure' to fit the linear model. For the spline method, we only fit the nonparametric additive model for convenience. We use a thin plate spline and use the 'gam' function of R package 'mgcv' for estimating, where the effective degrees of freedom of each covariate function are automatically selected. For neural networks, we pre-set several different network structures (different neuron and layer numbers), and then train the models to obtain their estimates respectively, and then calculate negative log-likelihood of the test set with the estimates, and finally we choose the network structure with minimum negative log-likelihood of the test set. To reduce computing cost, for each scenario, we only take three datasets to select the structure. The selected structure will be applied to all replicates of each scenario. Detailed neural network structure of each scenario is shown in Web Table 4. In the practical use of neural networks, some measures to prevent over-fitting need to be adopted. Amari et al. (1997) pointed out that cross-validation early-stopping is effective with the intermediate range sample size. Hence we use early-stopping to choose a suitable steps to stop training. See Web Appendix E for the pseudo code for early-stopping.

We use mean area under curve (AUC) of the non-cured probability and mean square error (MSE) for estimator $S_p(t)$ to compare the performance of the three models. The AUC here is about the predicted $\hat{\theta}(\boldsymbol{x}_i)$ and the true non-cured indicator $1 - Y_i$, and we take the average based on the sample size n and replicates number $R = 500$. The MSE for $\hat{S}_p(t)$ is defined as $\frac{1}{R \cdot n} \sum_{r=1}^{R} \sum_{i=1}^{n} (\hat{S}_p(t_{ir}) - S_p^0(t_{ir}))^2$, where $S_p^0(t)$ is true population survival function at t. Table 1 presents the prediction accuracy for three estimation methods. In general, the AUC of neural network (Nnet) methods is higher than that of the nonparametric spline (Nonpar) and the Linear method. When the sample size is small ($p = 10, n = 100$ and $p = 30, n = 300$), the AUCs of the three models are all low, but only the S_p of the Nnet has an MSE less than 0.2. When the sample size increases, the AUC of the Nnet increases notably, and the difference between Nnet and Nonpar also increases. For the models with 10 covariates, the Nnet method shows a higher AUC (0.720 vs. 0.621) and lower MSE of S_p (0.147 vs. 0.155) when compared with the Nonpar method. When the true model has complex interaction terms ($p = q = 30$), even with a sample size of 10,000, the AUC of the Nonpar still does not exceed 0.55, but the AUC of Nnet can reach 0.685. In general, as the sample size and the model complexity increase, the Nnet methods shows a greater advantage than the Nonpar and Linear methods. For Scenario 3 with $n \leq p$, the Nonpar

Table 1 Prediction results of linear method (Linear), nonparametric spline method (Nonpar) and neural network method (Nnet) for structured data on test set

Scenario	p	n	AUC			MSE of Sp		
			Linear	Nonpar	Nnet	Linear	Nonpar	Nnet
Scenario 1	10	100	0.538	0.547	0.556	0.215	0.210	0.189
		1000	0.513	0.593	0.599	0.187	0.161	0.153
		5000	0.506	0.617	0.673	0.180	0.156	0.150
		10, 000	0.504	0.621	0.720	0.720	0.155	0.147
Scenario 2	30	300	0.522	0.525	0.535	0.232	0.219	0.179
		1000	0.513	0.526	0.552	0.225	0.186	0.173
		5000	0.505	0.535	0.660	0.222	0.181	0.165
		10, 000	0.504	0.538	0.685	0.222	0.181	0.170
Scenario 3	500	300	–	–	0.620	–	–	0.136
		500	–	–	0.650	–	–	0.128

and Linear methods are not applicable, but the Nnet still has good prediction power (AUC=0.65) and small MSE of S_p (0.128).

We then show the estimation performance of the survival component. Web Table 1 to 3 present the mean estimator, empirical standard deviation (ESD) and mean of the estimated asymptotic standard error (ASE) over 500 replicates for Cox model coefficients. Except for the case where the sample size is small, the estimated $\boldsymbol{\gamma}$ agrees with its true value very well in other settings and overall ESDs and ASEs are similar.

4.2 Unstructured predictors

In this subsection, predictor $\boldsymbol{z}$ of survival component is still structured with dimension $q = 10$, while predictor $\boldsymbol{x}$ of the cured probability $\theta(\boldsymbol{x})$ is unstructured (images and text data).

4.2.1 Data generation

When $\boldsymbol{x}$ is an image predictor (marked as 'model image-gray'), the used gray-scale images of handwritten digits (28 by 28 pixels) including 10 categories (0–9) are obtained from Mixed National Institute of Standards and Technology (MNIST). Hence the dimension of $\boldsymbol{x}$ is 784. The dataset can be accessed in the 'keras' package of R software by the command: *dataset_mnist()*. We associate image $\boldsymbol{x}$ with the non-cured probability: $\theta(\boldsymbol{x}) = 1$ if the input is a handwritten image of '8' and $\theta(\boldsymbol{x}) = 0$ if an image of '6'. Other categories of handwritten images are not included. Failure times are generated from a Weibull distribution $S(t; a, b) = \exp(-\frac{t}{b})^a$, where a is the shape parameter, $b = \exp(\boldsymbol{\gamma}^T \boldsymbol{z})^{-\frac{1}{a}}$. Here $\boldsymbol{z}$ is a 10-dimensional structured predictor, and $\boldsymbol{\gamma}$ is the same as the setting of Scenario 2 in Sect. 4.2.1. The censoring times are generated from exponential distribution independent of the time to failure, achieving a total censoring rate of 60% in average. The sample size is 10,000 for the training set and 1500 for the test set. We generate 500 replicates for each scenario.

When $\boldsymbol{x}$ is a text predictor (marked as ‘model text’), the used text data is obtained from Internet Movie Database (IMDB) including 50,000 highly polarized reviews. The dataset can be accessed in the ‘keras’ package of R software by the command: *dataset_imdb()*. The dataset including 50% negative and 50% positive reviews. We associate the attitude $\boldsymbol{x}$ of the review with the non-cured probability: $\theta(\boldsymbol{x}) = 1$ if $\boldsymbol{x}$ corresponds to a positive commentpositive review and $\theta(\boldsymbol{x} = \text{negative review}) = 0$. The reviews, that is, the sequence of words has been preprocessed and converted into a sequence of integers, where each integer represents a word in the pre-prepared dictionary. We only keep the first 5000 most common words in the training set, and the low-frequency words will be discarded. Then these features will be converted into one-hot encoding, and each feature indicates whether it belongs to one of the 5000 words. As a sequence, $\boldsymbol{x}$ is a 5000-dimensional one-hot encoded predictor. Failure times are generated from a Weibull distribution $S(t; a, b) = \exp(-\frac{t}{b})^a$, where a is the shape parameter, $b = \exp(\boldsymbol{\gamma}^T \boldsymbol{z})^{-\frac{1}{a}}$. Here $\boldsymbol{z}$ is a 10-dimensional structured predictor, and $\boldsymbol{\gamma}$ is the same as the setting of Scenario 2 in Sect. 4.2.1. The censoring times are generated from exponential distribution independent of the time to failure, achieving a total censoring rate of 60% in average. The sample size is 10,000 for the training set and 3000 for the test set. A total of 500 replicates are generated.

In practical applications, $\boldsymbol{x}$ may also be RBG images. At this time, the dimension p of $\boldsymbol{x}$ is usually much larger than the sample size n. We also conduct a simulation study on such situation (mark as ‘model RBG’). The used images are from a Kaggle competition aims at creating an algorithm to distinguish dogs from cats. The original image is 150 by 150 pixels with three channels. In other words, the dimension of $\boldsymbol{x}$ is $p = 67{,}500$. The dataset and its detailed description can be obtain from https://www.kaggle.com/c/dogs-vs-cats/data. We associate the images $\boldsymbol{x}$ with the non-cured probability: $\theta(\boldsymbol{x} = \text{dog}) = 1$ and $\theta(\boldsymbol{x} = \text{cat}) = 0$. We randomly selected 500 observations from the original dataset as the training set and 300 observations as the test set. Hence in this scenario, the sample size $n = 500$ is much smaller than the dimension p of $\boldsymbol{x}$. Failure times are generated from a Weibull distribution $S(t; a, b) = \exp(-\frac{t}{b})^a$, where a is the shape parameter, $b = \exp(\boldsymbol{\gamma}^T \boldsymbol{z})^{-\frac{1}{a}}$. Here $\boldsymbol{z}$ is a 10-dimensional structured predictor, and $\boldsymbol{\gamma}$ is the same as the setting of Scenario 2 in Sect. 4.2.1. The censoring times are generated from exponential distribution independent of the time to failure, achieving a total censoring rate of 60% in average. A total of 500 replicates are generated. Due to the small sample size in this scenario, we use the method of transfer learning (Yosinski et al. 2014) to preprocess the data with the VGG16 network before modeling it. See the uploaded code for details.

4.2.2 Simulation results

The prediction and estimation results of model image-gray, model text and additional model image-RBG are presented in Table 2. Nonpar and Linear models are not applicable to the unstructured predictor, hence are not included. For the model image-gray, the AUC for the non-cured probability is close to 1, which indicates that for this sce-

Table 2 Prediction and estimation results for unstructured data on test set

Model	AUC	MSE of Sp	Estimator	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
Image-gray	0.999	0.041	$\boldsymbol{\gamma}_0$	0.800	1.200	0.500	1.100	– 0.600	– 1.400	– 0.500	– 0.800	0.500	1.800
			Est	0.798	1.197	0.500	1.101	– 0.600	– 1.399	– 0.498	– 0.799	0.500	1.799
			ESD	0.033	0.037	0.033	0.036	0.017	0.024	0.016	0.018	0.017	0.027
			ASE	0.037	0.042	0.037	0.034	0.020	0.020	0.020	0.020	0.020	0.020
Text	0.907	0.134	$\boldsymbol{\gamma}_0$	0.800	1.200	0.500	1.100	– 0.600	– 1.400	– 0.500	– 0.800	0.500	1.800
			Est	0.800	1.200	0.497	1.098	– 0.599	– 1.395	– 0.499	– 0.798	0.499	1.796
			ESD	0.035	0.035	0.032	0.038	0.017	0.022	0.017	0.019	0.016	0.027
			ASE	0.031	0.035	0.031	0.028	0.017	0.017	0.017	0.017	0.017	0.018
Image-RBG	0.872	0.153	$\boldsymbol{\gamma}_0$	0.800	1.200	0.500	1.100	– 0.600	– 1.400	– 0.500	– 0.800	0.500	1.800
			Est	0.816	1.250	0.511	1.125	– 0.618	– 1.441	– 0.519	– 0.819	0.508	1.854
			ESD	0.162	0.183	0.167	0.183	0.085	0.116	0.085	0.099	0.084	0.134
			ASE	0.141	0.164	0.141	0.130	0.078	0.080	0.078	0.079	0.078	0.081

$\boldsymbol{\gamma}_0$ is the true value of survival parameter $\boldsymbol{\gamma}$, Est is the estimate of $\boldsymbol{\gamma}$, ESD is empirical standard deviation and ASE is the mean of the estimated asymptotic standard error nario, the neural network can almost perfectly predict the cured probability. The MSE of the estimated $S_p(t)$ over 500 replicates is small (0.041). Estimated bias of survival component are also small and ESD and ASE are close to each other. For very sparse text data, our model also shows satisfactory performance in terms of high AUC (0.907) in prediction and small bias for survival components. This shows that for unstructured predictors, with sufficient sample size, our cure rate model can achieve good prediction results.

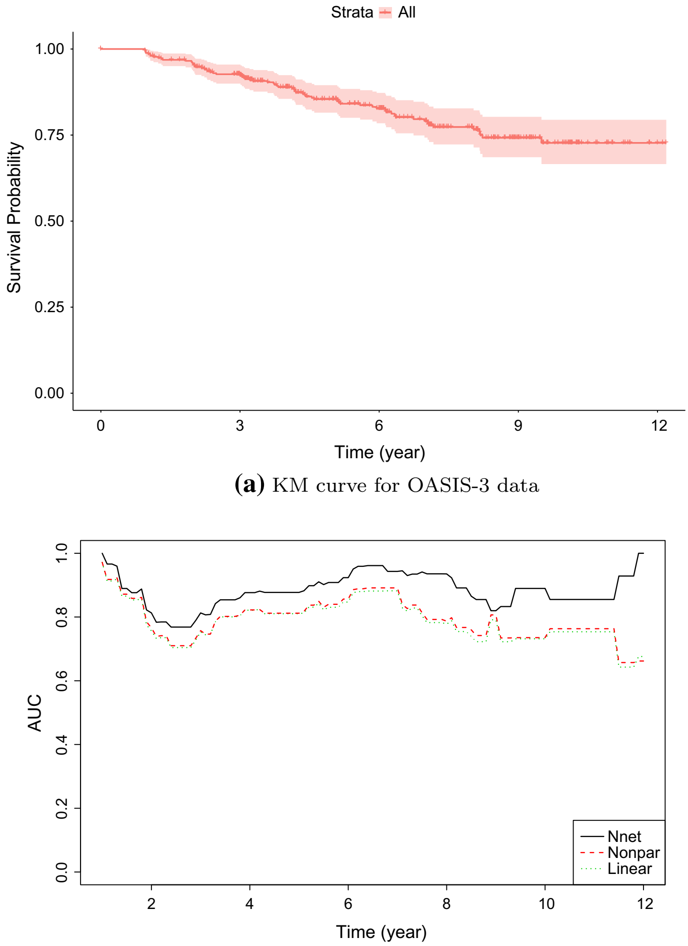
For the image-RBG model with a sample size much smaller than the number of covariates, the AUC for the non-cured probability is 0.872, the MSE of the estimated $S_p(t)$ over 500 replicates is still satisfactory (0.153). The parameters of the survival component are estimated correctly for the small bias. ESD and ASE are generally close. This indicates that our method is also applicable to unstructured $\boldsymbol{x}$ with small n large p, which will facilitate the analysis of clinicopathological images.

In summary, the performance of Nnet models are better than that of Nonpar and Linear models, which is reflected in the highest AUC and the lowest MSE of S_p. For structured predictors, when the true model contains very complex interaction terms and the sample size is small, the AUC of Nnet model and Nonpar model are not much different. When the sample size is sufficient (such as $n = 10{,}000$, $p = 30$), the prediction ability of the Nnet model is significantly higher than that of Nonpar model. This shows that for a sufficient sample size, the neural network can fit the data well. For unstructured predictors, neural networks have shown satisfactory prediction capabilities. For gray image predictors, very sparse text predictors, and RBG image predictors with larger number of covariates than the sample size, the resulting AUCs are all above 0.87. This shows that the Nnet models can fit unstructured data well and are suitable for predicting problems. Next, we apply our method to a real case with image predictors.

5 Application

In this section, we used our method to analyze a real dataset with image predictors. We put additional figures of this section into Web Appendix D. Access to application code is provided in Supplementary Material.

Open Access Series of Imaging Studies (OASIS) is a series of neuroimaging datasets that aimed at making neuroimaging datasets freely available to the scientific community. Wherein OASIS-3 is a longitudinal follow-up dataset for normal aging and Alzheimers Disease (AD) participants, including clinical data and Magnetic Resonance Imaging (MRI) images. The data and the detailed descriptions are available via www.oasis-brains.org. On one hand, we were interested in the hazard of the time from enrollment to AD onset and whether the subject's clinical indicators at baseline would affect the risk of AD. Since not all older people would suffer from AD, we referred to subjects who will not suffer from AD as 'non-susceptible' or 'cured'. On the other hand, it was also of interest to know whether the MRI at baseline was predictive to the susceptibility of AD. Clinical Dementia Rating (CDR) is commonly used to distinguish subjects with and without AD, and CDR > 0 represents demented. Hence we restricted the sample to subjects that were non-demented at the baseline (time at



(a) KM curve for OASIS-3 data

(b) AUC(t) for test set

Fig. 1 Top panel: Kaplan–Meier estimates of OASIS-3 patients. Bottom panel: AUC at different time points for test set

registration). The raw data was cleaned up by retaining MRI of the same size, removing the missing observations and eliminating the outliers. The remaining sample size was 352, 130 of whom were tested for AD at a later follow-up. The enrollment age of the subjects was 42–96 years old, and the mean follow-up time was 6 years. Figure 1 showed a plateau of Kaplan–Meier plots in the later stage of follow up. This indicated that there was a subgroup of patients who were not susceptible to the AD. Therefore, we used the proposed method to analyze the data. The neural network component was employed to process the MRI images $\boldsymbol{x}$ and to calculate the subject's cure probability, the Cox component was used to process the clinical data and assess the hazard of AD.

The structured covariates $\boldsymbol{z}$ used in the Cox component were Mini-Mental State Examination score (MMSE, range was from 0 to 30 with a larger value indicating better mental state), weight (Weight), apolipoprotein E (APOE), logarithm of Geriatric Depression Scale (log(GDS), a high score represents severe depression), intracranial volume (IntraCranialVol), subcortical gray matter volume (SubCortGrayVol). The outcome of survival was duration (years) from registration to event time (the first time observed CDR > 0). The unstructured covariates used in the neural network component was 160 by 200 pixels images taken from MRI images perpendicular to the spine and showed the largest ventricle. Both structured and unstructured data were normalized in range 0 to 1. We randomly divided the dataset into a training set of 280 observations and a test set of 72 observations.

We applied the proposed method to the OASIS-3 data. For model comparison, we also fitted the traditional model. However, for the aforementioned Nonpar and Linear model, image predictors were too cumbersome. Therefore, for the traditional models, we fitted the case where the structured predictor $\boldsymbol{x} = \boldsymbol{z}$ was used, and compared the results with the neural network model using unstructured MRI predictors.

In order to evaluate the difference in prediction ability of different models, we characterized ROC curves between the true event occurrence and the estimated population survival probability $\hat{S}_p(t)$ at different time points and calculate AUC for the test set. We plotted AUC(t) for the Nnet, Nnonpar and Linear cure rate model at different t. Bottom panel of Fig. 1 showed AUC of the Nnet model was basically above 0.8, and it was superior to the AUC of the Linear and Nonpar cure rate model at all time points, which showed that our method performed well in prediction. This indicated that using image predictors instead of structured predictors can improve the prediction accuracy of the cure rate model for OASIS-3 data.

Results for parameter estimation of survival components were presented in Table 3. An older adult with lower MMSE score, higher weight, lower APOE, higher GDS, higher IntraCranialVol and lower SubCortGrayVol was associated with a higher risk of AD. Estimated confidence interval indicated that the effects of MMSE, IntraCranialVol and SubCortGrayVol were significant at level 0.05. Our results are in line with the results of previous studies. Roth (1986), Mortimer (1988) and Satz (1993) showed that a person with higher premorbid brain reserve will have more functional brain tissue remaining at a given level of pathology and will thus develop clinical symptoms at a more advanced biological stage. This is consistent with our finding that the AD risk increases as IntraCraniaVol increases. Hesse et al. (2000) found that the level of APOE in the cerebrospinal fluid reduced in AD patients, however, due to the measurement error and the individual difference, it is not significant in our analysis. Prospective studies of Ganguli et al. (2006) have shown that depressive symptoms cross-sectionally associated with cognitive impairment but not the risk factor for AD. Similarly, GDS was not a significant risk factor of AD in our analysis. The estimated $\boldsymbol{\gamma}$ of the three models was basically the same, but our model had a smaller standard error. In our method, IntraCranialVol and SubCortCrayVol were identified as significant covariates, but they were not significant in spline cure rate model.

Based on the Nnet model, let other covariates be fixed at their median level, we plotted the estimated marginal survival probability $\hat{S}_p(t)$ when MMSE, IntraCranialVol and SubCortGray-Vol at different levels in Web Figure 1. From Web Figure 1a, when MMSE at the normal level, AD survival was high, while when MMSE was lower than 27, in other words, there was cognitive dysfunction, the risk of AD was significantly

Table 3 Survival parameters estimation result of neural network method (Nnet), nonparametric spline method (Nonpar) and linear method (Linear) for the OASIS-3 data

	Model	MMSE	Weight	APOE	log(GDS)	IntraCranialVol	SubCortGrayVol
Nnet	Estimate	−2.732	0.303	−0.655	0.661	4.560	−6.046
	Std.Error	0.697	0.881	0.742	0.469	1.102	1.573
	exp(Estimate)	0.065	1.353	0.520	1.936	95.605	0.002
	95% CI lower bound	−4.099	−1.425	−2.110	−0.258	2.401	−9.129
	95% CI upper bound	−1.365	2.030	0.801	1.580	6.719	−2.964
Nonpar	Estimate	−2.605	0.215	−0.566	0.367	3.335	−3.837
	Std.Error	0.746	0.995	0.939	0.545	1.839	2.807
	Exp(Estimate)	0.074	1.239	0.568	1.443	28.086	0.022
	95% CI lower bound	−4.067	−1.736	−2.405	−0.701	−0.269	−9.339
	95% CI upper bound	−1.144	2.165	1.274	1.434	6.940	1.664
Linear	Estimate	−2.645	0.187	−0.560	0.392	3.432	−3.829
	Std.Error	1.763	1.638	0.864	0.767	1.701	2.347
	Exp(Estimate)	0.071	1.205	0.571	1.480	30.935	0.022
	95% CI lower bound	−6.101	−3.023	−2.253	−1.112	0.098	−8.430
	95% CI upper bound	0.810	3.396	1.132	1.896	6.766	0.771

increased. IntraCranialVol and SubCortGrayVol had a higher AD risk at their 90% and 10% quantiles, respectively. When the covariates were all at the median level (red dotted line in Web Figure 1 (b) or (c)), the survival probability decreased rapidly from 4 years, and finally reached a steady level at 0.2.

Finally, Web Figure 2 explored whether MRI images present different patterns at different $\hat{\theta}(\boldsymbol{x})$ levels. The ventricles and cerebral cisterns of subjects with higher non-cured probability had an expanding trend, and the sulci had a tendency to deepen. These were the phenomena of the elderly brain.

6 Discussion

We proposed mixture cure rate models with a cure rate component accommodating complex predictors, which can handle cure rate component with structured or unstructured predictors like images. The method provide higher prediction accuracy while maintain the interpretation for the survival component. Estimators are showed to be consistent. Our simulation studies for structured data show that our model outperforms the linear cure rate model for all settings. When compared with the nonparametric cure rate model, the neural network method has a wider application even when the predictors are structured covariates. As shown in our simulation studies, when $p = 10$ or 30, if the additive model ignoring the interaction terms is used, the results may be biased. For the case of $n \leq p$, our proposed model still has acceptable prediction ability, while the traditional models are not applicable at this time. On the other hand, our application for OASIS-3 data shows our approach has high and stable predictive accuracy for AD prediction, which can clinically promote the diagnosis and prevention of Alzheimer's disease in the elderly.

In spite of the improvement of our method, some issues deserve more discussions. We only extend the cure rate component of the traditional MCM with a neural network. In fact, it makes sense to introduce unstructured predictors into both cure rate and survival components. Besides, although our simulation and application section use more than one hidden layer neural networks, we only derive the asymptotic results of the neural network with a single hidden layer. And in this article, we only estimate the asymptotic variance for the survival parameters, but not for the neural network parameters. Due to the difficulty of theoretical proof, the above-mentioned problems are limitations that have not been considered in this paper and are worthy of further research. Our simulation studies show that if θ is estimated to be more accurate, the accuracy of estimation for $\hat{\boldsymbol{\gamma}}$ and $\hat{S}_p(t)$ will also increase. Actually, except the fully connected neural networks, other complex deep learning networks such as CNN or some integrated networks (such as AlexNet, VGG, ResNet et al.) can be adopted to accommodate larger, more complex data, which may bring better prediction results. In our simulation and application section, CNN and integrated networks are used for processing image predictors, and the resulting prediction and estimation results are satisfactory. However, this paper only proves the convergence of parameters when using a fully connected network. The theoretical proof under other networks merits further research. On the other hand, neural networks also have their limitations, such as the difficulty and arbitrariness in the selection of the optimal network and the optimal number of training steps, and there is no relatively complete theoretical support; the back-propagation algorithm of training neural network is easy to fall into local minimum, and so on. The limitations of neural networks need further research and discussion, but recently some scholars have made efforts to overcome them, such as Jiang et al. (2003) and Hinsbergen et al. (2009). Next, in this article, we only consider the case where $\boldsymbol{x}$ is structured and unstructured predictor respectively. In practical applications, there may be cases where $\boldsymbol{x}$ has both types of data. For such cases, relevant proofs and simulations are being investigated. Furthermore, Chen and Du (2018) proposed a mixture model with a nonparametric accelerated failure time model (AFT) for the survival component. The model provides a more direct physical interpretation than the proportional hazards and an additional scalar parameter with more flexibility. Similarly, we can extend our model to have an AFT survival component with a neural network in logistic components. Extension to the neural network based nonparametric accelerator factor is to be further explored. On the other hand, promotion time cure rate model is another approach for modeling cure rate data. It is also desirable to develop the promotion time cure rate model with complex predictor. These are of our future research interest.

Supplementary Material

Web Appendices, Tables, and Figures referenced in Sects. 3, 4 and 5 along with code for simulations and applications, are available with this paper at website on Electronic Supplementary Material.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00180-021-01086-3.

Acknowledgements The research was supported in part by National Natural Science Foundation of China (11671256 Yu), and by the University of Michigan and Shanghai Jiao Tong University Collaboration Grant (2017 Yu). OASIS-3 data were provided by Principal Investigators: T. Benzinger, D. Marcus, J. Morris; NIH P50AG00561, P30NS09857781, P01AG026276, P01AG003991, R01AG043434, UL1TR000448, R01EB009352. AV-45 doses were provided by Avid Radiopharmaceuticals, a wholly owned subsidiary of Eli Lilly.

References

Amari S, Murata N, Muller KR, Finke M (1997) Asymptotic statistical theory of overtraining and cross-validation. IEEE Trans Neural Netw Learn Syst 8:985–996

Cai C, Zou Y, Peng Y, Zhang J (2012) Smcure: an R-package for estimating semiparametric mixture cure models. Comput Methods Programs Biomed 108:1255–1260

Chen T, Du P (2018) Mixture cure rate models with accelerated failures and nonparametric form of covariate effects. J Nonparametric Stat 30:216–237

Ching T, Zhu X, Garmire LX (2018) Cox-nnet: an artificial neural network method for prognosis prediction of high-throughput omics data. PLoS Comput Biol 14:e1006076

Cho J, Lee J, Bang H, Kim S, Park S, An J et al (2017) Programmed cell death-ligand 1 expression predicts survival in patients with gastric carcinoma with microsatellite instability. Oncotarget 8:13320–13328

Csàji BC (2001) Approximation with artificial neural networks. Dissertation, Eotvos Lorànd University

Cybenko G (1989) Approximation by superpositions of a sigmoidal function. Math Control Math Control Signals Syst 2:303–314

Fang HB, Gang L, Sun J (2005) Maximum likelihood estimation in a semiparametric logistic/proportionalhazards mixture model. Scand J Stat 32:59–75

Faraggi D, Simon R (1995) A neural network model for survival data. Stat Med 14:73–82

Farewell VT (1986) Mixture models in survival analysis: are they worth the risk. Can J Stat-Revue Canadienne de Statistique 14:257–262

Ferrara R, Pilotto S, Caccese M, Grizzi G, Sperduti I, Giannarelli D et al (2018) Do immune checkpoint inhibitors need new studies methodology? J Thorac Dis 10:S1564–S1580

Fine TL, Mukherjee S (1999) Parameter convergence and learning curves for neural networks. Neural Comput 11:747–769

Fleming TR, Harrington DP (2005) Counting processes and survival analysis. Wiley, New York

Ganguli M, Du Y, Dodge HH, Ratcliff GG, Chang CCH (2006) Depressive symptoms and cognitive decline in late life. Arch Gen Psychiatry 63:153

Gu C (2013) Smoothing spline ANOVA models, 2nd edn. Springer, New York

Halbert W (1990) Connectionist nonparametric regression: multilayer feedforward networks can learn arbitrary mappings. Neural Netw 3:535–549

Hesse C, Larsson H, Fredman P, Minthon L, Andreasen N, Davidsson P et al (2000) Measurement of apolipoprotein e (apoe) in cerebrospinal fluid. Neurochem Res 25:511–517

Hinsbergen CPIV, Lint JWCV, Zuylen HJV (2009) Bayesian committee of neural networks to predict travel times with confidence intervals. Transp Res Part C Emerg Technol 17(5):498–509

Hornik K (1991) Approximation capabilities of multilayer feedforward networks. Neural Netw 4:251–257

Jiang W, Liu Q, Liu T (2003) Drawbacks of neural network learning algorithms and countermeasures. Mach Tool Hydraul 5:29–32

Katzman JL, Shaham U, Cloninger A, Bates J, Jiang T, Kluger Y (2018) Deepsurv: personalized treatment recommender system using a cox proportional hazards deep neural network. BMC Med Res Methodol 18:24

Kuk AYC, Chen CH (1992) A mixture model combining logistic regression with proportional hazards regression. Biometrika 79:531–541

Law NJ, Taylor JM, Sandler H (2002) The joint modeling of a longitudinal disease progression marker and the failure time process in the presence of cure. Biostatistics 3:547–563

Liu X, Peng Y, Tu D, Liang H (2012) Variable selection in semiparametric cure models based on penalized likelihood, with application to breast cancer clinical trials. Stat Med 31:2882–2891

Louis TA (1982) Finding the observed information matrix when using the EM algorithm. J R Stat Soc Ser B Stat Methodol B 44:226–233

Lu W (2010) Variable selection in semiparametric cure models based on penalized likelihood, with application to breast cancer clinical trials. Stat Sin 20:661

Masud A, Tu W, Yu Z (2016) Variable selection for mixture and promotion time cure rate models. Stat Methods Med Res 27:2185–2199

McCulloch WS, Pitts W (1943) A logical calculus of the ideas immanent in nervous activity. Bull Math Biophys 5:115–133

Mortimer JA (1988) Do psychosocial risk factors contribute to Alzheimer's disease? Etiology of dementia of Alzheimer's type. Wiley, New York

Murphy SA, Rossini AJ, van der Vaart AW (1997) Maximum likelihood estimation in the proportional odds model. J Am Stat Assoc 92:968–976

Nielsen GG, Gill RD, Andersen PK, Sørensen TIA (1992) A counting process approach to maximum likelihood estimation in frailty models. Scand J Stat 19:25–43

Othus M, Li Y, Tiwari RC (2009) A class of semiparametric mixture cure survival models with dependent censoring. J Am Stat Assoc 104:1241–1250

Roth M (1986) The association of clinical and neurological findings and its bearing on the classification and aetiology of Alzheimer's disease. Br Med Bull 42:42

Rumelhart DE, Hinton GE, Williams RJ (1986) Learning representations by back-propagating errors. Nature 323:533–536

Satz P (1993) Brain reserve capacity on symptom onset after brain injury: a formulation and review of evidence for threshold theory. Neuropsychology 7:273–295

Sy J, Taylor J (2001) Standard errors for the Cox proportional hazards cure model. Math Comput Model 33:1237–1251

Tandon R, Adak S, Kaye JA (2006) Neural networks for longitudinal studies in Alzheimer disease. Artif Intell Med 36:245–255

Tsiatis A (1981) A large sample study of cox's regression model. Ann Stat 9:93–108

Wang L, Du P, Liang H (2012) Two-component mixture cure rate model with spline estimated nonparametric components. Biometrics 68:726–735

Yin G, Ibrahim JG (2005) Cure rate models: a unified approach. Can J Stat-Revue Canadienne de Statistique 33:559–570

Yosinski J, Clune J, Bengio Y, Lipson H (2014) How transferable are features in deep neural networks? Int Conf Neural Inf Process Syst 2:3320–3328

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.